The 'Soft-Push': Mining Internet Data for Marketing Intelligence

Maurice Mulvenna[*], Alex Büchner[*], Marian Norwood[‡], Caroline Grant[*]

[*]Northern Ireland Knowledge Engineering Laboratory
[‡]School of Commerce and International Business Studies

University of Ulster, Shore Road, Newtownabbey, Co. Antrim
Northern Ireland, BT37 0QB
UK
{md.mulvenna, ag.buchner, mt.norwood, c.grant}@ulst.ac.uk

Abstract

This paper examines the challenges that on-line shopping and other commercial transactions on the Internet pose for marketers in the retail industry. The Internet constitutes a whole new marketplace in its own right, and evidence exists that the traditional manipulation of the marketing mix has to be modified for this new environment. To place this paper in context, the authors describe the potential growth of electronic commerce on the Internet, how traditional marketing strategies can be adapted for the Internet and current Internet marketing techniques. This paper highlights how marketing professionals and retailers can exploit the tools of the Internet to move closer to their customers and add value to their products. It outlines the potential for applying data mining technology to the data that is collected as consumers browse and purchase goods and services in on-line shopping malls. The authors introduce a variation of the push promotional strategy - 'the soft-push' - a sales promotion strategy based on the navigational and purchasing behaviour of on-line shoppers. Data mining allows marketers to reveal customer profiles, helping to identify appropriate market segments. Different data algorithms are described and the data mining process is explained. A comparison is made between the 'soft-push' approach and other web mining approaches. In conclusion, key ethical questions are posed and the implications of data mining for marketers are summarised.

1. Introduction

The growth of the Internet has led to a critical mass of companies and consumers participating in a global on-line marketplace. This has spurred companies world-wide to experiment with innovative ways of marketing to consumers in computer-mediated environments [Hof95]. The exponential growth of the Internet has acted as a catalyst for the growth of electronic commerce. Through this relatively new commercial medium, businesses can market directly on-line with little advertising to a potential 40 million consumers connected across 130 countries [Mel95].

The Internet has acted as a catalyst for the growth of electronic commerce. An estimated base rate of 1 million new users a month is projected to connect 500 million users world-wide by the end of the millennium [DeA96]. The vast amount of data on the Internet has left experts with the requirement to devise new strategies for Internet marketing. That is, precise targeting of prospective consumers and tailoring of product offerings are crucial in achieving competitive advantage in such a large electronic marketplace [Lev96]. Companies are now finding that in order for their Internet presence to fulfil its promise as a competitive business tool - or in order to attract advertising revenue - they must learn more about the individuals visiting their site.

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 business

The main objective of this paper is to present a strategy for the use of data mining of Internet data for marketing purposes. The outline of the paper is as follows. In Section 2, we show the urgent need for improved marketing support techniques due to the rapid growth of the Internet. In Section 3 we present typical scenarios of electronic commerce on the Internet in form of shopping malls. Section 4 describes marketing techniques that are currently in use on the Internet. Section 5 outlines the types of

available data that are accumulated when customers visit an electronic mall. Sections 6 and 7 describe the data mining process briefly and illustrate its application to Internet data sources. Finally, in Section 8, related work is presented before conclusions are drawn.

2. The Growth of Business on the Internet

An increasing number of shopping malls, selling an ever-widening selection of products are becoming available on the Internet. Business-orientated activities on the Internet are entering a period of rapid growth, at present, some 80,000 businesses are trading electronically using this global medium [And96]. The business of on-line retail selling is poised to take a mammoth leap within the next five years with total sales of goods via the Internet expected to leap from its 1995 level of $200 million to $189 billion annually by the year 2000 [Von96]. This is the prediction of a survey released in 1996 by the US research firm International Data Corporation (IDC). IDC state that their projection of $189 billion in on-line sales by the year 2000 is realistic. However, a report released by Forrester Research Inc., estimated on-line retailing revenues reached $518 million in 1996 and predict that it will grow to $6.58 billion by the year 2000 [For96].

