Data Mining for Marketing Intelligence on the Internet

Maurice Mulvenna

MD.Mulvenna@ulst.ac.uk

Northern Ireland Knowledge Engineering Laboratory

University of Ulster

ESPRIT Electronic Commerce Project - MIMIC

Contributors:

- **Alex Büchner, Marian Norwood, Louis Moussy, Julian Clinton, Dave Jones, Paddy Ryan, Fergal Byrne**

Abstract

This paper outlines the sources of data available in on-line retail sites, and explores how Internet marketing may be enhanced by using data mining techniques to discover behavioural and access patterns in the data sources.

Data mining is the automated discovery of non-obvious, potentially useful and previously unknown information from large data sources. It use includes heuristic and artificial neural network techniques and induction algorithms to generate rules and associations that may be both useful *and* actionable. Within the context of relationship marketing data mining can provide knowledge about the unique characteristics of identified customer segments, so that business decisions may be made in relation to customer value and appropriate loyalty incentives can be developed.

The potential of data mining is enormous, but its market application may be tempered by customers and consumer organisations who may react negatively to the collection and 'mining' of aggregated personal information.

The implications are far reaching for Internet marketers, since data mining can improve their understanding of Internet consumer behaviour. It seems evident then, that Internet marketing activities will be characterised by sophisticated targeting of consumers. Ultimately, competitive advantage on the Internet may be determined by the ability of Internet marketers to collect and manage customer databases.

The paper concludes by describing the research objectives of MIMIC, a new ESPRIT research project funded under the Electronic Commerce thematic call. The MIMIC (Mining the Internet for Marketing IntelligenCe) project applies data mining techniques to Internet data.

Introduction

Electronic commerce has been described as “one of those rare cases where changing needs and new technologies come together to revolutionise the way in which business is conducted” (http://www.cordis.lu/esprit/src/ecomint.htm). Currently, relatively few consumers and businesses are connected to the Internet. Projections estimate over 50 million on-line by the year 2000, but is electronic commerce ready for the rigours of international retail trade?

There is now a plethora of on-line Internet shopping provider's world-wide, with thousands more appearing each month. Clients of the on-line retailers generate data at these sites by their interaction with the site in browsing and buying. Clearly these large data sets, which may be distributed and heterogeneous, will contain useful information helpful to business marketing strategies, both for retrospective analyses as well as data-driven forecasting.

The potential for this new form of marketing is enormous – browsers and buyers of products and services can be identified and targeted with attractive offers and sales promotions.

Sources of Data

The data generated by the main processes that occur at an Internet retailing web site may be described along the following dimensions:

1. Server data – data generated by the interactions between the persons browsing an individual site, and the web server. For example, the IP address of a computer browsing retailing sites.
2. Marketing data – the data stored by the Internet retailer on products, customers, suppliers, etc. For example, the consumer responses to discounting.
3. Site ‘meta’ data – the data about the site, usually generated dynamically and automatically after a site update. For example, location of product pages on a site, which are ‘leaf’ pages, navigational pages, etc.

Server Data

The httpd process that runs on web servers provides a facility to log information on accesses to the server (see Figure 1).

The W3C’s (www.w3.org) working draft on an Extended Logfile format is an attempt to map out an improved logging format in response to changing technical requirements (for example, proxy caches) and users’ needs, for example, demographic analysis.

The new format is similar to the current common logfile format in that data is stored as ASCII character sequences - essentially a ‘flat file’ format. The following additions have been made with the Extended Logfile format. The line by line format now includes *directives* as well as normal *entries*. The W3C draft describes the following directives:

