Gaining Insights into Web Customers using Web Intelligence

Sarabjot S. Anand[1], Matthias Baumgarten[2], Alex Büchner[1], Maurice Mulvenna[1]

Abstract. The web is fast maturing into an important marketing medium that provides businesses with the ability to undertake one-to-one marketing and provide personalised services to their customers. Log files provide a rich source of information about customers that can be used to achieve these services. However, the knowledge required to undertake such activity is far removed from what is provided by the current breed of analysis tools for web log files. In this paper we describe an innovative Web Intelligence tool, Easyminer[3], that provides marketers in e-commerce businesses with access to data mining technology that can sift through large amounts of data collected automatically on customer interaction with the businesses web site and bring to light useful marketing knowledge.

1 PROBLEM DESCRIPTION

As web sites develop into a mature medium for customer interaction, organisations are realising the gap between the information they require to support electronic customer relationship management (eCRM) and the information provided by the current suite of software aimed at analysing web site traffic, called log analysis tools. While organisations want to know how to increase profit per checkout, increase retention of customers, increase browser to buyer conversion rates and reduce clicks-to-close rates, log analysis tools provide information such as frequency of page accesses to individual pages on the web site and counts of high level IP domains.

Organisations have typically invested a lot of money into developing their web sites and web strategy. Now they are looking 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]. So clearly the answer to this problem lies elsewhere.

The technique used for measuring the success of a web site clearly depends on what the goal of setting up the web site is in the first place. A web site is commonly used for:

- Selling products and services
- Providing product/ company information
- Providing customer support online to reduce customer service costs

---

[1] MINEit Software Limited, Newtownabby, Conty Antrim, Northern Ireland, {sarab, alex, maurice}@mineit.com .
[2] University of Ulster at Jordanstown, Newtownabby, County Antrim, Northern Ireland, M.Baumgarten@ulst.ac.uk .
[3] MINEit and Easyminer are registered trademarks of MINEit Software Limited. All rights reserved.

Using page hits 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.

Businesses also want to measure the success of advertising on the web. According to the Forrester report, 86% of the time banner adds that have the highest click rates do not have a high browser-to-buyer conversion rate [1]. So click-through based metrics are not necessarily measures of advertising success. A better measure is the percentage of customers attracted to the site through the banner ad that are retained as customers (that is, they return to the web site again within a certain time frame).

In addition to measuring the success of their web sites and the success of banner ads, organisations also see the web as a medium for one-to-one marketing and providing personalised services. However, to achieve this they need to glean as much information about the customer from the interaction of the customer with the web site.

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 allowing them to achieve their goals of personalised services and one-to-one marketing. These tools are collectively referred to as Web Intelligence tools.
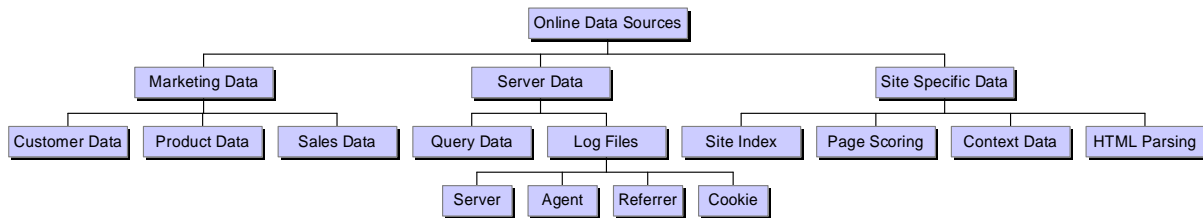
2 EASYMINER

Easyminer is a web intelligence tool that allows user to discover knowledge that they can use to measure the effect of e-marketing campaigns, discover segments in their e-customer base based on interaction with the web site, discover patterns of customer behaviour across web browsing sessions and report on these findings. The typical questions that a marketer in an e-retailer organisation needs to answer are:

- How can I increase my browsers-to-buyer conversion rate? This is the most direct measure of return on investment for an e-retailer measuring what percentage of browsers actually buy something off the web.
- How can I increase my retention rate? Retention in this case may be defined as the ratio of the number of browsers that return to the site within a predefined time window to the total number of browsers
- How can I reduce the clicks-to-close value? Clicks-to-close may be used as a metric used to measure how easily customers can find what they came to the site for. Personalisation of web-based services should reduce this value. As browsing the web using mobile phone becomes more common this measure may become a critical factor for retaining customers.

Figure 1: Data available in an Electronic Commerce Environment

- Does my web site design satisfy the needs of my various customer segments or are their certain segments that have a lower retention rate?

