Discovering & Deploying Marketing Knowledge through Web Usage Mining

M.D. Mulvenna, A.G. Büchner, S.S. Anand and J.G. Hughes

MINE*it* Software Ltd, Faculty of Informatics, University of Ulster
Shore Road, Newtownabbey, Co. Antrim, BT37 0QB, N. Ireland
email: maurice@mineit.com { ag.buchner, ss.anand, jg.hughes}@ulst.ac.uk

phone: +44 (0)28 90368394   fax: +44 (0)28 90366068

Abstract

In the present competitive environment, organisations need to retain existing high-value customers to remain competitive. One technique that can be used to achieve greater loyalty from customers is to personalise services provided. Such customisation of services not only helps customers, by satisfying their needs, but also results in customer loyalty. Electronic commerce sites provide organisations with a lot of information about their customers - information that can be used to personalise services to customers. Web usage mining is a new discipline that addresses these needs, whose key principles are presented in this paper. They include different types of online data, novel kinds of domain knowledge, as well as the discovery of marketing intelligence itself. All concepts have been incorporated within a commercial product called EasyMiner and real-world experiments have been carried out.

1 Introduction

Electronic commerce sites not only provide an additional channel for marketing and sales, they also provide a rich source of information about an organisation's customers. The key customer-related key disciplines in marketing discussed in this paper are attraction, retention, relationship, and reward. Data collected at electronic commerce sites can help organisations to be more effective in attracting new customers, retaining high-value customers, cross sales and pre-empting departure. This paper introduces the concept of web log usage mining, describes discrepancies between data and domain knowledge in traditional marketing and web log usage mining exercises, and outlines the discovery and deployment of discovered online marketing intelligence.

The outline of the paper is follows. In Section 2, the processing of data found in online sites and its pre-processing is described. In Section 3, typical Internet domain knowledge is presented, including a mechanism how to incorporate such expertise in data mining exercises. Section 4, describes procedures of discovering marketing intelligence in the form of navigational customer behaviour, before, in Section 5, the discovered patterns are employed in a real-world scenario. In Section 6, related work is evaluated, before conclusions are drawn in Section 7. All concepts have been incorporated within a commercial product called EasyMiner.

2 Online Data Processing

2.1 Online Data Sources

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 (Büchner & Mulvenna, 1998; Mulvenna, Norwood, Büchner, 1998). 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.

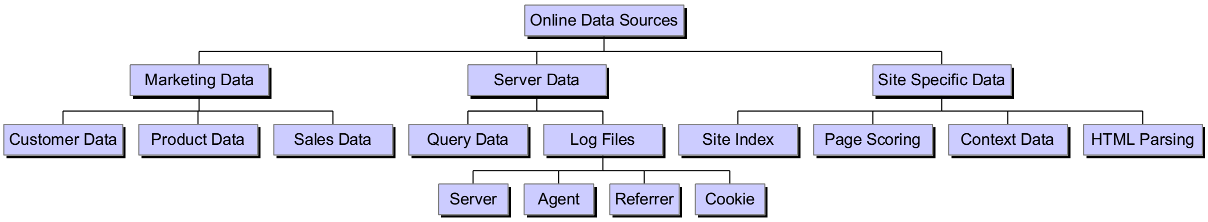


Figure 1. Data Available for Web Mining

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 (W3C, 1995). This format combines the server and error logs into one file. More recently, the Extended Logfile Format (W3C, 1996) has been used, which consolidates the Common format with additional information, namely the referrer (misspelled 'referer' in the original standard) and cookie information. The incorporation of referrer information results in the output of the mining of these logfiles 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. Figure 2 illustrates an entry from a typical Extended Logfile containing search engine information.

```
195.152.154.6 - - [10/Aug/1999:13:51:50 +0000] "GET /career.htm HTTP/1.0" 200 5288
"http://www.excite.co.uk/search.gw?look
=excite_uk&search=career+mine&tsug=1&csug=10&lang=en&c=web.noporn&sorig=rpage&trace=b"
"Mozilla/4.0 (compatible; MSIE 4.01; Windows 95)"
```

Figure 2. Example Entry from an Extended Logfile

Cookies are tokens generated by the web server and held by the clients. The information stored in a *cookie* helps to ameliorate the explicitly 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. Usually the following information is contained in a cookie: User ID, source IP address, TTL (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 generally generated when users of the web site use search facilities on the web site to search for relevant pages/products.

