DATA MINING AND USER PROFILING FOR AN E-COMMERCE SYSTEM

Ken McGarry, Andrew Martin, Dale Addison and John MacIntyre

School of Computing and Technology,
University of Sunderland, St Peters Campus,
St Peters Way, Sunderland,
United Kingdom SR6 ODD
e-mail: ken.mcgarry@sunderland.ac.uk

ABSTRACT

Many companies are now developing an online internet presence to sell or promote their products and services. The data generated by e-commerce sites is a valuable source of business knowledge but only if it is correctly analysed. Data mining web server logs is now an important application area for business strategy. We describe an e-commerce system specifically developed for the purpose of demonstrating the advantages of data mining web server logs. The data generated by the server logs is used by a rule induction algorithm to build a profile of the users, the profile enables the web site to be personalized to each particular customer. The ability to rapidly respond and anticipate customer behaviour is vital to stay ahead of the competition. The e-commerce site is written in Java and uses an off-the-shelf data mining package and a freely available web-server.

1. INTRODUCTION

This paper describes the development of a software tool for the exploratory analysis of website data. The project was initiated by the Centre for Adaptive Systems (CAS), which is a part of the School of Computing, Engineering and Technology at the University of Sunderland in the UK. The need for this project arises from the Centres consultancy activities, which consists of advising small to medium enterprises (SME) on Information Technology and Artificial Intelligence techniques. Many companies are interested in improving their web-based profitability by profiling and targeting customers more effectively. The system used as an exploratory tool to develop an understanding of the issues involved in mining website data. In addition, it is used to demo these principles to SME's to encourage take-up of the technology. The system consists of three main stages:

- The development of a small mock e-commerce system.
- The analysis of server logs, cookies, user generated forms using data mining techniques such as rule induction.
- The development of suitable models and rules to predict future user activity.

Data mining and knowledge discovery is generally understood to be the search for interesting, novel and useful patterns in data on which organisations can base their business decisions [3]. Although statistics has a place within the data mining paradigm, most of the techniques used are based on algorithms devised by the machine learning and artificial intelligence community. A wide variety of techniques have been successfully used such as neural networks [9, 10, 8], decision trees/rule induction [13], association rules [1] and clustering [4].

Another aspect of the data mining process that needs to be addressed is the identification of what constitutes an interesting and useful pattern. There are two main methods of determining this. The first uses objective mathematical measures to assess the degree of interestingness, many such measures exist and are domain dependent [5, 2]. The second method is to incorporate the users subjective knowledge into the assessment strategy. Each of these approaches has various characteristics for example the subjective method requires access to a domain expert [6, 7]. The determination of interestingness is likely to remain an open research question.

Many companies are now using data mining techniques to anticipate customer demand for their products and as a result these companies are able to reduce overheads and inventory costs. Companies such as Walmart are well aware of the benefits of data mining and

have used the technology to modify their business strategy. The introduction of online transactions and the nature of the data involved with this form of business has given data mining increased scope. The internet is a medium for business transactions that is available 24 hours a day, all year round and is world-wide in scope. The potential for increased business opportunities is enormous but many technical challenges need to be addressed.

The remainder of this paper is structured as follows: section two discusses the advantages and techniques used for profiling online customers, section three describes the WEKA data mining package and how it is used within our system, section four presents the details of the server platform implemented, section five describes the development and experimental work performed on the system, section six discusses the conclusions.

2. DATA MINING THE WEB

A web site is often the first point of contact between a potential customer and a company. It is therefore essential that the process of browsing/using the web site is made as simple and pleasurable as possible for the customer. Carefully designed web pages play major part here and they can be enhanced through information relating to web access. The progress of the customer is monitored by the web server log which holds details of every web page visited. Over a period of time a useful set of statistics can be gathered which can be used for trouble shooting e.g. when did the customer abandon the shopping cart? or why do customers view a page but do not buy? Therefore, problems with poorly designed pages may be uncovered.

However, the main advantage of mining web server logs relate to sales and marketing. Sites like Amazon hold individual customer's previous product searches and past purchases with which to target this particular individual. In essence the data residing within the web server log can be used to learn and understand the customers buying preferences. This information can be actively used to target similar types of customers for example in promoting special offers or advertising new products to customers most likely to respond. For a good general introduction to web mining see the work of Mena [11, 12]. Data gathered from both the web and more conventional sources can be used to answer such questions as:

- Marketing - who's likely to buy?
- Forecasts - what demand will we have?
- Loyalty - who's likely to defect?
- Credit - which were the profitable loans?
- Fraud - when did it occur?

