Discovering Internet Marketing Intelligence through Web Log Mining*

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

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

phone: +44 (0)1232 368394 fax: +44 (0)1232 366068

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 high-value customers. Web log 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 an architecture 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 the organisations customers. The four customer-related key disciplines in marketing are attraction, retention, cross-sales, and departure. 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 mining, describes discrepancies between data and domain knowledge in traditional marketing and web log mining exercises, and outline 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.

* This research has partly been funded by the ESPRIT project Nº 26749 (MIMIC — Mining the Internet for Marketing IntelligenCe).

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.

2 Online Data Processing

2.1 Online Data Sources

The data available in electronic commerce environments is three-fold and includes server data in the form of log files, web meta data representing the structure of the web site, and marketing information, which depends on the products and services provided (see Figure 1 below and Büchner & Mulvenna, 1998).

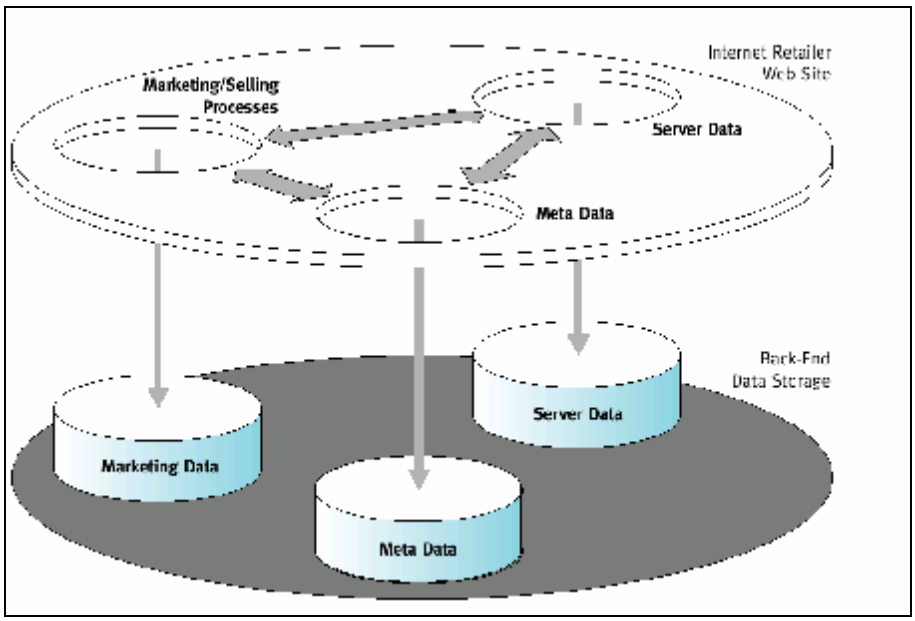


Figure 1. Internet Retailer Web Site Processes

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. There are three types of log files, namely server logs, error logs, and cookie logs. *Server logs* are either stored in the Common Logfile Format or the more recent Extended Logfile Format. The Extended Logfile Format supports additional directives that provide meta information about the log file, such as version, start and end date of session monitoring, as well as the fields which are being recorded in common log files. *Error logs* store data of failed requests, such as missing links, authentication failures, or timeout problems. Apart from detecting erroneous links or server capacity problems — which, when satisfactorily corrected, can be seen as a form of

customer satisfaction — the usage of error logs has proven to be rather limited for the discovery of actionable marketing intelligence. Cookies are tokens generated by the web server and held by the clients. The information stored in a *cookie log* helps to ameliorate the transactionless state of web server 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. 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 least 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, sociographic information, and so forth.

The last source is 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.

2.2 Online Data Preparation

In addition to 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 2.

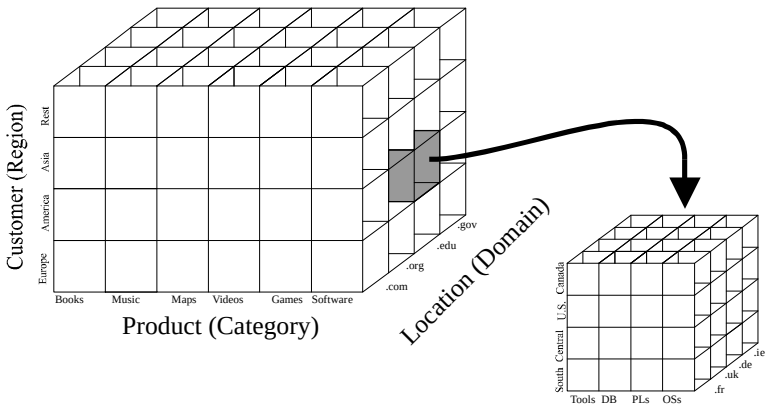


Figure 2. Web Log Data Cube

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 2), which can be used as input for data mining exercises.