Any organisation that uses the Internet to trade in services and products uses some form of information system to operate Internet retailing. Clearly, some organisations use more sophisticated systems than others. The lowest common denominator information that is typically stored is about customers, products and transactions, each in different levels of detail. More sophisticated electronic traders also keep track of customer communication, distribution details, advertising information on their sites associated with products and / or services, demographic and other information from providers such as Axciom and Experian.

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 (W3C, 1997).

2.2 Online Data Preparation

As well as standard semantic and schematic heterogeneity resolutions across Internet data (see Büchner & Mulvenna for details), online information is ideally represented in a data warehousing environment. A typical web log data hypercube is depicted in Figure 3.

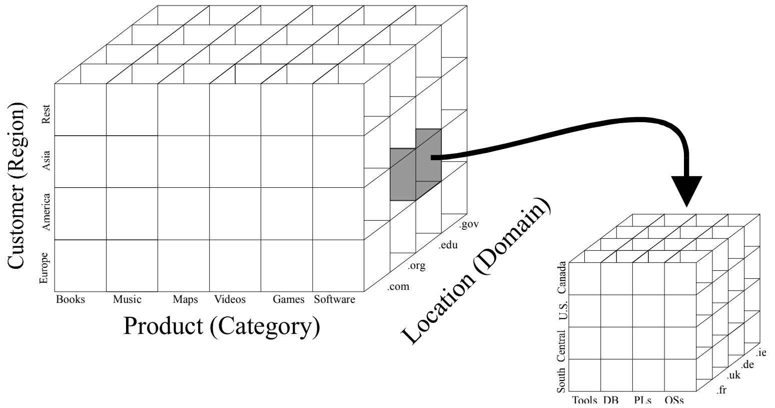


Figure 3. Web Log Data Cube

From this cube, which is based on the example web log snowflake schema below, it is a straightforward procedure to create multiple materialised views using basic OLAP functionality (see Figure 4), which can be used as input for data mining exercises.

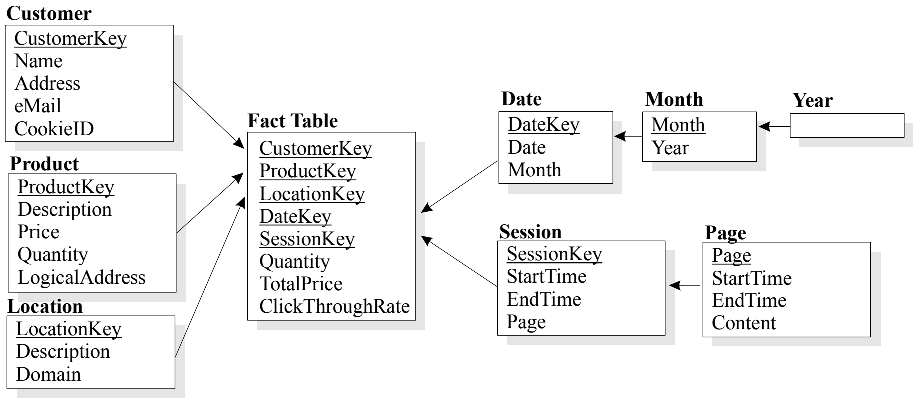


Figure 4. An Example Web Log Snowflake Schema

3 Domain Knowledge Incorporation

As in most knowledge discovery domains, there are two types of domain knowledge that are relevant for web log usage mining: methodology and algorithm dependent thresholds (not further discussed in here) as well as problem- and domain-specific general knowledge and constraints. For the purpose of discovering marketing intelligence from Internet log files, two types of web-specific (problem/domain specific) domain knowledge are outlined in this paper, namely navigation templates and topology networks. More general domain knowledge like concept hierarchies are also supported but not discussed here.

Domain knowledge is used to constrain the search space of navigational patterns of interest and to reduce the granularity of the data so as to increase the visibility of sequences within the data.

3.1 Navigation Templates

In order to perform goal-driven navigation pattern discovery it is almost always necessary that a virtual shopper has passed through a particular page or a set of pages. Navigation templates describe the form of sequences of interest to any level of specificity as required by the user. The template can be used to specify start pages, end pages, middle pages as well as pages that should not appear in a sequence of interest. A typical start item is the home page of an electronic commerce site, a middle item a page connected to a search engine, and a regularly specified end item, where a purchase can be finalised.