Internet Service Providers (ISP) on the other hand want to know the amount of traffic on the web site, so website statistics of the form provided by log analysis tools are useful to them for ensuring that their infrastructure can handle these volumes of web traffic. However, increasingly ISPs are transforming into a dual role of portal sites too. This implies that now they need to understand their customers better to personalise the information provided to individual customer needs. Once again analysis of customer navigation through the web can provide valuable information for carrying out this personalisation with minimal user input.

2.1 WEB DATA PRE-PROCESSING

The first step towards acquiring web intelligence is the collation of data and pre-processing it into a form that can be used for discovering knowledge.

The data available in electronic commerce environments is three-fold (Figure 1) 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 [4,5]. 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, registration data 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 [6]. This format combines the server and error logs into one file. More recently, the Extended Logfile Format [7] has been used, which consolidates the Common format with additional information, namely the referrer and cookie information. Easyminer also provides facilities for user defined log files to be read in case none of the pre-defined log file formats have been used. For example, a business may also want to store customer identification data with the log file to provide a key to joining this data to other customer data collected through customer interaction using channels other than the web.

The incorporation of referrer information results in the output of the mining of these log files being much more useful and actionable in marketing terms. For example, people often locate a site using search engines. The referrer field contains the search engine, and the search terms. Easyminer supports data input from each of these web log formats and provides access to the data in a relation form (see Figure 2). This allows users to drill-down into graphs generated from the log files and enrich log file data by joining them to web-site registration data [2], marketing databases, customer databases and external demographic and psychograhic data sets, such as those that are provided by Axiom and Experian. It also allows the identification of navigational sessions rather than treating individual page hits as separate events and various analyses on this data, such

as the discovery of navigational sequences and segmentation to be carried out on the data.

Figure 2: Relational view of Log file data

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 [8].

2.2 SEGMENTING VISITORS

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. E-retailers are not limited to analysing sales data. They can also discover similarity between customers based on their navigational behaviour, that is, products they may have browsed but not necessarily bought. This is particularly useful in the light of a recent statistic that internet shoppers are three times more likely to buy offline than online [9]. This means that while a number of online shoppers make their decision to buy online the actual commerce takes place in shops. If this commerce takes place in a competitors store the only way of segmenting the browser base is to use navigational information. Surveys of online shoppers also indicate that the percentage of online shoppers that actually make purchases online is increasing, thus retaining customer interest is an important goal of e-retailers.
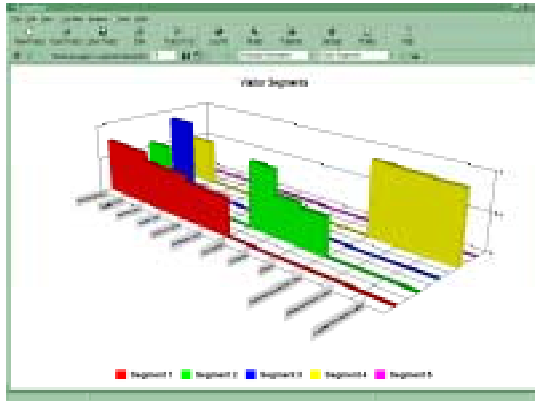
**Figure 3:** Browser Segments based on Navigational Behaviour

The implication of these survey results on advertising is also interesting (see section on Application Building) as the justification for advertising must be based on navigation behaviour rather than sales.
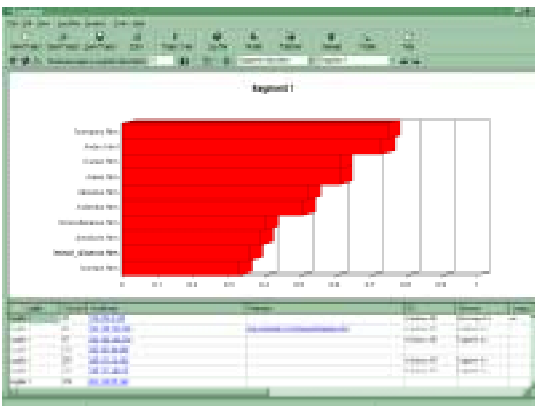
Figure 4: Description of two Session Segments

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. Presently 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 logfile data and the minimum number of sessions needed for a