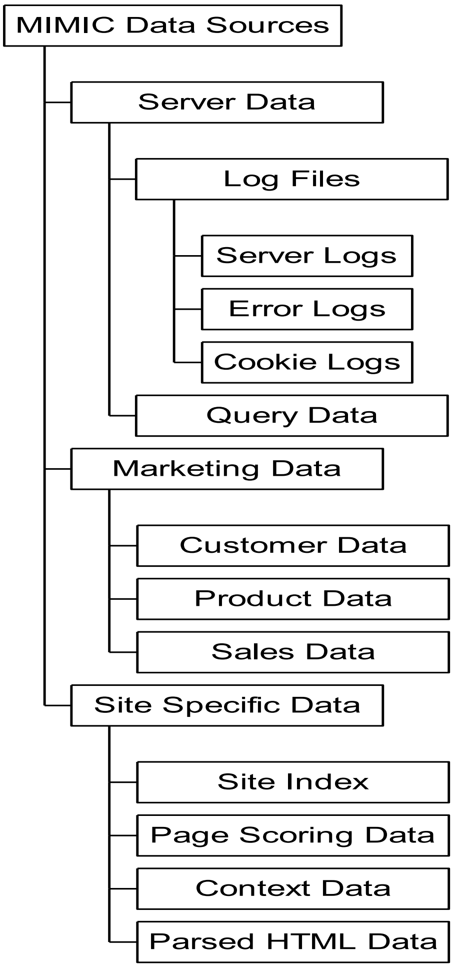

Figure 1. Sources available for Data Mining in On-line Retailing

Directive	**Data type**	**Comment**
Version	*<integer>.<integer>*	Version of extended logfile format used
Fields	[<*specifier*>*…*]	Specifies fields recorded in log
Software	*String*	Identifies the software which generated the log
Start-Date	*<date> <time>*	Date and time at which log was started
End-Date	*<date> <time>*	Date and time at which log was finished
Date	*<date> <time>*	Date and time at which entry was added
Remark	*<text>*	Comment information

Table 1. Directives defined in W3C ‘Extended Logfile Format’ Specification’

The version and field directives are mandatory. Essentially the Extended Logfile format uses new field identifiers to enable the capture of additional information.

Additionally, summarisation of the Extended Logfile format is now possible including sampling and selection. For example, selection of entries with a URI or URI stem portion:

www.ulst.ac.uk or `/pages/products/books`

The following fields are included in the Extended Logfile Format:

Date	Time	Client IP Address
User Name	Service Name	Server Name
Serve IP	Server Port	Method
URI Stem	URI Query	Http Status
Win32 Status	Bytes Sent	Bytes Received
Time Taken	Protocol Version	User Agent
Cookie	Referrer	

Table 2. Fields included in the Extended Logfile Format

Marketing Data

Any organisation that uses the Internet to trade in services and products uses some form of information system to operate Internet retailing. Usually, such a database will contain information on customers, products and sales, including transactional information.

Site Specific data

The third data source is data about a web site. This includes data about the architecture of a site, and the contents of HTML pages, whether static or generated dynamically. This is ‘meta data’ – data about the data making up a web site. There are many ways in which an Internet retailer can gather information on their web site by including information on context (for example, page scoring) or by dynamically creating and updating a database of the topology of the Internet site.

Marketing & The Internet

(Dibb; Simkin; Pride; Ferrell 1997) define marketing as consisting of “individuals and organisational activities that facilitate and expedite satisfying exchange relationships in a dynamic environment through the creation, distribution, promotion and pricing of goods, services and ideas”. This definition highlights the need for the development of the ‘right’ marketing mix (product, price, promotion, distribution, etc.). According to (Kotler 1997) this is “a set of marketing tools that the firm uses to pursue its marketing objectives in the target market”. The development of the ‘right’ mix is governed by an understanding of customers’ needs and wants.

Normally, all marketing communications are subject to measurement and control in order to monitor the success or otherwise of campaigns. Traditionally effectiveness is measured by tracking awareness and image, by measuring response rate to all direct communications and of course by looking at sales volume. It involves a complex series of tasks. This is not necessarily so for Internet marketing. Organisations that market on the Internet can now measure directly the response generated by the campaign, advertisement, discounting, etc. In the case of online 'banner ads', the success is measured by the 'click through' rate. This is the number of times that, when someone is presented with an ad, they then choose to click on that ad to obtain further information. A company hosting its own merchant server can measure the number of 'hits', and break down this information by domain (e.g., .com for commercial).

The above examples illustrate that each interaction between consumer and seller is recorded in digital form. This is the important difference between conventional selling and selling on the Internet. At this stage in the Information Age, each actor in the information cyberspace, for example, consumers and retailers in Internet retailing, generates digital trails and data that may be stored and analysed by the organisation operating the web server. Data mining algorithms are the new actors in this information cyberspace.