According to IDC's latest survey, 29% of Internet users have already shopped on-line and 91% plan to make purchases on-line in the future. Recent demographics show that 39% of US households owned a personal computer in 1996 and this figure is steadily increasing as more people invest in tele-working. Furthermore, analysts predict that 85% of household PC owners will use on-line services by the year 2000, which will have obvious positive implications for home shopping sales [Pal96]. Home users logged an average purchase total of $50 monthly in 1996, while business users logged a monthly purchase total of $200. For major UK companies with a turnover of £200 million, electronic business represents just 3% of sales in 1996 at present. However, these companies believe that on-line sales through the Internet will be as high as 20% by the year 2000 [KPM96].

It can be assumed that by 1998, the largest obstacles to commercial development of the Internet will have been resolved [Gar96]. Major breakthroughs are expected in backbone performance, local access performance and most importantly, in reliable cost-effective electronic payment systems that will cause the most fundamental change in trade since paper money has been introduced [For94]. Various companies such as Digicash, CyberCash and FirstVirtual are currently deploying and testing several mechanisms for conducting secure business transactions over the Internet. 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).

3. Electronic Commerce in On-line Shopping Malls

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. The catalysts are the widespread international adoption of the networking standard TCP/IP, World Wide Web (WWW) browsers and the HyperText Transfer Protocol (HTTP), which are the de facto standards facilitating EC.

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. The figure also shows that there are many possible permutations of interactions. 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.

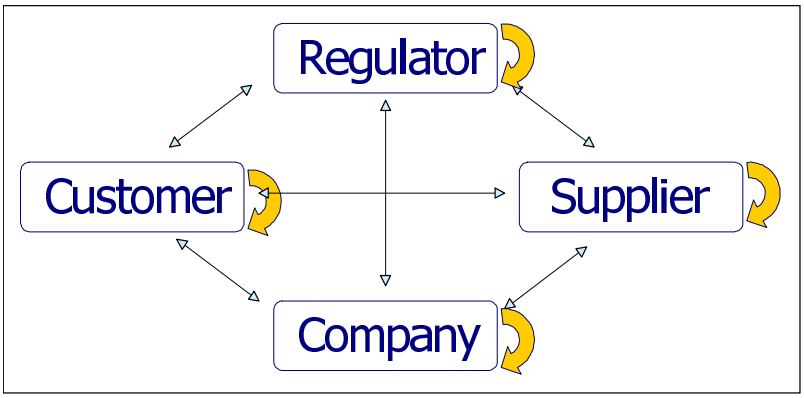

Figure 1: Inter-Organisational Communication Paths in Electronic Commerce

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. This company-to-customer or the on-line shopping mall model is the permutation selected in this paper for the application of data mining. These primary reasons for selecting the on-line shopping mall model are that:

Consumers exhibit a wide range of differing behavioural patterns;
Consumer descriptive information provides a rich source of useful data;
Products and services in on-line malls are varied.

Each of the above provide consumer data, which is at the core of successful target marketing and fully suitable for data mining.

4. Internet Marketing Techniques

[Dib97] 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 [Kot97] 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.

Sophisticated successful marketing takes this a step further. An organisation must understand its customers' needs and wants, and it must satisfy them with a product or service that displays some form of competitive advantage. With an understanding of not only the customer but also the marketplace (the business environment and competition) as well as internal capabilities, a marketing strategy can be developed. In developing the strategy, an organisation attempts to identify groups of customers where each separate group, or market segment, has 'similar' needs. Each group of customer can then be offered a specifically tailored programme, also known as target marketing.

However, it is a well known fact that success in marketing is never guaranteed. In today's dynamic business environment, marketing decisions are extremely complex. Such decisions as 'what customers to target' or 'what products or services to offer targeted customers' demand reliable and accurate knowledge. As [Dru83] states, "The aim of marketing is to make marketing superfluous. The aim is to know and to understand the customer so well that the product or service fits him/her and sells itself".

4.1. The Challenge of 90s Customer

[Mur96] describe the customer of the 90s as "becoming increasingly sophisticated, sceptical and literate - in short, harder to persuade". They argue further that faced with 'advertising clutter' the consumer simply 'switches channel' or moves to a different Internet on-line site. The problem of 'switchers' became evident during the 1980's when research revealed that most shoppers were 'notoriously' fickle, and would switch from store to store at will [Kau84]. Media fragmentation, market demassification and in particular the advent of interactive technologies have heightened the problem of 'switchers'. According to [Cro95] "interactive technologies transform and redefine relationships between customer

and the companies with whom they do business". In direct opposition to traditional methods of marketing communications, on-line shopping empowers consumers "to choose which site they visit" [Wil96].