cluster to be assumed to be valid. Output from the clustering is shown in Figure 3. The pie-chart on the right side shows the number of sessions within each of the segments while the graph on the left shows the spatial orientation of the segments, the 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. Figure 4 shows the descriptions of two of the segments found in a MINEit Software Limited Web Site. As we can see, segments are described in terms likelihood of a web page being visited during the sessions that belong to the segment. In the example descriptions in Figure 4 two distinct segments of navigation sessions are shown. The first one (labelled as Segment 3) consists of sessions aimed at learning more about the company while the second segment (labelled as Segment 1) consists of sessions that belong to browsers who are clearly interested in the products offered by MINEit software limited. Below each description is the list of sessions that belong to the segment. The table displaying all the data available on the sessions including referrer information, query strings etc. On further investigation of the segments we discovered that a large percentage of sessions within Segment 2 had as referrer the kdnuggets web site where MINEit has links to its product pages. This shows that the links on the kdnuggets site is useful in attracting product oriented browsers onto the MINEit web site, resulting in very precise sessions.
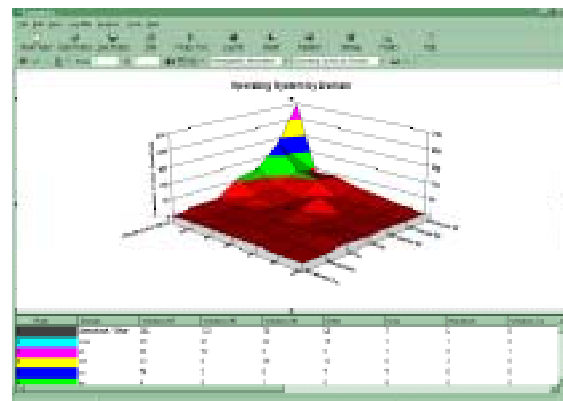
Figure 5 Characterising Segments using high level domain names and Operating System

Once the segments have been discovered, the sessions belonging to a cluster can be extracted into a new log data table and can be analysed further using further segmentation or characterised using various summary graphics as shown in Figure 5. Further investigation of the sessions within the cluster may be carried out using drill-down facilities that result in more detailed graphs as shown in Figure 6.

Up until this point we have only discussed the clustering of sessions rather than visitors (browser and buyers). A visitor is characterised by a number of sessions. Each of these sessions provide information of the interests of the customer. Thus, to discover clusters of customers, different sessions associated with a single customer need to be identified and treated as one entity. In general the IP address of the browser provides a means for identifying a customer and their associated sessions. However, this technique fails in a majority of cases, for example, browsers accessing the web site through their ISP would not have unique IP addresses. To get around this, most web sites use unique cookies that are deposited on the client machine on the first visit to the web site. While this is generally satisfactory, a recent survey suggested

that as many as 10% of internet browsers set up the client software to block cookies.

Figure 6: More detailed view of segment characteristics, shown in Figure 5, using drill-down

However, where cookies are available they are at present the most reliable way of tying together sessions belonging to the same customer and segmentation must then be carried out at the customer rather than at the session level.

2.3   DISCOVERING SITE HIGHWAYS

Navigation of a web site is temporal in nature, thus, one of the basic forms of knowledge that needs to be discovered from data collected in web logs is navigational sequences that describe the most commonly tread pathways (defined based on a threshold value of sessions that follow the pathway, referred to as support) through the web site. Easyminer uses the Capri sequence discovery algorithm [10] for discovering sequences.

Two types of sequences may be discovered within Easyminer: Open sequences and clickstreams. 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 clickstream is a special type of sequence where the pages accessed are contiguous navigations. Thus a clickstream of the form <index.html, orderform.html> does imply that a direct link exists between the index.html and orderform.html page and this link was navigated by the customers 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 middle page of a site, a URL providing information about a special marketing campaign, and a regularly specified end page, where a purchase can be finalised.
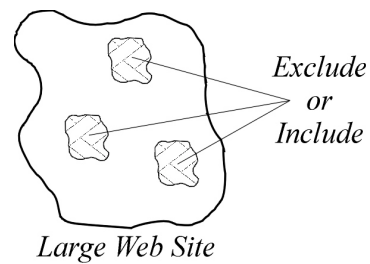

Figure 7: Usage of Network Domain Knowledge.

The second type of taxonomical domain knowledge is that of *network* topologies, which is useful when the topology of web site or only a sub-network of a large site is of interest to the user for the discovery of sequences. The second type of taxonomical domain knowledge is that of *network* structures, which is useful when the topology of web site or only a sub-network of a large site has to be used for the discovery of knowledge. This domain knowledge it is used to include or exclude certain parts of a web site from the analysis, as shown in Figure 7.
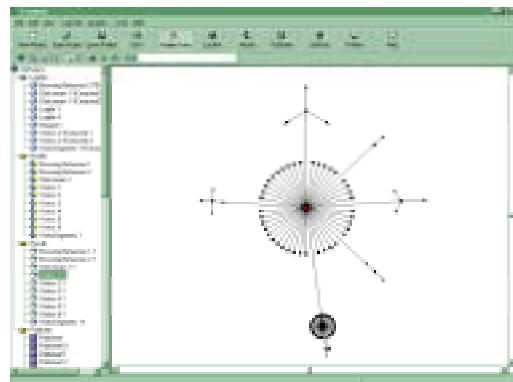

Figure 8: Site Map constructed from logfile.