An example illustrates the concept of navigation templates. Envisage the analysis of a pre-Christmas marketing campaign in an online bookstore that has introduced reduced gift items. The template is shown in Figure 5 below.

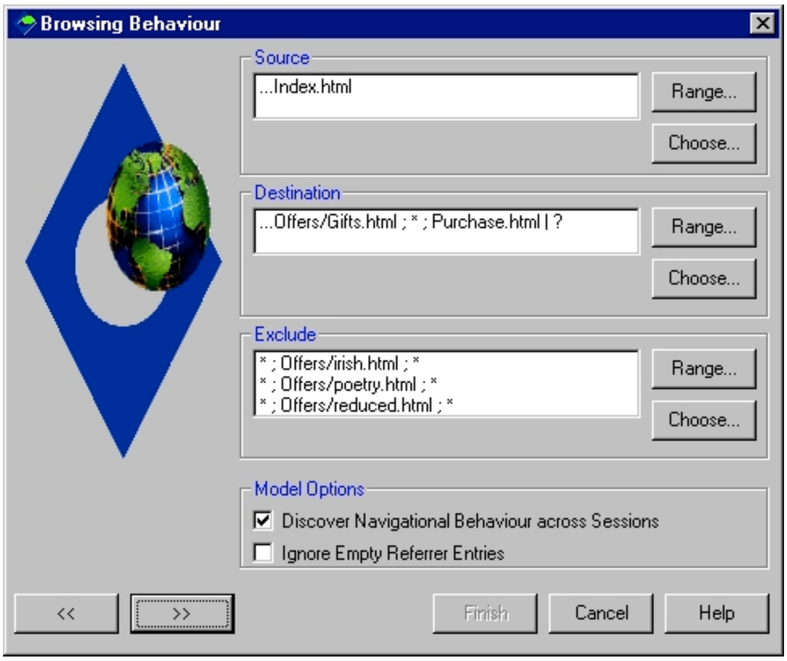
```
[
< index.html | * | offers/gifts.html ; * ; purchase.html | ? >
^< * ; offers/reduced.html ; * >
^< * ; offers/poetry.html; * >
^< * ; offers/irish.html ; * >
]
```

Figure 5**. Example Navigation Template**

Here, the asterisk (*) is a placeholder for a number of web pages while the ‘?’ is a placeholder for a single page. A semi-colon indicates the end of a navigation session while ‘|’ indicates the continuation of a navigation session. Finally the symbol ‘^’ symbolises a negation. Thus, the interpretation of the template in Figure 5 would be as follows:

> *We are interested only in navigation sequences that start at the home page, “index.html” and end at “offers/gifts.html” and are then followed by new navigation by the same customer, resulting in a purchase. However, any navigation that includes “reduced.html”, “irish.html” or “poetry.html” is ignored in the analysis.*

Figure 6 illustrates the navigation template usage in the EasyMiner product.



Source
...Index.html

Destination
...Offers/Gifts.html ; * ; Purchase.html | ?

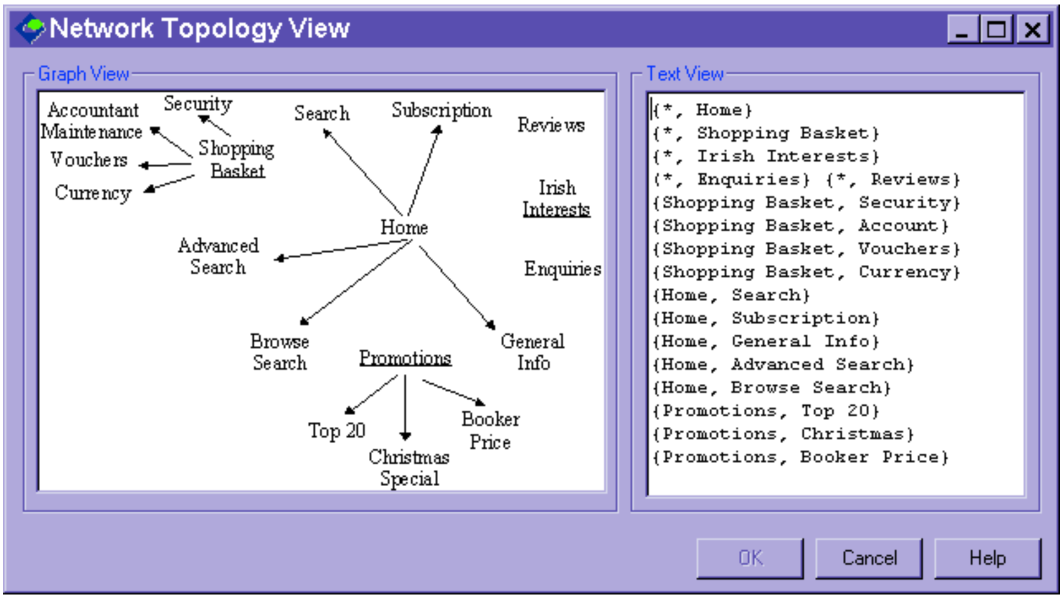
Exclude
* ; Offers/irish.html ; *
* ; Offers/poetry.html ; *
* ; Offers/reduced.html ; *