Information is available from:

- Registration forms, these are very useful and the customers should be persuaded to fill out at least one. Useful information such as age, sex and location can be obtained.
- Server log, this provides details of each web page visited and timings.
- Past purchases and previous search patterns, useful for personalization of web pages.
- Cookies, these reside on the customers hard drive and enable details between sessions to be recorded.

3. WEKA DATA MINING PACKAGE

WEKA was developed at the University of Waikato in New Zealand [14]. WEKA is written in Java and implements many AI algorithms (classification and clustering). In this project its role is primarily to mine combined user profiling and buying/browsing data patterns. The classifier used is called Prism and is based on ID3 [13]. The software can be used as a standalone package or implemented within the users system through DLL's. For an example of Prism rules see Figure 1.

If occupation = execman
and age = 18
and salary = 2 then ALC1
If occupation = compeng
and age = 119
and salary = 8 then ALC2
If occupation = ostdnt
and age = 18
and salary = 1 then ALC2
If occupation = execman
and age = 18
and salary = 2 then ALC2
If age = 21
and salary = 8
and occupation = retired then ALC3

Figure 1: Prism rules

The generated rules are propositional $< IF..THEN >$ classification rules, they have several conditions in each antecedent and refer to a single consequent or class.

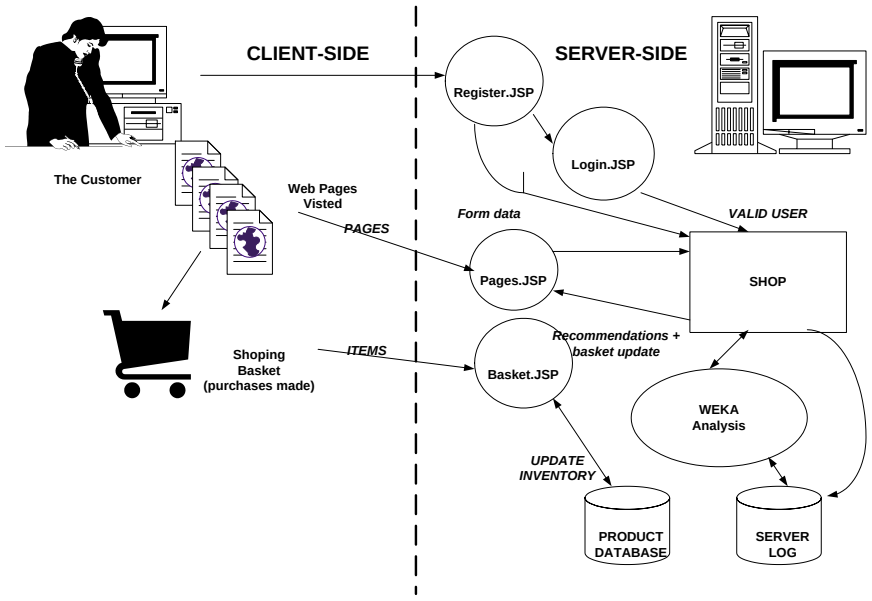

Figure 2: System Overview of the Simulated Shop

4. SYSTEM DEVELOPMENT

The system is comprised of a number of interacting Java modules. Figure 2 shows how the modules interact. Data from the customer is gained from the initial action of registering, here details such as name, occupation, age, sex and other details are entered. Such forms are very useful but users will only fill them out given sufficient incentive such as free gifts, discounts etc. Other data becomes available when users make purchases and browse web pages, this is reported by the server log and shopping basket Java servlet. After the initial trials with a student user base we gathered enough data to train the Prism rule induction algorithm.

The shop consists of three levels of pages, the main shop page which welcomes the user and offers them a choice of products. The subcategory pages which offer specific types of products e.g bread.jsp and memory.jsp. The products pages, such as memsdram.jsp and breadgranary.jsp these display several items which may be purchased. These pages are linked into the database and update the stock count when purchases are made. The final page is the checkout.jsp page which allows the users to buy their selected products. The system gives each registered user some e-money or credit from which to make purchases. Checks are made to ensure the user is registered and logged in before a session is valid. As the users make purchases a shopping basket implemented by a Java servlet maintains track of all details e.g. cost, number of items. The servlet ensures that the user is still in credit and checks that items are in stock before purchases can occur. Valid transactions also update the users credit and the database. Any such conflicts are reported to advise the user. Figure 3 shows the main page of the simulated shop.

