Data Mining & Electronic Commerce

Maurice D. Mulvenna, Alex G. Büchner

Northern Ireland Knowledge Engineering Laboratory
University of Ulster, Shore Road, Newtownabbey, Co. Antrim
Northern Ireland, BT37 0QB
UK
{md.mulvenna, ag.buchner}@ulst.ac.uk

Abstract

Many users of the Internet are aware that each time they connect to an on-line shopping server, they leave behind a 'footprint' in the site's server logs. The information contained in this footprint is innocuous, but it can be 'mined'. Data mining is the 'non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data'. This paper outlines the types of data available for mining, describes operation and limitations of the mining algorithms, and discusses the marketing and ethical issues that arise.

1 Types of Electronic Commerce

Electronic Commerce (EC) is defined as "any activity that utilises some form of electronic communication in the inventory, exchange, advertisement, distribution or payment of goods and services" ([Cha96]). This is a broad definition and indicates that EC encompasses much more than its predecessor Electronic Data Interchange (EDI). Although still in its infancy, the rapid growth in the use of EC is recognised as being fuelled by Internet technology, in particular the widespread international adoption of the networking standard TCP/IP, World Wide Web (WWW) browsers and the HyperText Transfer Protocol (HTTP). These are the *de facto* standards facilitating Electronic Commerce.

Figure 1 enumerates the possible interaction paths between the main actors in Electronic Commerce: companies; suppliers; customers; and regulators. Of course, many of these roles are interchangeable; for example, suppliers are normally also companies. In the figure, there are many possible permutations of interactions. Only the most important at this stage in the evolution of EC will be considered in following sections, namely company-to-company and compnay-to-customer.

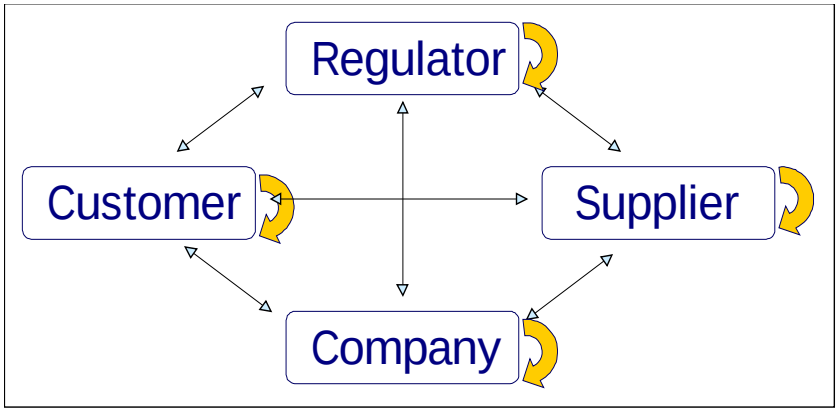

Figure 1. Electronic Commerce: Inter-Organisational Communication Paths

The curved arrows indicate that interactions can be between similar organisations. For example, regulators may exchange information with each other. A planning authority may exchange information with a regional tourist authority to ensure that planning applications for new housing comply with regional aesthetics.

1.1 Company-to-Company

The company-to-company interaction subsumes the concept of supply chains. Many organisations which are a part of a supply chain already utilise some form of EDI. For example, in Northern Ireland, the clothing company Desmonds, supplies exclusively to Marks & Spencers in England. The two companies are so closely coupled that invoicing is not required, and the supplying company's stock levels are made available electronically to the retail company.

The impact of EC can be amplified by the re-engineering of the processes in the participating companies. This enables the companies to identify value-chains and then add further value. Davenport ([Dav90]) has identified the five major steps to be addressed in re-engineering. These are to develop business vision and process objectives; identify processes to be redesigned; understand and measure existing processes; identify IT levers; and design and build a prototype of the process (Figure 2).