Model Options
☑ Discover Navigational Behaviour across Sessions
☐ Ignore Empty Referrer Entries

Figure 6. Using Simplified Navigation Templates in EasyMiner

3.2 Topology Networks

The second type of domain knowledge is that of network structures, which is useful when the topology of web site has to be represented, or only a sub-network of a large site is to be analysed. A network can theoretically be replaced by a set of navigation templates. However, navigation templates are of a more dynamic nature, whereas networks stay static over a longer period of time. An example network provided by the domain expert of one of the largest online bookstores in Ireland is shown in Figure 7, where an underlined word describes a page that can be reached from any other page on the site. The textual counterpart is also depicted in Figure 7, where an asterisk denotes the set of all pages.



```
{*, Home}
{*, Shopping Basket}
{*, Irish Interests}
{*, Enquiries} {*, Reviews}
{Shopping Basket, Security}
{Shopping Basket, Account}
{Shopping Basket, Vouchers}
{Shopping Basket, Currency}
{Home, Search}
{Home, Subscription}
{Home, General Info}
{Home, Advanced Search}
{Home, Browse Search}
{Promotions, Top 20}
{Promotions, Christmas}
{Promotions, Booker Price}
```

Figure 7. Example Network Topology

4 Discovering Internet Marketing Intelligence

Marketing experts divide the customer relationship life cycle into distinct steps, which cover attraction (acquisition), development, retention, relationship, and reward (Figure 8). In each of these life cycle stages the potential customer profile may be defined as shown (eg., in Relationship stage → Client).

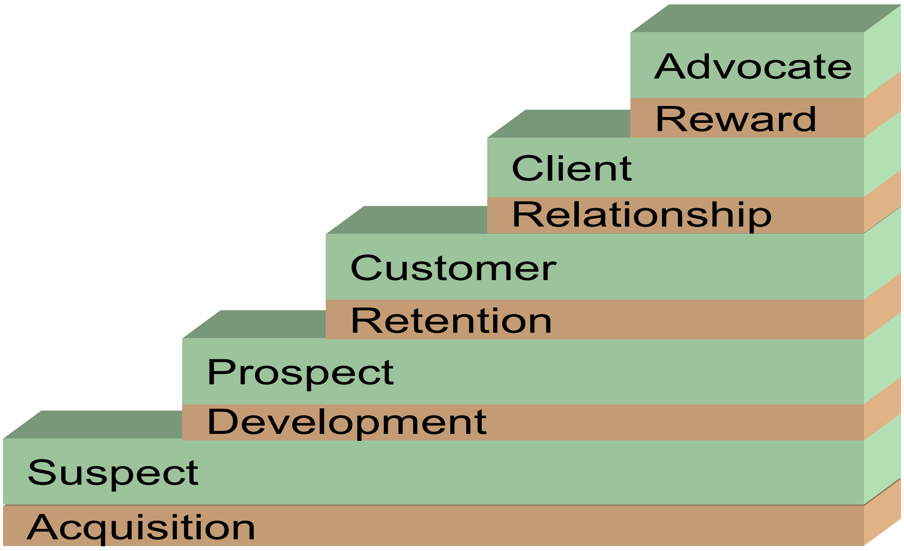


Figure 8. Customer Relationship Life Cycle

It has been recognised that mass marketing techniques are generally inappropriate for e-commerce scenarios. Direct marketing strategies, supported by knowledge discovery techniques are generally more successful (Ling & Li, 1998). In this section, a knowledge discovery scenario is presented for illustrative purposes from selected marketing life cycle stages, each of which defines a discovery goal, marketing strategy, and data mining approach