Therefore, the exchange relationship at the heart of marketing is evolving and "giving more control to the customer" [For96]. In the context of on-line shopping, the current population of users of the Internet reflects the demographics of medium to high-income, consumers with little patience and no allegiance to particular on-line malls. This represents a challenge to the marketer to develop highly targeted marketing programmes and in particular effective marketing communications.

4.2. Marketing Communications

Today's on-line consumer is astute, discerning and empowered. The result is that the era of mass marketing communication has largely ended. No more hard sell advertising, but rather marketing communication which is targeted and customised to the needs of consumers. In this section, a general explanation of marketing communication is provided and is followed by an introduction to targeted marketing communications most applicable to on-line shopping.

The role of marketing communication is to inform the market clearly, persuasively and accurately about the company, its products and services. To do this, marketers typically use four major tools: advertising; sales promotion; public relations; and personal selling. Whilst advertising is a recognised tool of marketing - public relations, sales promotion and personal selling, are, generally not so well recognised despite the fact that they each achieve unique communication objectives. Public relations involves publicity and can be used to manage the image of a company. Personal selling is the most direct form of communication and involves face-to-face selling. It is when sales promotion is examined in closer detail that a striking trend of the 1990s is revealed. For, whilst it is accepted that advertising remains the most important *form* of communication, current figures (UK) show that *actual expenditure* on sales promotion exceeds that for advertising. The growth of on-line shopping and Internet marketing is fuelling such expenditure as on-line marketers realise the benefits of offering highly targeted sales promotions.

Sales promotion can help on-line marketers overcome the 'problem' of impatient, 'switching' consumers by converting 'switchers' into loyal and repeat purchasers. According to [McD84] sales promotion is a "specific activity, which can be defined as the making of a featured offer to defined customers with a specific time limit". In general terms, a sales promotion helps to encourage patronage, a high level of product awareness and can even encourage trial and repeat purchase of a product or service. Despite criticism regarding its short-term focus, sales promotion can in fact help a company to achieve strategic, or long-term, marketing success. For example, through a series of strategic sales promotions market share can be increased cumulatively. It is appropriate now to examine the strategic choices to be made in implementing a sales promotion campaign.

4.3. Promotional Strategy: Push and Pull

Irrespective of the type of promotion or communication technique, there are two basic promotional strategies: Push and Pull. The *push* strategy is aimed at intermediaries (such as wholesalers) in the channel of distribution. The *pull* strategy is aimed at the ultimate customer or user and directly attempts to create demand for the manufacturers' goods to pull goods through the shop from the consumer end of the channel.

The Internet provides a pull environment whereby competing Internet sites try to *attract and maintain* consumer interest and attention. However, according to [Vas96] there are inherent limitations to the existing Internet pull environment: "to maximise their impact, marketers need to exploit a push strategy as well". Vassos describes this stage as the "Outbound Stage of Internet site development" and cites American Airlines (AA) as a useful example. AA has launched an outbound offering (sales promotion) called 'Net SAAver Fares', where each week AA will send customers a notification about the availability of discount air fares for the upcoming weekend[1].

In the case of on-line shopping, a push strategy would involve a targeted sales promotion aimed at the ultimate consumer. The retailer can employ a variety of techniques in a short-term or tactical way to increase sales without the commitment of longer-term changes. For example, a price mark-down can be used to clear out stocks of winter fashion at the end of a season, this does not constitute a permanent

[1] http://www.americanair.com/aa_home/net-saavers.html

reduction in price. There is a wide range of sales promotion types, such as consumer incentives, dealer incentives, sales force incentives, merchandising, window displays, exhibitions, packaging, sponsorship, etc. One of the most advanced types is direct marketing, which includes forms such as database marketing and on-line marketing

4.4. The Direct Marketing Process

As a form of sales promotion, direct marketing is ideally suited to Internet marketing. It is a tool for direct communication with target customers of a special product and promotion offer that has a high chance of appealing to customers [Dic97]. With the creation of a database, targeted sales promotions can be offered on-line to 'hot prospects' rather than, as in the case of the Internet pull environment, promotions aimed at disinterested, impatient, unspecified and more importantly unlikely to buy groups of consumers. Hence, the opportunity is provided for a push strategy. The traditional process of direct marketing is outlined in Figure 2.

