Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining*

Alex G. Büchner
Northern Ireland Knowledge Engineering Laboratory
University of Ulster
ag.buchner@ulst.ac.uk

Maurice D. Mulvenna
School of Computing and Mathematics
University of Ulster
md.mulvenna@ulst.ac.uk

Abstract

This article describes a novel way of combining data mining techniques on Internet data in order to discover actionable marketing intelligence in electronic commerce scenarios. The data that is considered not only covers various types of server and web meta information, but also marketing data and knowledge. Furthermore, heterogeneity resolution thereof and Internet- and electronic commerce-specific pre-processing activities are embedded. A generic web log data hypercube is formally defined and schematic designs for analytical and predictive activities are given. From these materialised views, various online analytical web usage data mining techniques are shown, which include marketing expertise as domain knowledge and are specifically designed for electronic commerce purposes.

1 Introduction

Numerous Internet-based business models have been developed recently, of which electronic commerce is playing a key role ([MNB98]). The transition from vast amounts of Internet server and transaction data to actionable marketing intelligence is one of the major challenges in this research field. The conglomerate of quasi standardised log file formats and highly domain-specific marketing data requires a holistic approach to collect, pre-process, and consolidate available web site information, in order to provide flexible materialised views for explorative operations such as online analytical processing or data mining (consequently named *online analytical web usage mining*). The objective of this article is to propose an environment that allows the discovery of patterns from trading-related web sites, which can be harnessed for electronic commerce activities, such as personalisation, adaptation, customisation, profiling, and recommendation.

The approach outperforms recent endeavours in several issues: it views web mining holistically, that is, the entire process from the data organisation and consolidation to the applicability of the discovered knowledge; it supports all data sources in an electronic commerce scenario, not only a problem-specific or technology-orientated subset; it integrates data warehousing and data mining techniques; and it allows a flexible creation of multiple problem-orientated materialised views.

The paper is structured as following. In Section 2 the three types of data sources that can typically be found on electronic commerce-related web sites are briefly described, which cover server data, marketing data, and web meta data. In Section 3, two major data preparation activities are proposed. The first deals with different types of schematic and semantic heterogeneities which arise among entities in such an environment and resolving mechanisms are provided. The latter describes Internet- and electronic commerce-specific pre-processing operations. In Section 4, flexible materialised views are created, which comprise the formal specification of a generic web log data hypercube, as well as schematic designs. In Section 5, a web-enabled knowledge discovery process is proposed and techniques of pattern discovery are described, which cover the typical customer relationship life cycle. In Section 6, related work is evaluated and compared against our approach, before, in Section 7, conclusions are drawn and further work is outlined.

2 Data Sources

We have developed an on-line retailing model, which contains the processes on an Internet retailer web site, as well as the back-end data storage components, which is depicted in Figure 1 ([MNB98]). The on-line retailing model is the permutation selected to research the efficacy of data mining. The primary reasons for selecting the on-line retailing model are behavioural (consumers exhibit a wide range of differing behavioural patterns), marketing-orientated (consumer and product information provides a rich source of

* This research has partly been funded by the ESPRIT project Nº 26749.

useful data), and descriptive (products and services in on-line retail sites are varied).

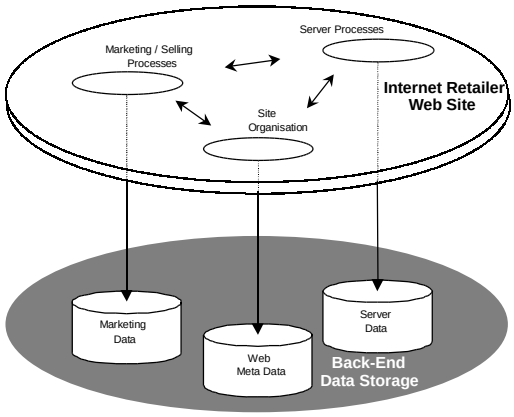

Figure 1. Online Retail Model

The three types of data are described in more detail in the following sub-sections.

2.1 Server Data

Server data is generated by the interactions between the persons browsing an individual site and the web server. The *httpd* process that runs on web servers provides a facility to log information on accesses to the server. That data can be divided into log files and query data.

There are three types of log files, namely server logs, error logs, and cookie logs. Server logs are either stored in the Common Logfile Format or the more recent Extended Logfile Format. The available field information from the two files is shown in Table 1, where the gray shaded cells indicate support in the extended format only.

Field	Description
date	Date, time, and timezone of request
client IP	Remote host IP and / or DNS entry
user name	Remote log name of the user
bytes	Bytes transferred (sent and received)
server	Server name, IP address and port
request	URI query and stem
status	*http* status code returned to the client
service name	Requested service name
time taken	Time taken for transaction to complete
protocol version	Version of used transfer protocol
user agent	Service provider
cookie	Cookie ID
referrer	previous page
…	…

Table 1. Server Log File Formats

Additionally, the Extended Logfile Format supports directives which provide meta information about the log file, such as version, start and end date of session monitoring, as well as the fields which are being recorded.