4.1. Web Server Platform

The Apache tomcat server is a free program from the Apache organisation and is used to serve static HTML, dynamic JSPs and Java Servlets, making it an important part of the system. Tomcat is written in Java and provides two important aspects to any Java web server. First, it can serve static pages making it highly useful and simpler for development purposes (avoiding the need to set up a second server like apache to serve static HTML). Second, it can use Java Servlet technology to allow logic intensive code to be run native to the system as well as being able to serve JSP pages to allow dynamic content to be added to a site.

The Tomcat server was chosen to provide Servlet and JSP support; it is currently the best Java web server available and therefore made it the primary choice for serving the dynamic content within the shopping component of this system. The source code is available, bugs and bug fixes are well known and it is a very stable platform.

An example of the server log is shown in Figure 4,

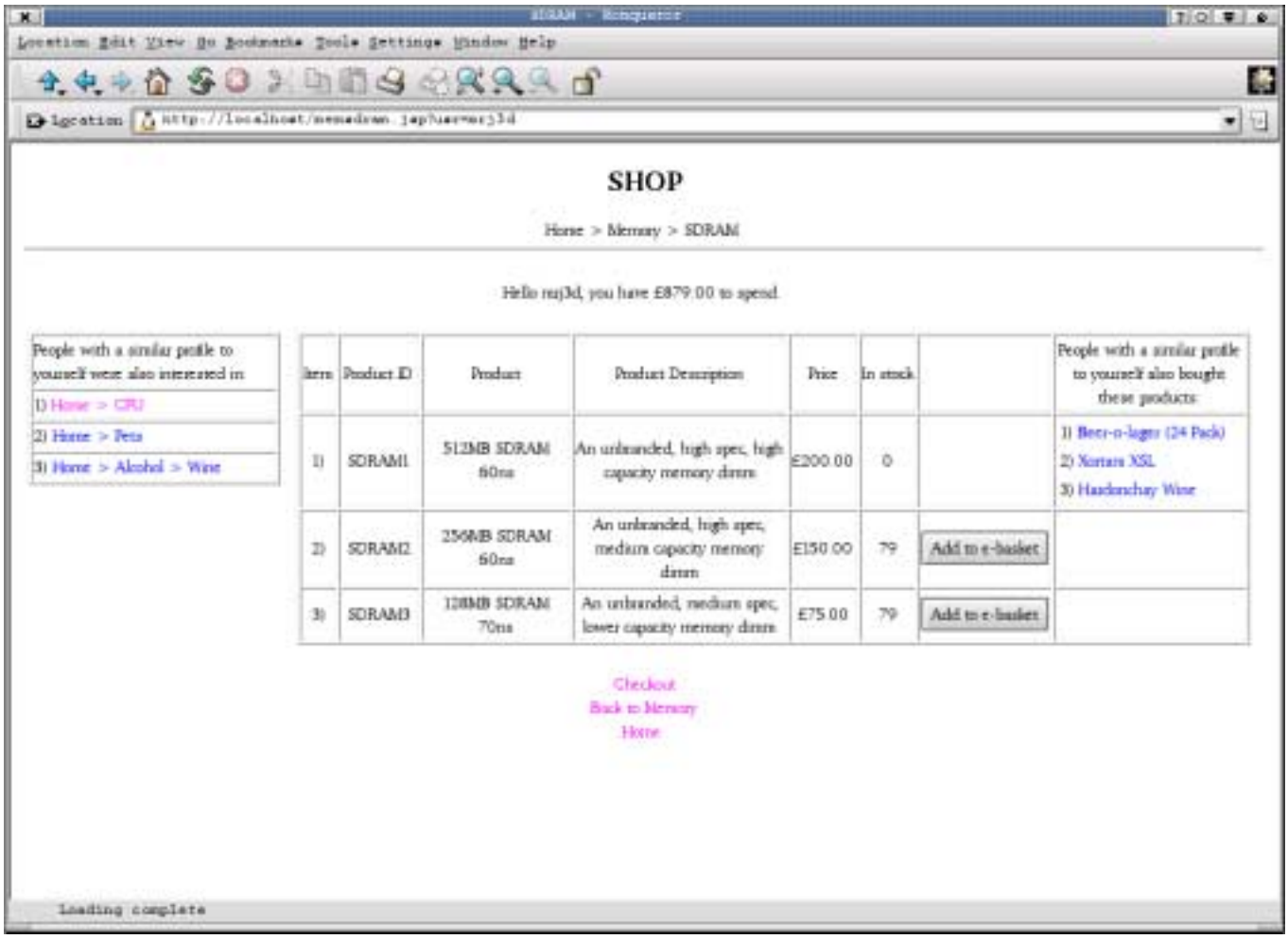

Figure 3: Homepage of the Simulated Shop

194.82.103.77 - - [19/Feb/2002:18:06:52 00] "GET /checkout.jsp HTTP/1.0" 200 793
194.82.103.78 - - [19/Feb/2002:18:07:06 00] "GET /shop.jsp HTTP/1.0" 200 1062
194.82.103.74 - - [19/Feb/2002:18:07:19 00] "GET /checkout.jsp HTTP/1.0" 200 793

Figure 4: Web server extract

here we present a snapshot of a test session where several users were logged in. The IP address of each machine is captured, along with the date and time stamp, next the web pages each user visited is displayed.

4.2. Test Results

The initial system was based on the data generated by 25 students. The site therefore had to cater for the particular preferences held by this group (mainly aged between 20-35). The rules generated after parsing the web server and items purchased generally tended to agree with the expected market segment for this particular group. The web site had to be large enough to give the necessary variability i.e. enough products to differentiate the users. Users logging in subsequently were presented with a personalized web page based on previous purchases and web pages visited. The data mining software was then used off-line to create a user profile based on the form, purchase and server data. The generated rules were later incorporated to provide a more robust user model.

5. CONCLUSIONS