(Büchner, Mulvenna, Anand, Hughes, 1999a). The selected stages: are attraction (acquisition), retention, relationship, and reward

4.1 Customer Attraction

The two essential parts of attraction are the selection of new prospective customers and the acquisition of selected potential candidates. One possible marketing strategy to perform this exercise is to find common characteristics in already existing visitors' information and behaviour for the classes of profitable and non-profitable customers. These groups are then used as labels for a classifier to discover Internet marketing rules, which are applied online on site visitors. Depending on the outcome, a dynamically created page is displayed, whose contents depends on found associations between browser information and offered products / services.

The three classification labels used were 'no customer', that is browsers who have logged in, but did not purchase, 'visitor once' and 'visitor regular'. An example rule is as follows.

```
if Region = IRL and
   Domain1 IN [uk, ie] and
   Session > 320 Seconds
then VisitorRegular
Support = 6,4%; Confidence = 37,2%
```

This type of rule can then be used for further marketing actions such as displaying special offers to first time browsers from the two mentioned domains after they have spent a certain period of time on the shopping site.

4.2 Customer Retention

Customer retention is the step of managing the process of keeping the online shopper as loyal as possible. Due to the non-existence of physical distances between providers, this is an extremely challenging task in electronic commerce scenarios. One strategy is similar to that of acquisition, that is dynamically creating web offers based on associations. However, it has been proven more successful to consider associations across time, also known as sequential patterns. Typical sequences in electronic commerce data are representing navigational behaviour of shoppers in the forms of page visit series (Chen, Park, Yu, 1996).

Agrawal & Srikant (1995)'s a priori algorithm has been extended so it can handle duplicates in sequences, which is relevant to discover navigational behaviour. The M*i*DAS (Mining Internet Data for Associative Sequences) algorithm (Büchner, Baumgarten, Mulvenna, Anand, Hughes, 1999b) also supports domain knowledge as specified in Section 3. An example sequence is as follows.

```
{
ecom.infm.ulst.ac.uk/,
ecom.infm.ulst.ac.uk/News_Resources.html,
ecom.infm.ulst.ac.uk/Journals.html,
ecom.infm.ulst.ac.uk/,
ecom.infm.ulst.ac.uk/search.htm,
}
Support = 3.8%; Confidence = 31.0%
```

The discovered sequence can then be used to display special offers dynamically to keep a customer interested in the site, after a certain page sequence with a threshold support and / or confidence value has been visited.

4.3 Customer Relationship

The objective of relationship is to diversify selling activities horizontally and / or vertically to an existing customer base. It is normally executed by using cross-selling, right selling and up-selling techniques, which mean respectively, selling other products to the customer, offering (marketing/selling) the right product to the customer, and selling more products to the customer. The objective is to maximise the value of the customer to the merchandiser. In this short paper, we have adopted our traditional generic cross-sales methodology (Anand, Patrick, Hughes, Bell, 1998), in order to illustrate cross-selling in an electronic commerce environment.

To discover potential customers, characteristic rules of existing cross-sellers have to be discovered. This was carried out using attribute-orientated induction. For a scenario in which the product CD is actively being cross-sold to book sellers, an example rule is

```
if Product = book then
   Domain1 = uk and
   Domain2 = ac and
   Category = Tools
Support = 16.4%; Interest = 0.34
```

Deviation detection is used to calculate the interest measure and to filter out the less interesting rules. The entire set of discovered interesting rules is then used as the model to be applied at run-time on incoming actions and requests from existing customers.

4.4 Customer Reward

Rewarding customers who actively promote or implicitly support your e-commerce site through referral programmes (put in place by the e-commerce site) is a means of retaining those customers who are most loyal and who most frequently purchase high value items. These customers are sometimes labelled Most Valued Customers (MVCs). One reward mechanism that has been employed is to provide a lock-in strategy for MVCs. Essentially, this is a variation of the R-F-M model (Recency-Frequency-Monetary value of purchases). A RFM value is computed for each customer of the web site. Those customer who ranks highest is served with targeted promotions, for example, discount options on products. Those who rank lowest may not be offered any promotions.