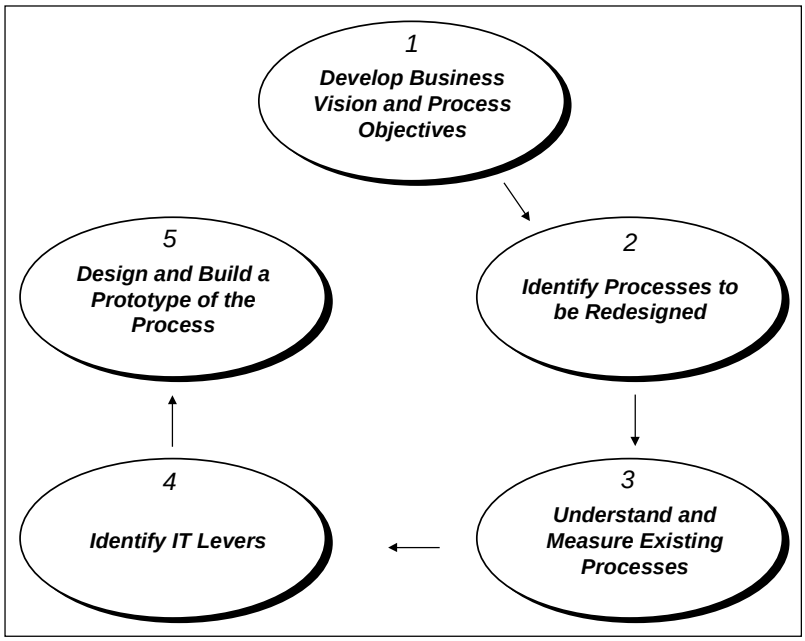

Figure 2: Five Steps in Re-engineering

The vision and objectives in the first of these steps is the balancing of factors such as cost and time reduction, quality of output of process, improvements in worklife, and the empowerment of employees through process buy-in. In the second step there are alternate strategies that may be used. The most popular of these is to focus on those processes that can have the highest impact on an organisation. Another option is to use techniques that identify those processes that may be most data-rich, and can perhaps benefit from the application of IT-enabled BPR. Measurement of process is the third vital step. If there is no way to gauge how a new re-engineered process has impacted an organisation, either financially or in other more indirect ways, then it may be difficult to champion the changes that BPR may bring. The fourth step can, if directed correctly, lead to massive improvements in how a company carries out its business. IT is central to the success of a great majority of BPR operations, and with technologies such as the Intranet, OLAP and knowledge based systems available, quantum

leaps can be made in addressing major corporate problem areas. The fifth step is important because it provides the organisation with the visible results of a BPR exercise. It also enables the BPR team to develop a more generic model for further BPR application.

The re-engineering of organisations is frequently a necessary prerequisite to successful implementation of Electronic Commerce technology. It is often required in order to disassemble the 'functional silos' that exist after an organisation has used and grown accustomed to EDI over time ([Mar96]).

1.2 Company-to-Customer

Much of the excitement of EC is focused on the delivery of products and tradable services to consumers in their own homes. This is usually known as 'on-line shopping' and is seen as the area of Electronic Commerce where many of the following issues will be debated first: standards; international trade agreements; security; and trust.

However, the interaction between customer and company also includes the area of on-line banking and publishing. Many of the companies traditionally involved in these areas are investing heavily in new technology and business pilots in order to assess the likely future market. For example, banks are aware that organisations such as Microsoft that have not ventured into the retail financial domain previously, are keen to position themselves as 'Trusted Third Parties' (TTP). Initially, these newcomers may restrict their operations as TTPs to the verification of digital signatures, etc. However, some TTP organisations already offer some form of software 'credits' in return for the customer carrying out a service for them. These credits may soon turn into dollar fractions, and as soon as that happens, the distinction between traditional banks and TTP organisations will become blurred. Recently, Digital announced Millicent, a patent-pending micro-payment system for the Internet ([Keh97]).

On-line shopping is normally hosted on a WWW Internet server. Customers connect to the server using their own PC, and interact with the system. The interaction usually takes the form of browsing and searching for products. Once a product is located, and the customer decides to buy, the item can be added to a 'shopping trolley' or other supermarket metaphor. Payment options are numerous, and the security issues are outside the scope of this paper, but a common option is to pay using an internationally-recognised credit card using secure HTTP (S-HTTP). In order to verify the cardholder information with the credit card company and to obtain an address for delivery of the items, a WWW on-line site will often request further details such as address. Frequently, additional information will be captured at this stage, . for example age, sex, interests, likes and dislikes, etc.

2 Data Sources in On-Line Shopping

2.1 Server Logs

On-line shopping servers produce several network protocols which have the potential to be mined: The *Common Log Format (CLF)* provides information about physical connections in the form 'host ident authuser date request status byte'. The *Custom Log Format* provides logs about logical connections, as well as software and hardware profile of the user. Frequently, the information on the originator will be in the form of an IP number. Using reverse DNS, it is possible to obtain the full domain name which can then be parsed. For example, the domain name www.rasta.ac.jp can be resolved to provide information that identifies the originator as an academic from Japan. However it must be recognised that not

all domain names can be resolved. *Error protocols* record all occurring faults while being connected to an on-line shopping server.

2.2 Cookie Logs

HTTP is a 'transactionless' protocol, i.e. each interaction with an Internet server is independent from any that precede or follow it. This causes problems with many aspects of EC, where transaction processing is required. For example, the shopping trolleys in on-line malls need to be able to store your intended purchases, even if you leave the site and re-visit several days or weeks later. This problem is overcome by the use of 'cookies'. These are software components introduced by Netscape[1] which can store information about a client's access to a server, on the client's computer. Cookies are normally used to store state information like the contents of a shopping trolley, or the pages accessed when a client last connected to an on-line mall. Increasingly, Internet server software[2] contains extensions that enable the storage of information about cookies, which are called *cookie logs*. Cookie logs contain generic information in the form 'name expiry_date, path domain and security_level', which can be customised depending on the applied domain.