Client companies are generally more inclined to take-up a technology if they can see the end product actually working. The use of a demo system can highlight the potential to clients better than mere sales literature or site visits alone can. The system enables us to experiment and show to clients the possibilities of web server based data mining. Future work will see the addition of a client email list to contact specific customers with details of special offers etc. Also, better use will be made of the support provided by "cookies" to track individual users more closely. The ability to perform the data mining activity on-line without user intervention is also a good candidate for implementation.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD-93*, pages 207–216, 1993.

[2] C. Fabris and A. Freitas. Discovering suprising patterns by detecting occurences of simpson's paradox. In *Development in Intelligent Systems XVI (Proc. Expert Systems 99, The 19th SGES International Conference on Knowledge-Based Systems and Applied Artificial Intelligence)*, pages 148–160, Cambridge, England, 1999.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthursamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI-Press, 1996.

[4] D. Fisher. Knowledge acquistion via incremental concept clustering. *Machine Learning*, 2:139–172, 1987.

[5] A. Freitas. On rule interestingness measures. *Knowledge Based Systems*, 12(5-6):309–315, 1999.

[6] B. Liu, W. Hsu, and S. Chen. Using general impressions to analyze discovered classification rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 31–36, 1997.

[7] B. Liu, W. Hsu, L. Mun, and H. Y. Lee. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):817–832, 1999.

[8] K. McGarry and J. MacIntyre. Data mining in a vibration analysis domain by extracting symbolic rules from RBF neural networks. In *Proceedings of 14th International Congress on Condition Monitoring and Engineering Management*, pages 553–560, Manchester, UK, 4th-6th September 2001.

[9] K. McGarry, S. Wermter, and J. MacIntyre. The extraction and comparison of knowledge from local function networks. *International Journal of Computational Intelligence and Applications*, 1(4):369–382, 2001.

[10] K. McGarry, S. Wermter, and J. MacIntyre. Knowledge extraction from local function networks. In *Seventeenth International Joint Conference on Artificial Intelligence*, volume 2, pages 765–770, Seattle, USA, August 4th-10th 2001.

[11] J. Mena. *Data Mining Your Website*. Digital Press, 1999.

[12] J. Mena. *Web Mining for Profit*. Digital Press, 2001.

[13] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[14] I. Witten and E. Frank. *Data Mining*. Morgan Kaufmann Publishers, San Franciso, 2000.