1. Database elements are specified, such as name, address, phone, fax, gender, income, hobbies, product usage situations, benefits sought etc. The source of record and date are added.
2. Customer sends in warranty card, enters a competition, or has name on list of company selling related products or services. Permission is given and information is used for direct marketing.
3. Record is added to database using database software.
4. Target segment is specified and statistical analysis performed. Hobbies, interests, usage situation, and promotion preference identified.
5. Most appealing direct-marketing campaign is designed to promote most appealing products and features to target segment.
6. Database is updated, recording response of customer to campaign.

Figure 2: The Direct Marketing Process, adapted from [Dic97]

4.4.1. Database Marketing

It can be seen from Figure 2, that **Database Marketing** is critical to the process of direct marketing. The database can be used in four major ways: to identify target groups or prospects; to decide which customers receive which offers; to build customer loyalty; and to reactivate customer purchases [Kot97]. When rating and selecting people from the database [Sto93] recommends applying the R-F-M formula (recency, frequency and monetary amount). The best customer targets are those who bought most recently, those who buy frequently and those who spend the most. Points are established at various R-F-M levels and individual customers are scored - the higher the score, the 'hotter' the prospect.

4.4.2. On-Line Marketing

The process of Direct Marketing can be adapted to **On-line Marketing**. One of the most recent developments in the direct marketing process, it is simply another channel of marketing where a firm can "meet and condition a prospect" [Kot97]. Not surprisingly Microsoft promotes its products on the Internet. In 1996, an extra dimension was added when the company provided a definitive on-line guide to the Euro '96 Soccer championships. A web site was set up for four weeks to give access to the information that only the press and the championship organisers had previously been able to obtain. The web site recorded 18 million hits during the championships, 1.26 million of them during the peak day. Over the same period, traffic to Microsoft's permanent UK web site increased by 78%.

4.4.3. Adapting the Direct Marketing Process to On-line Shopping

On-line retailers, in particular can benefit from the adaptation of the process of direct marketing. A sales promotion based on previous shopping behaviour and other descriptors can be offered at the end of an on-line shopping visit. Hence the emergence of the 'soft push' strategy. An illustration of how direct marketing might be adapted for on-line shopping is at Figure 3.

1. Database elements are specified, such as name, address, phone, fax, gender,web sites visited, recent purchases, frequency of web sites visits, expenditure etc.
2. Micro Marketing: Navigational and purchasing behaviour data gathered and added to database.
3. Target Segment is specified.
4. Data Mining performed. Customer profiles built, trends and behaviour patterns identified.
5. Mass Customisation: most appealing sales promotions offered to target segment.
6. Database is updated, recording response of customer to sales promotion.

Figure 3: The Soft-Push Marketing Process

There can be no doubt that technology is driving on-line shopping forward. "The ability to gather and analyse vast quantities of data is the bedrock of all targeted marketing, allowing marketers to understand and predict how their customers behave" [Cur96]. High marketing costs see marketers looking to cost effective technological developments to capture and store customer data. Given the increasing volume of such data, technology is providing the answers for many direct marketing questions. Consequently, the language of on-line shopping is peppered with terms such as: 'Mass Customisation' (Figure 3, stage 5) and 'Micro Marketing' (Figure 3, stage 2).

Traditionally referring to flexible production techniques, the term **mass customisation** has come to be associated with on-line marketing. Shikar Ghosh, chariman of US Internet company Open Markets has suggested that "with the ability to customise services at every level, the marketer on the Internet can for the first time target economically a segment of one person - and that person can interact". Hence, mass customisation refers to a customised and tailored marketing 'offer' aimed at segments consisting of one individual. For the first time, it is possible to consider a cost-effective means of individually customised marketing communication. Such customised marketing communication, however, relies on reliable and accurate information about the customer, and it is to micro marketing that we turn to provide such information.