4.5 Marketing Responses

Each of the stages of the marketing lifecycle has attendant marketing response strategies, which are shown in Figure 9. The marketing response is shown along with the data scope required to fulfil the response. For example, in order carry out loyalty marketing in the retention stage, up- cross- and right-selling techniques are used.

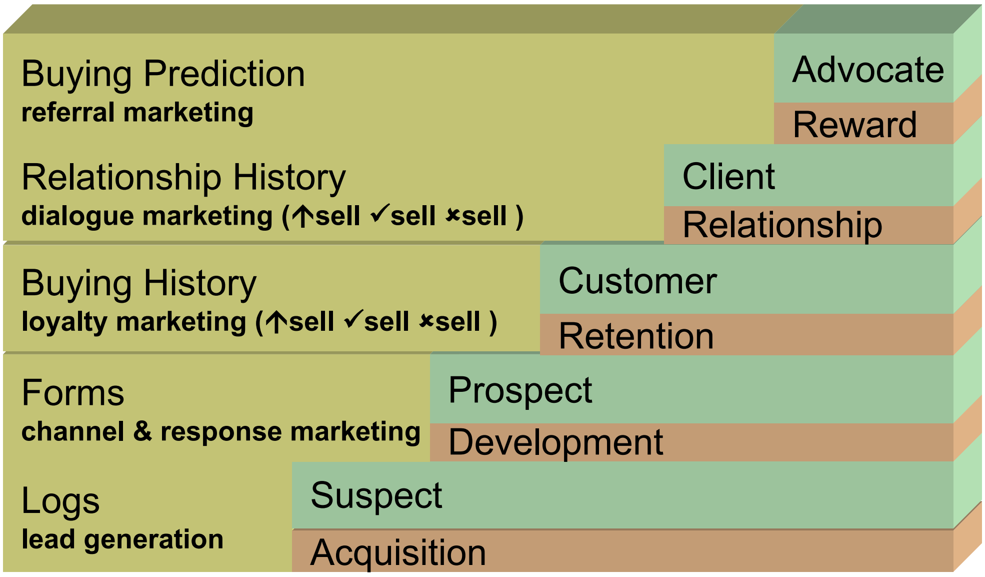


Figure 9. Marketing Responses

5 Deployment of Discovered Marketing Intelligence

In order to deploy discovered marketing intelligence, navigational behaviour - based on M*i*DAS/MKS, (Anand, Scotney, Tan, McClean, Bell, Hughes, Magill, 1997) - and predictive models based on sequences, classification, clustering and Bayesian techniques are represented in Predictive Modelling Markup Langauge (PMML) (Magnify, 1999) and made available to the web server through NSAPI, ISAPI and Apache http proxy agent filters. Figure 10 shows the web mining deployment schematic, where an online customer interacting with the web browser updates log files transparently by interacting with the web site. The personalised content is created dynamically based on retail data (product and service information, prices, order, etc.,) and existing domain knowledge.