Data Mining

Data mining has been defined as "the automated discovery of non-obvious, potentially useful and previously unknown information from large data sources" (Frawley; Piatetsky-Shapiro; Mattheus 1991). It applies various artificial intelligence techniques to discover patterns that may be both useful and actionable.

Data mining software consists of algorithms that can discover *associations* between data, help to *classify* items, and discover *sequences* hidden in the data (Anand; Büchner 1998). An example of a discovered association could be the types of products that consumers usually buy at the one time. An example of classification would be the grouping of a particular type of buyer, such as 'high frequency but low value' purchaser. An example of discovered sequences could be a buyer periodically returning to purchase other novels by a particular author.

Within the context of data analysis of Internet retailing sites, data mining can discover knowledge about the unique characteristics of identified customer segments, so that business decisions may be made in relation to predictions about customer value and appropriate loyalty incentives can be developed (Mulvenna; Büchner 1997).

Data mining allows marketers to reveal layers of information about markets (or subsets of markets) in ever increasing detail, enabling customer or prospect profiles to be built and the identification of the segments upon which marketing activities can focus.

So what are the benefits of using data mining techniques to gather intelligence about a market? It allows on-line retailers to 'fine tune' their selling strategy. This gives them a greatly enhanced insight into the number and types of lines in stock, the best electronic shop front format, and which offers should be directed at which customers and how should they be communicated (Humbry 1996). Using direct marketing, heavy buyers of a product can be identified and targeted with attractive offers and sales promotions. Targeting does not have to involve discounted offers alone. It could mean making certain categories of customer aware of products or services that might be of interest to them, or inviting them to special on-line events.

The application of data mining techniques to the data of on-line retailers provides high-level knowledge – for example, in the form of rules - that describes consumer navigational and purchasing behaviour. These rules capture trends and behaviour patterns that may be applied within a marketing strategy. The authors propose that the high-level, descriptive, behavioural rules supply the marketing specialists with the means to feedback directly the aggregated behavioural responses of the clients of an online service. The high-level mined rules may be incorporated into the architecture of an on-line retailing system. When each person interacts with the on-line service to navigate and purchase goods or services, their on-line behavioural patterns are identified, and the on-line retailing system may react to change dynamically the information presented to that consumer. This is the goal of MIMIC.

The MIMIC Project

MIMIC ("Mining the Internet for Marketing IntelligenCe") is a project funded by ESPRIT IV Framework Thematic Programme for Electronic Commerce (MIMIC 1997). The MIMIC Consortium consists of four partners, represented by the University of Ulster, and three small-to-medium enterprises (SME).

The technical objective of the MIMIC project is to develop a data mining toolkit that will make possible the mining of the data generated by online retailers (Mulvenna; Büchner; Norwood; Grant 1997). The business objective is to provide European on-line shopping retailers (of which many are SME companies) with advanced technology that will enable them to maintain their competitiveness in the global Internet marketplace. The approach of the MIMIC project is to capture the requirements from SME on-line shopping providers, and develop specialised Internet-capable data mining algorithms. The three areas of data mining outlined in the previous section – classification, association and sequences – will require additions and specialised extensions to cope with the specific nature of Internet data. These will be validated by the users and incorporated ultimately into a special version of existing data mining software.

The target market segment is on-line shopping retailers, and the goal of the MIMIC project is to give these retailers the ability to market their products and services effectively. Such customised and directed marketing communication, however, relies on accurate information about the customer. MIMIC will provide, for the first time, the capability for retailers to consider a cost-effective means of customised data-driven marketing communication.

Although it is widely believed that mass marketing techniques such as advertising are a major proportion of a marketing budget, many companies now spend more on sales promotion methods, which include direct marketing techniques. The benefits of data-driven marketing are the ability to measure directly the response to the marketing effort, and the resultant cost-effectiveness of this technique. Data-driven marketing also enables on-line retailers to monitor their customers' behaviour patterns and to detect and entrap potential 'switchers' – those people who show no allegiance to a particular retailer. Electronic commerce dispenses with geographic boundaries, and empowers customers to shop where they obtain the cheapest/best bargain. The marketing intelligence information from the MIMIC data mining toolkit may be used to provide a competitive edge for on-line shopping retailers, allowing them to capture, retain and satisfy customers.