Internet technology offers businesses the opportunity to gather market intelligence and monitor customer choice through customers revealed preferences in navigational and purchasing behaviour on the web. This process of gathering information on customers is termed **micro marketing**. It can enable a business to customise future offerings for a particular customer, i.e., it is a proactive approach which anticipates customers' future needs [Nor97]. Customer data is the "DNA of the organisation, it is the blueprint for the customer's relationship with the brand and it's right at the heart of strategy" [Mom96].

4.5. Data Mining - the 'Soft-Push' Approach

Mining or analysing consumer navigational and purchasing behaviour is the main focus of this paper. 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 which segments 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 [Hum96]. The potential for direct marketing is enormous - in theory 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 on-line shopping mall data provides high-level knowledge - 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 implement a *soft-push* marketing strategy.

The high-level mined rules may be incorporated as a rule-based system into the architecture of an on-line mall service. When each new consumer interacts with the on-line mall to navigate and purchase goods or services, their on-line behavioural patterns are identified by the rule-based system. When this

happens, the on-line mall system immediately reacts to change dynamically the information presented to that consumer.

It should be noted that the terms 'push' and 'pull' already exist in Internet terminology. However, when used in this paper the context of the terms comes from their use in marketing.

5. Data Sources in On-Line Shopping

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. Either way, all actions are being kept track in log files. In addition, external information is being collected and accessed during a shopping mall visit. The structure and connotation of the different types of data sources are outlined in the following sub-sections.

5.1. Server Logs

On-line shopping servers produce several network protocols which have the potential to be mined: The *Common Log Format* 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 that can then be parsed. For example, the domain name www.kaist.ac.kr can be resolved to provide information that identifies the originator as an academic from the Republic of Korea. However it must be recognised that not all domain names can be resolved. Alternatively, on-line resources such as InterNic may be searched for domain names. *Error protocols* record all occurring faults while being connected to an on-line shopping server.

5.2. 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[2] 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[3] 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 security_level', which can be customised depending on the applied domain.

5.3. 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 depend on the nature of the on-line shopping business and the intended usage of the data. Typical requested information is about age, gender, interests, likes, and dislikes, etc.

5.4. Miscellaneous Sources

It is common practise to obtain demographic data from third party suppliers. This 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] http://home.netscape.com/newsref/std/cookie_spec.html

[3] For example, Apache Internet server

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 that provides such information is DoubleClick[4].

6. The Data Mining Process

Data mining should be viewed as a process involving several automated and non-automated steps rather than a single step. In this section we describe each of the steps in the data mining process [Ana96], which is depicted in Figure 2, and discuss the different aspects of each of these steps.