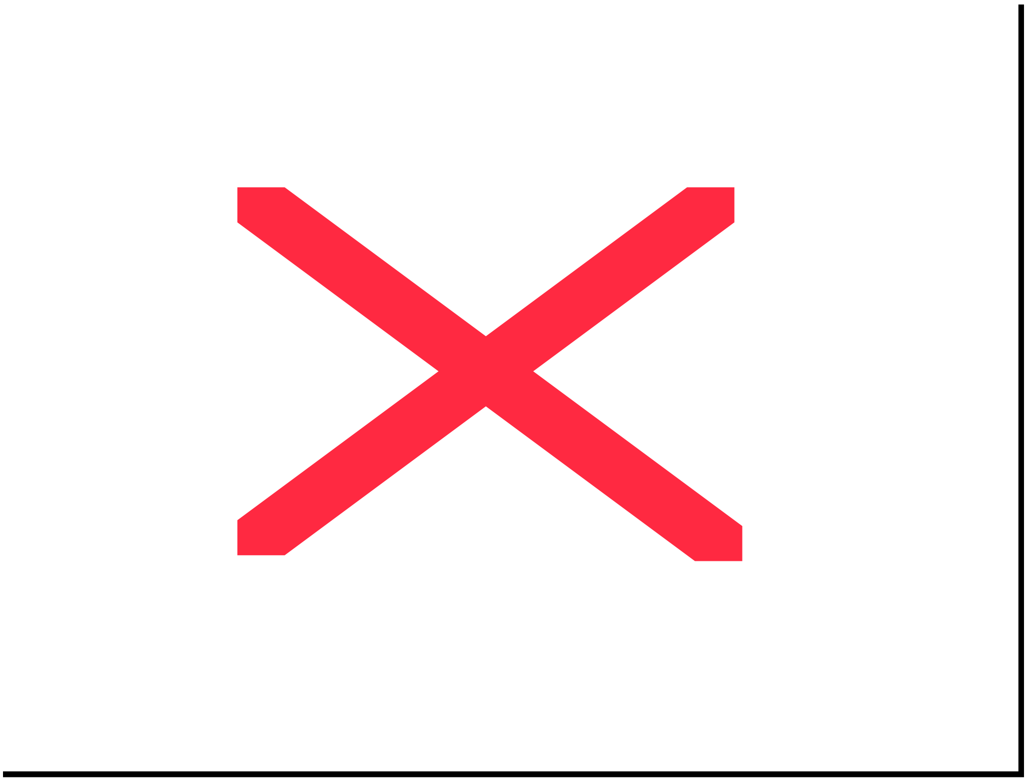


Figure 10. Web Mining Deployment Schematic

A project has been carried out with one of the biggest Irish online book shops, where currently about 2% of the overall sales are from Internet users. The objective was to establish the usability of existing customer, transactional and browsing data, in order to discover Internet marketing intelligence. Sequences, discovered by M*i*DAS, were employed as decision criteria for dynamically creating online promotions. A sample sequence containing 4 items is shown below, where two fields (HTTP_REFERER and URL) have been considered. The discovery was intended to find sequences, which show the success of the Christmas campaign. It shows that on that particular day, 16 people came from a banner advertisement in the business section of yahoo.co.uk, via the home page, to the special offer page, which led to an enquiry at a later stage. The coverage is 0.18% of chosen data set.[1]

[1] For reasons of confidentiality, more detailed information cannot be provided publicly.

```
HTTP_REFERER=http://www.yahoo.co.uk/Regional/Countries/Ireland/Business_and_Economy/Com
panies/Books/Shopping_and_Services/Booksellers/ | URL=/index.html |
URL=/searcher.phtml?area=christmas , URL=/searcher.phtml?area=enquiry (4, 16, 0.18%)
```

Figure 11. Sample M*i*DAS Sequence

The most interesting sequences (chosen by the domain expert) are now used in the creation of dynamic web pages customised for the current navigator. Data is presently being collected to measure the benefits of web log usage mining at the bookshop.

6 Related Work

Etzioni (1996) has suggested three types of web mining activities, viz. *resource discovery*, usually carried out by intelligent agents, *information extraction* from newly discovered pages, and *generalisation*. For the purpose of the discussion of related work only the latter category is considered, since it covers web log usage mining.

Zaïane et. al. (1998) have applied various traditional data mining techniques to Internet log files in order to find different types of patterns, which can be harnessed as electronic commerce decision support knowledge. The process involves a data cleansing and filtering stage (manipulation of date and time related fields, removal of futile entries, etc.) which is followed by a transformation step that reorganises log entries supported by meta data. The pre-processed data is then loaded into a data warehouse which has an *n*-dimensional web log cube as basis. From this cube, various standard OLAP techniques are applied, such as drill-down, roll-up, slicing, and dicing. Additionally, artificial intelligence and statistically-based data mining techniques are applied on the collected data which include characterisation, discrimination, association, regression, classification, and sequential patterns. The overall system is similar to ours in that it follows the same process. However, the approach is limited in several ways. Firstly, it only supports one data source — static log files —, which has proven insufficient for real-world electronic commerce exploitation. Secondly, no domain knowledge (marketing expertise) has been incorporated in the web mining exercise, which we see as an essential feature. And lastly, the approach is very data mining-biased, in that it re-uses existing techniques which have not been tailored towards electronic commerce purposes.