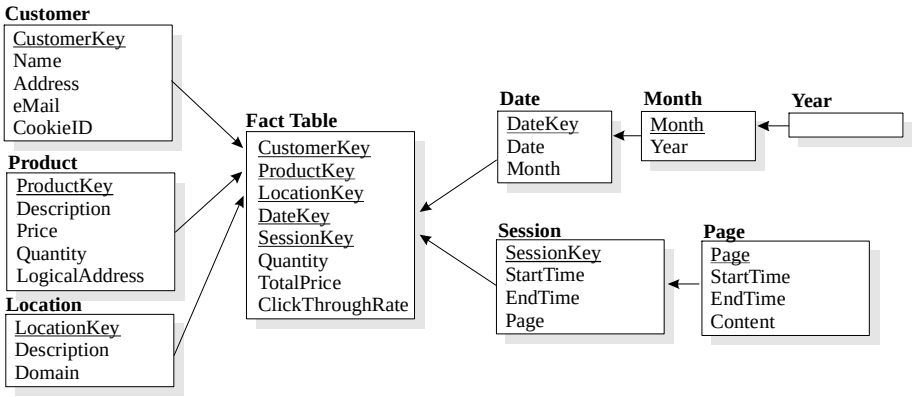


Figure 3. An Example Web Log Snowflake Schema

3 Domain Knowledge Incorporation

As in most knowledge discovery domains, there exist two types of domain knowledge that is relevant for web log 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 supported, 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 shall illustrate the concept of navigation templates. Imagine the analysis of a pre-Christmas marketing campaign within an online bookstore that introduced reduced gift items. The template is shown in Figure 4 below. 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 4 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 navigations by the same customer, resulting in a purchase. However, a navigation that includes "reduced.html", "junk.html" or "secondhand.html" are ignored in the analysis.

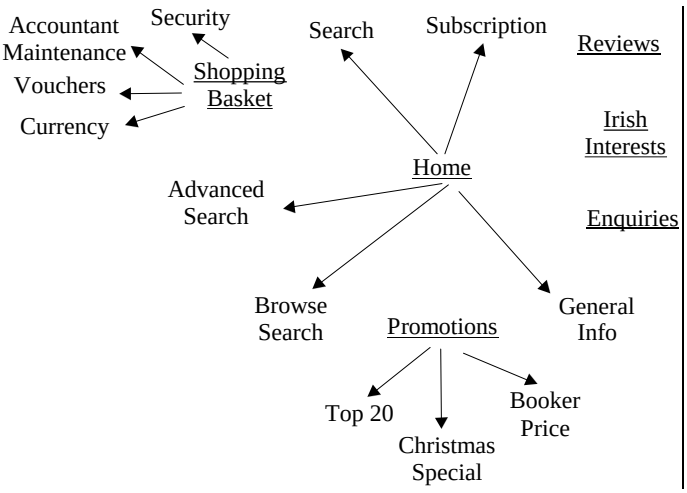
```
[
< index.html | * | offers/gifts.html ; * ; purchase.html | ? >
^< * ; offers/reduced.html ; * >
^< * ; offers/junk.html; * >
^< * ; offers/secondhand.html ; * >
]
```

Figure 4. Example Navigation Template

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 dealt with. 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 biggest online bookstores in Ireland is shown graphically in Figure 5(a), where an underlined

word describes a page that can be reached from any other page on the site. The textual counterpart in depicted Figure 5(b), where an asterisk connotes the set of all pages.



(a) Graphical Representation

```
{*, 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}
```

(b) Textual Representation

Figure 5. Example Network Topology

4 Discovering Internet Marketing Intelligence

Marketing experts divide the customer relationship life-cycle into four distinct steps, which cover attraction, retention, cross-sales, and departure. 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 all four marketing disciplines, each of which defines a discovery goal, marketing strategy, and data mining approach (Büchner *et al.*, 1999a).

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 *et al.*, 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 *et al.*, 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 Cross-Sales

The objective of cross-sales is to diversify selling activities horizontally and / or vertically to an existing customer base. We have adopted our traditional generic cross-sales methodology (Anand *et al.*, 1998), in order to perform the given task in an electronic commerce environment.

For discovering potential customers, characteristic rules of existing cross-sellers had to be discovered, which was performed through the application of attribute-orientated induction. For a scenario in which the product CD is 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 can then be used as the model to be applied at run-time on incoming actions and requests from existing customers.