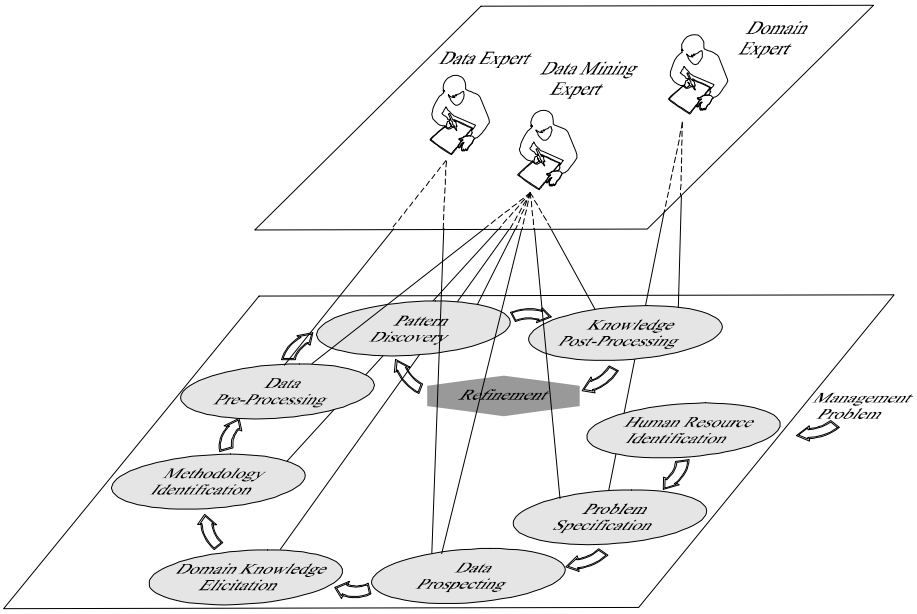

Figure 4: The Data Mining Process

After a problem has been identified at the management level of an organisation **human resource identification** is the first phase of the data mining process. In most real-world data mining problems the human resources required are: the *domain expert*, the *data expert* and the *data mining expert*. Normally, data mining is carried out in large organisations where the prospect of finding a domain expert who is also an expert in the data stored by the organisation is rare. The need has to be stressed for bringing together these human resources early into the process as any project that does not bring together these expertise right at the beginning of the process will very likely encounter problems later on.

Problem specification is the second phase of the data mining process. Here a better understanding of the problem is developed by the human resources identified in the human resource identification component of the project. Different techniques are used to tackle each of these tasks. Therefore, identifying the task type that the problem falls into is important. 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.

The second part of the problem specification phase is to identify the ultimate user of the knowledge. Clearly if the knowledge discovered is to be used by a human, it must be in a format that the user can

[4] http://www.doubleclick.com

understand and is used to. However, if data mining is only a small part of a larger project and the output from data mining is to be interpreted by a computerised system, the format of the discovered knowledge will have to strictly adhere to the expected format.

Data prospecting is the next phase in the process. It consists of analysing what is the state of the data required for solving the problem at hand. There are three main considerations within this phase: What are the relevant attributes? Is the data required electronically stored and accessible? Are the data attributes required populated?

Knowledge discovery is not totally automated and independent of user intervention, and hence the importance of **domain knowledge elicitation**. The domain expert can often provide domain knowledge that can be used by the discovery algorithm for making patterns in the data more visible, pruning of the search space or for filtering the discovered knowledge based on a user driven interest measure.

The main task of the **methodology identification** phase is to find the best knowledge discovery methodology to solve the specified mining problem. Often a combination of methodologies is required to solve the problem at hand. The most commonly used technologies are rule induction, derivatives of traditional statistics, genetic algorithms, evidence theory, case-base reasoning, Bayesian belief networks, fuzzy logic, rough sets and neural networks. The chosen paradigm depends on the type of information that is required and the domain of knowledge being discovered. The selected technique also influences the format of the input data, whose preparation is part of the data pre-processing phase of the data mining process.

The next phase is that of **data pre-processing**. Depending on the state of the data this process may or may not constitute the phase where most of the effort of the data mining process is concentrated. Data pre-processing involves removing outliers in the data, predicting and filling-in missing values, noise modelling, data dimensionality reduction, data quantisation, transformation and coding, as well as heterogeneity resolution. Data pre-processing in data mining has often been considered the equivalent of the 'stick of dynamite' within the mining of ores. No pre-processing will result in the really useful knowledge remaining undiscovered. However, too much pre-processing may result in the discarding of the data in which the interesting knowledge is implicit. Thus, there is a balance required during pre-processing.

The **pattern discovery**[5] phase follows the data pre-processing phase. It consists of using algorithms that automatically discover patterns from the pre-processed data. The choice of algorithm depends on the mining task at hand. Due to the large amounts of data from which knowledge is to be discovered, the algorithms used in this phase need to be efficient. Techniques for learning with sampling and high performance computing are important considerations within this phase.

The last step of the data mining process is **knowledge post-processing**. Trivial and obsolete information has to be filtered out and discovered knowledge has to be presented in a user-readable way, using either visualisation techniques or natural language like constructs. Often the knowledge filtering process is domain as well as user dependent, and thus requires domain specific semi-automated techniques.

The last two steps of the data mining process usually form a refinement process, which must be iterated through until sufficient results have been achieved.

7. Applying the Data Mining Process to Internet Data Sources

Data, which is available from on-line shopping environments (see Section 5), can be utilised in the data mining process as following.

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. Examples are the detection of potential customers for a newly launched product, mutually

[5] This phase is often referred to as data mining, however, we use data mining to describe the overall knowledge discovery process, as it reflects the industrial usage.

exclusive products to be price-reduced, or re-arrangement of product pages (virtual shelves) depending on user's behaviour. More task types have been given in Section 6.