2.3 Customer Information

The information that is captured using on-screen forms to enable secure credit authorisation is a rich source of additional data about the customer. This process is usually preceded by user identification, which can either be performed at the first usage of the on-line service, or at the first product purchase. The type of requested data depends on the type of the on-line shopping business and the nature of the usage of the data.

2.4 Miscellaneous Sources

2.4.1 Demographic Data

It is common practise to obtain demographic data from third party suppliers. These data has usually been collected over decades and provides valuable information for strategic decisions. Database marketing has become a very lucrative field, which offers such data, often tailored for specific requirements.

2.4.2 Internet-sourced

A novel data source is the Internet itself, which has been exploited by various parties. Information about users is recorded and offered to interested customers, either through intelligent agents (and equivalents) or cookies, An example of a company which provides such information is DoubleClick[3].

3 Data Mining

3.1 Introduction

Data Mining is the term given to the automated discovery of non-obvious, potentially useful and previously unknown information from large data sources. It enables industry in different sectors to utilise their most under-utilised resource i.e. data collected by them about various aspects of their businesses, by automatically discovering patterns in their processes which can allow them to gain a competitive edge over their competitors.

---

[1] http://home.netscape.com/newsref/std/cookie_spec.html
[2] For example, Apache Internet server
[3] http://www.doubleclick.com

3.2 Types of Patterns

Different types of patterns can be detected from given data. The most relevant types for Electronic Commerce are

- Association rules (relations and dependencies among fields in the data) for basket analysis on the on-line shopping mall,
- Classification rules (mapping data items into one or several predefined categorical classes) for finding potential customers for a certain product or classifying visitors' behaviours,
- Characteristic rules (discovering specifics of one data item) for tackling cross-sales problems,
- Sequential rules (modelling states and patterns of a process) to detect typical paths shoppers tend to chose, and
- Clustering (finding groups of similar entities) to find similar paths of visitors which lead to the same product interest or purchase.

3.3 The Data Mining Process

Data Mining is recognised to be a process, rather than a stand alone automated algorithm which discovers knowledge from data without human interference. A Data Mining process incorporates data, domain and Data Mining expertise (see Figure 3).

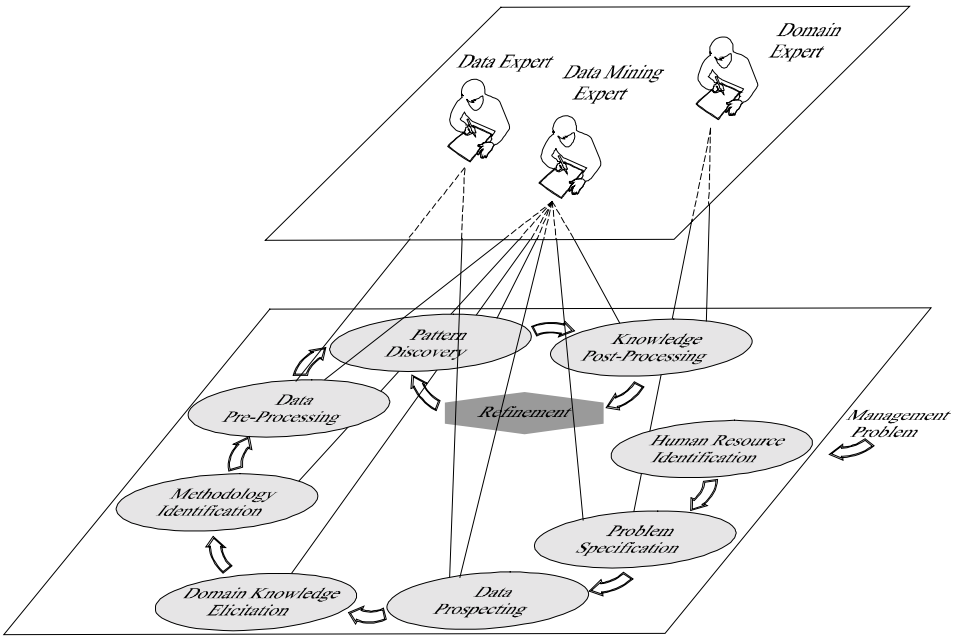

Figure 3: The Data Mining Process

The process is covering the following steps: *human resource identification* is concerned about finding appropriate expertise. In the *problem specification* phase various components of the business solution required as an output from the Data Mining process are identified. The *data prospecting* stage ensures that all required data (quantitatively and qualitatively) is available and accessible. The *domain knowledge elicitation* incorporates existing knowledge about the tackled problem in the Data Mining system. The *methodology identification* identifies the methodology utilised to tackle the problem most appropriately. *Pattern discovery* is what is often referred to as the most important Data Mining step, in which the actual knowledge is discovered. *Knowledge post-processing* is concerned about modifying discovered knowledge so it can be facilitated by the user, e.g. produce natural language like rules. These last two steps usually form a refinement process which has to be iterated through until sufficient results have been achieved.