Cooley et. al. (1997) have built a similar, but more powerful architecture. It includes an intelligent cleansing (outlier elimination and removal of irrelevant values) and pre-processing (user and session identification, path completion, reverse DNA lookups, etc.) task of Internet log files, as well as the creation of data warehousing-like views (Cooley, et. al., 1999). In addition to (Zaïane et. al, 1998)'s approach, registration data, as well as transaction information is integrated in the materialised view. From this view, various data mining techniques can be applied; named are path analysis, associations, sequences, clustering and classification. These patterns can then be analysed using OLAP tools, visualisation mechanisms or knowledge engineering techniques. Although more electronic commerce-orientated, the approach shares some obstacles of (Zaïane et. al, 1998)'s endeavour, is mainly the non-incorporation of marketing expertise.

Spiliopoulou (1999) have developed a sequence discoverer for web data, which is similar to our M*i*DAS algorithm. Their GSM algorithm uses aggregated trees, which are generated from log files, in order to discover user-driven navigation patterns. The mechanism has been incorporated in a SQL-like query language (called MINT), which together form the key components of the Web Utilisation Analysis platform (Spiliopoulou, Faulstich & Winkler, 1999).

7 Conclusions and Future Work

We have presented the concepts and benefits of web log usage mining in the context of electronic commerce, which includes the pre-processing of online data, the incorporation of domain knowledge, as well as the discovery of marketing intelligence itself. The concepts have been incorporated in the authors' MIMIC architecture and results of carried out experiments have been presented.

Further work in the area of discovering marketing-driven navigation patterns is twofold. First concentrates on practical issues, which include horizontal and vertical diversification of digital behavioural data (such as Web/DITV and WAP devices) and a smoother interface to a web-enabled data warehouse. The second area of future work is the standardisation of the deployment engine, which utilises PMML representations of the mined knowledge, and deploys this knowledge through proxy mechanisms to realise recommendation and personalization.

8 References

Agrawal, R. & Srikant, R. (1995) Mining Sequential Patterns, *Proc. Int'l Conf. on Data Engineering*, pp. 3-14.

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

Anand, S. S., Patrick, A. R., Hughes, J. G., Bell, D. A., (1998) A Data Mining Methodology for Cross-Sales, *Knowledge-based Systems Journal* **10**: 449-461.

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

Büchner, A.G., Mulvenna, M.D., Anand, S.S. & Hughes, J.G. (1999a) An Internet-enabled Knowledge Discovery Process, *Proc. 9th Int'l. Database Conf.*

Büchner, A.G., Baumgarten, M., Mulvenna, M.D., Anand, S.S. & Hughes, J.G. (1999b) Navigation Pattern Discovery from Internet Data, *ACM Workshop on Web Usage Analysis and User Profiling (WebKDD'99)*

Chen, M.S., Park, J.S. & Yu, P.S. (1996) Data Mining for Traversal Patterns in a Web Environment, *Proc. 16th Intl'l Conf. on Distributed Computing Systems,* pp. 385-392

Cooley, R., Mobasher, R. & Srivastava, J. (1997) Web Mining: Information and Pattern Discovery on the World Wide Web, *Proc. 9th IEEE Int'l Conf. on Tools with Artificial Intelligence*

Cooley, R., Mobasher, R. & Srivastava, J. (1999) Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information Systems*, **1**(1).

Etzioni, O. (1996) The World-Wide Web: Quagmire or Gold Mine?, *Comm. of the ACM*, **39**(11):65-68

Ling, C.X. & Li, C. (1998) Data Mining for Direct Marketing: Problems and Solutions, *Proc. 4th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 73-79.

Magnify, Inc., Chicago, (1999) www.magnify.com/pmml

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

Spiliopoulou, M. (1999) The laborious way from data mining to web mining, *Int'l Journal of Computing Systems, Science & Engineering*, March

Spiliopoulou, M., Faulstich, L.C. & Winkler, K. (1999) A Data Miner Analysing the Navigational Behaviour of Web Users. *Proc. ACAI'99 Workshop on Machine Learning in User Modelling*, forthcoming

Srikant, R. & Agrawal, R. (1996) Mining Sequential Patterns: Generalizations and Performance Improvements, *Proc. 5th Int'l Conf on Extending Database Technology*, pp. 3-17.

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

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

World Wide Web Consortium (W3C), Logging Control In W3C httpd, (1997)

http://www.w3.org/XML/

Zaïane, O.R, Xin, M. & Han, J. (1998) Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, *Proc. Advances in Digital Libraries Conf.*, pp. 19-29.