Data prospecting is concerned about the state of the different logs. Possible problems can be missing and false values in log files, incompatibilities and heterogeneity among different protocols or inaccessibility thereof. Additionally, external available data sources, which can be incorporated in the data mining process, have to be identified and identical problems perceived.

Available **domain knowledge** can be either known information about the problem being tackled or about the structure of the on-line shopping mall. For example, the logical topology of an Internet server can be modelled as a semantic network and incorporated as domain knowledge.

As described in the previous section, the **methodology identified** depends on the problem at hand, the data and knowledge available, as well as the type of information to be discovered.

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, filling in missing values, and removing false entries or outliers. The types of data pre-processing to be performed depend on the state of the available log file(s) as identified at the data prospecting stage.

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 these discovered patterns to a format that is understandable for humans or applicable for the system which generates user and task sensitive WWW pages dynamically. Depending on the quality of the discovered patterns (level of novelty, degree of certainty and confidence, quantity, applicability, etc.) latter two steps have to be repeated after thresholds have been refined accordingly.

8. Related Work

The potential for data mining of Internet data (also known as Web Mining) has been recognised by various parties, whose endeavours are described in this section and compared to our 'soft-push' approach. Web Mining can be sub-divided into three subtasks [Etz96], namely resource discovery, information extraction and generalisation.

Resource discovery is concerned about finding documents and services on the Internet, which have been unknown to the user beforehand. There are several search engines available on the Internet, which operate on different levels of sophistication. [Che97] proposes an intelligent agent based system, which assists users to locate documents related to their interests. The learning agents facilitates data mining algorithms using user access logs. [Spe97] is facilitating different structural Internet information, such as hypertext links, domain names, relationships between concepts represented by words and phrases, or paths travelled through Internet sites by visitors. These artefacts are then used to discover information interesting to the user, such as moved pages, related pages, persons, etc.

Information extraction is concerned about excerpting specific information from newly discovered Internet resources. This leads to the area of mining semi-structured data, i.e. structured text documents. Various specialised systems have been developed, e.g. for information extraction of HTML meta data. A more sophisticated approach has been presented by [Per95], which learns to extract information from unfamiliar resources by querying them with familiar objects and matching output returned against knowledge about the query objects.

Generalisation is concerned about discovering (behavioural) patterns at individual Internet sites. [Mac96] uses standard HTTP server log files and applies existing association discovery algorithms to detect behavioural patterns of users. The discovered rules can then be used to support marketing decisions or to redesign the Internet server. [Chu97] has built a system which not only generates categories from stored visitor sessions, but also provides a module which suggests actions dynamically depending on the visitor's actions as well as the generated categories. The advantage of this system is that it can adapt to the user's behaviour and suggest Internet locations dynamically.

Our approach supersedes the above endeavours, in that it embeds the knowledge discovery techniques in the data mining process. It also incorporates marketing specific expertise as domain knowledge and facilitates the discovered rules to tackle customers on an individual basis. The specific discovered knowledge can be used to generate Internet pages dynamically, and thus push the visitor softly towards his/her behavioural profile.

9. Conclusions

The urgent need for more advanced marketing techniques has been shown, which is caused by the exponential growth of the Internet and the more sophisticated customer of the 90s. The most appropriate form of marketing techniques – direct marketing – has been brought into context with electronic commerce and the on-line shopping mall model. Our 'soft-push' approach borrows from promotional strategies and enhances its features to fit into an Internet scenario. Information, in form of Internet logs, which is available from such sites has been outlined, and a data mining process has been applied to those sources to harness them accordingly.

One aspect that 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 answered satisfactorily, before data mining technology is coupled to on-line shopping systems.

The implications of data mining for marketers are far reaching, since it can improve understanding of on-line customer behaviour and allow for more efficient communication with such customers. However, the key for marketers is to remember the basic principles of what they are doing and know what consumer data is or is not important. Data mining can do the spadework whilst marketers concentrate on developing creative and imaginative communication campaigns.

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.

[And96] C. Anderson: Net Profits: A survey of the Internet, *The Economist*, 1996.