Conclusions

This paper has outlined how data mining may be applied to assist the marketing communication process for Internet retailers. The algorithms used in data mining are capable (with modification) of processing of large, distributed and often heterogeneous databases that store the transactional, behavioural and other data gathered by Internet retailers.

The benefit of data mining to on-line retailing is clear, but its market application may be tempered by customers and consumer organisations who may react negatively to the collection and 'mining' of aggregated or personal information (Toronto Star 1998). It is impossible to predict the future landscape of the Internet and the roles of the actors in this information cyberspace. However, it is becoming apparent that consumers value their rights and will only participate in marketing relations that benefit them as much as the retailer (Norwood; Mulvenna; Büchner; Grant 1997; Easton; Parker 1998). The onus is therefore on the retailers to provide assistance and engender trust. This may be done positively, as a value adding action (Thirkell 1997).

Internet retailers who do not attempt to address the trust issue directly face low turnover at best and vigorous collective reactions from consumers at worst. Ultimately, consumers and other actors in cyberspace will seek out lateral solutions to what they view as privacy intrusions (Hagel III; Rayport 1997). These may include a desire for anonymous transactions; for example, using digital cash equivalents.

Data mining algorithms and techniques may be applied to more than Internet retailing data. Other areas that involve Internet transactions include business-to-business electronic commerce. If the electronic commerce market grows as some commentators predict, there will be an explosion in the growth of digital transactional and behavioural data, and an increased demand for data mining tools and techniques.

The results of the MIMIC project should demonstrate the usefulness of data mining techniques and explore - to retailers and consumers alike - the implications of the application of these techniques.

The ultimate goal of MIMIC is to prepare the pre-competitive results of our work for commercial exploitation across Europe and world-wide. Additionally, four commercial users, who are potential customers of the output of the project, will pilot the MIMIC prototype. We already have received expressions of interest from Gourmet food producers, French wine chateaux, and book sellers.

References

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

Dibb, L. Simkin, W.M. Pride, O.C. Ferrell: Marketing Concepts and Strategies, 3rd European Edition, Houghton Mifflin, Boston, New York, 1997.

Easton, J., Parker, J., "Seeking the *There* There: Building Relationships is the Key to Building Profits", ZD Internet Magazine, February 1998

Frawley, W.J., Piatetsky-Shapiro G. and Mattheus, C.J., "An Overview: Knowledge Discovery in Databases", in Knowledge Discovery in Databases, AAAI/MIT Press, 1991, pp. 1-27

Hagel III, J., Rayport, J.F., "The Coming Battle for Customer Information", Harvard Business Review, January-February 1997

Humbry, C., Digging for Information, Marketing, November 21, pp.41-42, 1996

Kotler: Marketing Management: Analysis, Planning, Implementation and Control, 9th Edition, Prentice Hall International Inc, pp. 466-467, 1997.

Mining the Internet for Marketing IntelligenCe – MIMIC, Esprit Proposal N°:26749, 1997

Mulvenna, M.D., Büchner, A.G., "Data Mining and Electronic Commerce, in Overcoming Barriers to Electronic Commerce", (OBEC '97), Malaga, Spain,1997

Mulvenna, M.D., Büchner, A.G., Norwood, M.T., Grant, C., "The 'Soft-Push': Mining Internet Data for Marketing Intelligence", In: Working Conference: Electronic Commerce in the Framework of Mediterranean Countries Development, Ioannina, Greece, October 1997

Norwood, M.T., Mulvenna, M.D., Büchner, A.G., Grant, C., "Competitive Advantage on the Internet Through Relationship Marketing", 1st Berlin Internet Economics, October 1997

Thirkell, P.C., "Caught by the Web: Implications of Internet Technologies for the Evolving Relationship Marketing Paradigm", Conference of American Marketing Association, Dublin, Ireland, June 1997

“Data Mining’ Privacy Threat”, Toronto Star, Toronto, Canada, 14 January 1998