4 Applying the Data Mining Process to Internet Data Sources

The potential for so called Web-Mining has been recognised by various parties (e.g. [Che97], [Etz96] or Mac96]), but all approaches only use server logs as input data. However, as outlined in Section 2, more data is available from on-line shopping server logs, which can be utilised in the Data Mining process:

Typical *human resources* in an EC context are a WWW administrator (data expert), a marketing manager of the on-line shopping complex (domain expert) and a Data Mining provider (Data Mining expert). The *problem specification* depends on the nature of the on-line shopping mall and the problem to be tackled. An example is the detection of potential customers for a newly launched product. *Data Prospecting* is concerned about the state of the different logs, i.e. missing or false values and their accessibility. Available *domain knowledge* can be either available information about the problem being tackled or about the structure of the on-line shopping mall, e.g. the logical topology of an Internet server can be modelled as a semantic network and incorporated as domain knowledge. The *methodology* being chosen depends on the problem at hand and the data and knowledge available, as outlined in Section 3. *Data Pre-processing* includes filtering out irrelevant information, e.g. a user's name when navigation trends have to be discovered, solving semantic heterogeneity among incompatible logs and filling in missing values. The *pattern discovery* will then find patterns, based on the methodology identified, the data being pre-processed and the domain knowledge being elicited. *Knowledge post-processing* can either be the transformation of the discovered patterns to a format which is understandable for humans or applicable for the system which generates user and task sensitive WWW pages dynamically.

5 Conclusions

It has been shown what information is available from Internet logs and how this data can be harnessed by Data Mining techniques.

One aspect which has not been mentioned yet, but has raised major concerns in the Electronic Commerce community, are legal and ethical aspects ([Kal97]). Where are the boundaries (legally as well as ethically) of what information can be recorded? What kind of knowledge can be justified to be discovered? Who is taking control over that process? etc. As in the non-virtual world, people like anonymity when shopping, especially if private or business sensitive information is involved. Thus, those questions have to be answered satisfactory, before Data Mining technology is coupled to on-line shopping systems.

In addition to technical aspects of Web Mining, also marketing issues have to be tackled. The potential knowledge being discovered can be used as important criteria for further marketing strategies. The conclusions of those modified strategies can then be realised in on-line shopping systems by dynamically creating user and task sensitive pages. For instance, a user who has been identified as a potential customer for a particular product range – based on his or her previous records applied to knowledge discovered from existing data – can be offered those items in that product range, which are reduced at the minute.

6 References

[Ana96] S.S. Anand, A.G. Büchner, J.G. Hughes, D.A. Bell: Towards Real World Data Mining, *Proc. of the Workshop on Data Mining in Real World Databases at the 1st Int. Conf. on Practical Aspects of Knowledge Management*, Vol. 1, October 1996.

[Cha96] C.A. Charles, C.P. Foss, S. Dewan (Eds.): Globalizing Electronic Commerce, Report on the Int. Forum on Electronic Commerce, Beijing,

China, 20-21 March 1996, Center for Strategic & Intl Studies.

[Che97] D.W. Cheung, B. Kao, J. Lee: Discovering user Access Patterns on the World-Wide Web, in *Proc. 1st Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pp. 303-316, 1997.

[Dav90] T.H. Davenport, J.E. Short, The New Industrial Engineering: Information Technology and Business Process Redesign, in *Sloan Management Review*, Vol. 11, 1990.

[Etz96] O. Etzioni: The World-Wide Web: Quagmire or Gold Mine?, in *Communications of the ACM*, 39(11):65-68, 1996.

[Kal97] R. Kalakota, A.B. Whinston: Electronic Commerce: A Manager's Guide, Addison-Wesley, 1997.

[Keh97] L. Kehoe, L., Start Talking Cents, in *Financial Times*, London, 12 March 1997.

[Mac96] J. Mace,: Internet Usage Analysis: A detailed Study of an Electronic Commerce Web-Site, in *Proc. of the Workshop on Data Mining in Real World Databases at the 1st Int. Conf. on Practical Aspects of Knowledge Management*, Vol. 1, October 1996.

[Mar96] J. Martin, Cybercorp: the New Business Revolution, AMACOM, NY, 1996