Network topology domain knowledge within Easyminer, is specified through a site map that is constructed from the log file being analysed (see Figure 8). 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 run of Capri. A network on the other hand is based on the structure of the web site and so 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 the sequences exist within Easyminer. The first method uses the site map and overlays the sequences so that the user can see the sequences within the context of the web site, the log files of which, they have been discovered from. The alternative method is to use the sequence tree view as shown in Figure 9. The quadruple shown at the leaf node of each branch of the tree shows the size of the sequence (the number of pages within the sequence), the number of occurrences of the sequences within the log file, the support and confidence associated with the sequence.
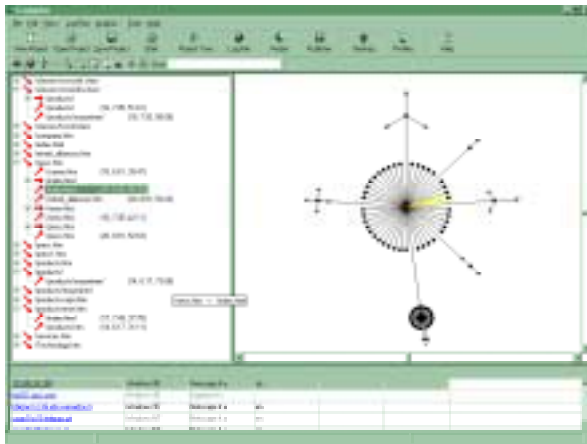
Figure 9: Sequences Tree View

2.4 REPORTING AND KNOWLEDGE DEPLOYMENT

All models developed within Easyminer can be exported to HTML and Microsoft Office documents. Future versions of Easyminer will contain a deployment engine that will allow knowledge discovered by Easyminer to be represented in a format that can be interpreted by the web server and deployed to provide personalised services to browsers of the web site.

3 APPLICATION BUILDING

In this section we describe how Easyminer may be used to provide solutions where traditional methods for analysing web logs have been found to be inadequate. Consider the following scenario.

WXYZBargains.com, a retailer on the web sells sports equipment, toys, books and music CDs. To promote their goods, the retailer decides to pay for banner ads to be placed on four web sites that it believes to be relevant to each of its product lines. In preparation for this advertising campaign marketers at WXYZBargains have set out what they believe would be routes that they would expect customers attracted from each of these banner adds to navigate, each path being distinct as the banner adds are expected to attract customers with different interests. For example, the sports banner ad advertises cheap skiing equipment. Once customers are attracted to the web site through this banner add, the marketers expect that they would investigate some other goods too, such as designer skiing clothes and instruction videos. They have also decided that they would measure the success of the banner ads based on the number of goods, other than those advertised in the banner ad, viewed by the customer. If this number is 10% greater than those viewed by customers that enter the web site through the WXYZBargains home page and view product ranges associated with skiing, the banner ad will be considered to be successful.

The first step is to segment the customer base. The expected result would be four distinct segments, one for each banner add and a number of segments defined on customers that enter the web site from the home page. The segments pertaining to the banner ads would be expected to have small diameters (see Figure 3) as sessions with one of the banner ads as the referrer would be expected to have much more similar navigational behaviour than customers arriving at the web site through the home page. Once the segments have been discovered they can be investigated further using summary graphics such of those shown in Figures 5 and 6.

Segments that display unexpected navigational behaviour may present opportunities for improved ad placement, improved ad content or an improved product line. For example, if the banner ad for CDs seems to result in sessions with navigational behaviour that is expected from the sessions referred by the banner ad for books clearly the banner ad for CDs is not effective. Also, the web site at which the CD banner ad was placed may actually be more effective as a location for a banner ad on books instead.

The sequence discovery algorithm can be used to discover pathways through the site using the different banner ads as the first page, defined using navigational templates. Once again, different pathways would be expected for customers who arrive at the web site through clicking on a banner ad. If the sequences discovered from sessions originating from the home page are similar to that discovered from sessions referred from one of the banner ads, clearly the ads are not being effective.

Easyminer is currently being beta-tested and initial feedback from the testers in very positive. Version 1 of the tool will be ready for release in the second quarter of 2000.

4 BENEFITS OF 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 for discovering the knowledge required to achieve goals such as one-to-one marketing and personalisation of web services. It also provides a means for measuring success of on-line business and return on investment into web strategies through the calculation of business goal defined e-metrics.

References

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

[2] Dan R. Greening. Data Mining on the Web – There's gold in the mountain of Data, Web Techniques, January, 2000-01-26

[3] Jesus Mena. Data Mining your Website, Digital Press, 1999

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

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

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

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

[8] World Wide Web Consortium (W3C), Logging Control In W3C httpd, 1997. http://www.w3.org/XML/

[9] Jeff Moore. Shopping online but buying offline, http://gt.clinckz.com, November, 1999

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