4.4 Customer Departure

Customers who depart have either stopped purchasing a certain service or product and / or have moved to a competitor, which is also known as churn. The goal of customer departure prediction is to take action in order to prevent the exit (for instance, through a targeted promotion) or to prevent further costs in case the customer will leave, no matter what action will be taken.

Since a customer in an electronic commerce scenario does not explicitly leave, a user-defined delta value has to chosen as a threshold in which no activities have been recorded (neither browsing nor purchases). Log files from a certain period previous to the last activity have then to be analysed similarly to the customer retention scenario, that is sequences are discovered in order to find characteristics of churners. In parallel, classification exercises can be performed on the customer data in order to distinguish leavers from current customers. The types of patterns discovered are similar to the ones shown in sections 4.1 to 4.3 and are omitted for reasons of brevity.

5 Deployment of Discovered Marketing Intelligence

In order to deploy discovered marketing intelligence, navigational behaviour (discovered by M*i*DAS) is used to present the outlined concepts. M*i*DAS is a key component of the MIMIC (Mining the Internet for Marketing IntelligenCe) architecture built on top of the Mining Kernel System, which has been developed at the authors' laboratory (Anand *et al.*, 1997). MIMIC contains a data warehouse for storing web logs as well as marketing information, which provides multiple views of the stored data. The objective of MIMIC is to deliver marketing intelligence, which can be used for marketing activities, such as customer attraction, retention, cross-sales, and so forth.

Figure 6 shows the MIMIC architecture, where log files are being created by an online customer interacting with the web browser. The personalised content is created dynamically based on retail data (product and service information, prices, order, et cetera) and existing domain knowledge. This knowledge is incorporated by a marketing expert who is supported by navigational patterns from MiDAS.

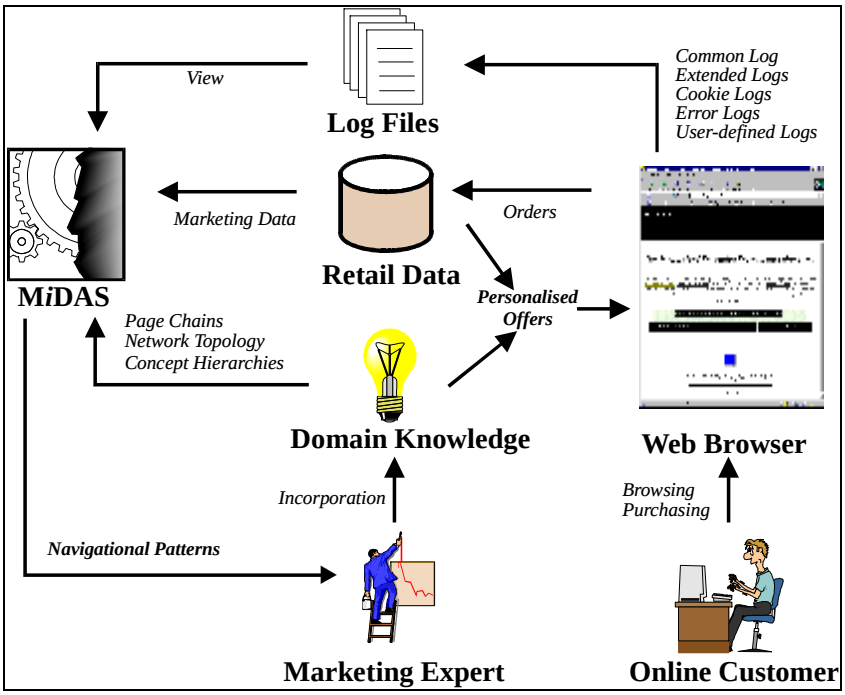


Figure 6. The MIMIC Architecture

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 MiDAS, 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/Companies/Books/Shopping_and_Services/Booksellers/ |
URL=/index.html | URL=/searcher.phtml?area=christmas ,
URL=/searcher.phtml?area=enquiry (4, 16, 0.18%)
```

Figure 7. 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 mining at the bookshop, however, as customer loyalty is intangible any such measurement is not going to be a true reflection of the overall benefits to the organisation.

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 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 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 TV and Internet channels) and a smoother interface to a web-enabled data warehouse. Second is concerned with the improvement of the algorithmic part, which includes the incorporation of more sophisticated types of domain knowledge (such as multi-level concept hierarchies) and tackling of performance issues.

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., A. R. Patrick, J. G. Hughes and D. A. Bell. 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. An Internet-enabled Knowledge Discovery Process, *Proc. 9th Int'l. Database Conf.*, forthcoming, 1999a.

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

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

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

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.

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

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

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

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.