[Cha96] C.A. Charles, C.P. Foss, S. Dewan (Eds.): Globalizing Electronic Commerce, Report *Int. Forum on Electronic Commerce*, Beijing, China, 20-21 March 1996, Center for Strategic & Int. Studies.

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

[Chu97] P. Chundi, U. Dayal: An Application of Adaptive Data Mining: Facilitating Web Information Access, *Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 31-37, 1997.

[Cro95] P. Cross, S. Smith: Internet Marketing, *Marketing Week*, July:37-42, 1995.

[Cur96] J. Curtis: Too Clever by Half, *Marketing*, May 9:29-32, 1996.

[DeA96] M. DeAngelis: Marketing on the Internet. NetResource.Com. 1996, http://netresource.com/itp/repch6.html

[Dib97] S. Dibb, L. Simkin, W.M. Pride, O.C. Ferrell: Marketing Concepts and Strategies, 3rd European Edition, Houghton Mifflin, Boston, New York, 1997.

[Dic97] P. Dickson: Marketing Management, 2nd Edition, The Dryden Press, USA, pg. 560, 1997.

[Dru83] P. F. Drucker, Managing for Results, Heinemann, London, 1983.

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

[For94] Forrester Research: The New Customer Connection, Computing Strategy Report, September 1994.

[For96] Fobairt Report: Ireland: the digital age, the Internet, Doing Business, 1996.

[Gar96] Gardner Group: Internet Strategies, http://www.gardner.com/whatsnew/inettv.html, 1996.

[Hof95] D.L. Hoffman, T.P. Novak: Commercial scenarios for the Web: Opportunities and challenges, *Journal of computer mediated communication* "Special Issue on Electronic Commerce" October, 1995.

[Hum96] C. Humbry: Digging for Information, *Marketing*, November 21: pp. 41-42, 1996

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

[Kau84] A.K. Kau, A.S.C. Ehrenberg: Patterns of Store Choice, *Journal of Marketing Research*, XXI:339-409, 1984.

[Kot97] P. Kotler: Marketing Management: Analysis, Planning, Implementation and Control, 9th Edition, Prentice Hall International Inc, pp. 466-467, 1997.

[KPM96] KPMG UK Management Consulting: Electronic Commerce: Key Kindings, 1996.

[Lev96] K. Levis: Electronic Commerce, *British Communications Engineering*, 14:281-285, 1996.

[Mac96] J. Mace: Internet Usage Analysis: A detailed Study of an Electronic Commerce Web-Site, *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.

[McD84] McDonald: Marketing Plans: How to Prepare Them, How to Use Them, Heinemann, London, pg. 110, 1984

[Mel95] K. Meltsner: Understanding the Internet: A Guide for Material Scientists and Engineers, *Journal of Management*, 47(4):9-10, 1995.

[Mom96] M. Momen as cited in C. Bond: The Heart of the Matter, *Marketing*, November 21:38-39, 1996.

[Mur96] J. Murray, A. O'Driscoll: Strategy and Process in Marketing, Prentice Hall International Inc, 1996

[Nor97] M.T. Norwood, C. Grant: The Internet: Interactive, Integrated, Inevitable, *Proc. 1st Academy of Marketing Conf.*, Manchester Metropolitan University, forthcoming, 1997.

[Pal96] P. Pallab: Marketing on the Internet, *Journal of Consumer Marketing*, 13(4):27-39, 1996.

[Per95] M. Perkowitz, O. Etzioni: Category translation: learning to understand information on the internet, *Proc. 15th Int. Joint Conf. on Artificial Intelligence (IJCAI'95)*, pp. 930-936, 1995.

[Spe97] E. Spertus: Parasite: Mining Structural Information on the Web, *Proc. 6th Int. World Wide Web Conf.*, http://www6.nttlabs.com, 1997.

[Sto93] R. Stone, Successful Direct Marketing, NTC Books, Chicago, 1993.

[Vas96] T. Vassos: Strategic Internet Marketing, Que Books, Canada, 1996.

[Von96] H. Vonder: Mammoth Growth Predicted for Cybershopping, *Interactive Week On-Line Magazine*, April 1996.

[Wil96] R. Wilson: BMW Steers into Cyberspace, *Marketing Week*, July 26:42, 1996.