Error logs store data of failed requests, such as missing links, authentication failures, or timeout problems. Apart from detecting erroneous links or server capacity problems — which, when satisfactorily corrected, can be seen as a compulsory form of customer satisfaction — the usage of error logs has proven to be rather limited for the discovery of actionable marketing intelligence.

Cookies are tokens generated by the web server and held by the clients. The information stored in a cookie log helps to ameliorate the transactionless state of web server interactions, enabling servers to track client access across their hosted web pages. The logged cookie data is customisable, which goes hand in hand with the structure and content of the marketing data (see Section 2.2).

A fourth data source that is typically generated on electronic commerce sites is query data to a web server. For example, customers to an online store may search for products, or clients to a research database may search for publications. The logged query data must be linked to the access log through cookie data and/or registration information. There are currently no formal drafts for standards for handling query data, although new specification suggestions have reached draft stage, for instance Resource Description Framework RDF ([W3C98a]).

2.2 Marketing Data and Knowledge

Any organisation that uses the Internet to trade in services and products uses some form of information system to operate Internet retailing. Clearly, some organisations use more sophisticated systems than others. The least common denominator information that is typically stored is about customers, products and transactions, each in different levels of detail. More sophisticated electronic traders keep also track of customer communication, distribution details, advertising information on their sites associated with products and / or services, sociographic information, and so forth.

A second type of information which is stored in modern environments is marketing knowledge. This type of domain expertise, obtained internally or externally, has usually been formulated by (human or artificial) marketing experts. It can be in the form of target-directed data collated from cross-fertilised sources, or the output of online analytical web mining activities carried out at an earlier stage. More formalised domain knowledge ([ABH95]) is expressed in hierarchical generalisation trees (for example, the topological organisation of Internet domains) or attribute relationship rules (marketing-specific constraints).

The commonality of both sources is its arbitrary nature and hence structure, which has to be considered in each electronic commerce scenario (see Section 3.1).

2.3 Web Meta Data

The last source is data about the site itself, usually generated dynamically and automatically after a site update. Web meta data provides the topology of a site, which includes neighbour pages, leaf nodes and entry points. This information is usually implemented as site-specific index table, which represents a labelled directed graph. Meta data also provides information whether a page has been created statically or dynamically and whether user interaction is required or not.

In addition to the structure of a site, web meta data can also contain information of more semantic nature. Examples are arbitrary or ontology-based content information, usually represented through HTML meta tags or XML ([W3C98b]) statements, the type of a page (root, navigational, content page, or a hybrid thereof), or page scores, which have been derived according to some pre-defined set of heuristics.

3 Data Preparation

The previous section has shown the arbitrary nature of available data sources in an electronic commerce environment. This section deals with the data preparation of that information. The data preparation step contains two major tasks, namely the resolution of schematic and semantic heterogeneities among the relevant data, and a battery of pre-processing activities, some of which are generic, but most are Internet- and electronic commerce-specific.

3.1 Heterogeneity Resolution

3.1.1 Schematic Heterogeneity

Due to the fact that most server data sources are standardised, relatively little schematic heterogeneity occurs in these files. Major discrepancies in log files are different physical arrangements of fields in server and error logs, and different field names recorded by different web servers. These naming mismatches are easily resolved using mapping files. These mapping files are either created manually or by (semi-)automated techniques as recommended in the literature (see for instance [KCGS93]). Cookie logs are either used to join various customer-related information sources or to identify reoccurring events. Depending on the structure of the cookie, the types of schematic heterogeneities that occur are naming conventions and different formats. Naming conflicts are resolved analogous to incompatibilities in server and error log files; formatting conflicts are resolved using parameterised conversion functions. Cookies can only be handled in a canonical form if they have been set up with the intention to do so — there is no possibility to resolve schematic conflicts generically among different cookie types or dangling cookies.

Less standardised and hence more unstructured is query data, since it is usually set up for individual, application- and domain-specific purposes. Depending on the design and the pre-planned activities, the type of schematic heterogeneity is of differing granularity. Assuming that some form of planning has been performed, typical types of discrepancies are naming conflicts and data representation on domain level, and schema isomorphism on entity level (different number of attributes representing the query and its result). Similar to log files, domain level conflicts can be resolved using mappings and conversion functions; deriving missing fields is the most common form to deal with schema isomorphism.

Similarly, type, content and structure of marketing information and meta data depend heavily on the electronic commerce domain, the topology of the site, the logical and physical interconnectivity with other sites (for instance on shopping malls), and so forth. Again, and ideally, in the case where the entire environment has been designed for the purpose of the collaboration of marketing and Internet data, the amount of schematic conflicts to resolve is minimal. But, more realistically, it is more likely that server and meta data has to be logically connected to an already existing marketing data pool. The most typical discrepancies encountered are representation mismatches of marketing information (for instance customer numbers) and log files (cookie identifiers as well as the identification of a specific customer through standard log file information), and also between marketing data and meta data. Most inconsistencies are resolved through mapping tables, whereas some require further pre-processing, which is outlined in further detail in Section 3.2.

3.1.2 Semantic Heterogeneity

After the creation of a conflict-free schema, various interpretations of available data sources are necessary. Typical examples in log files are date and time formats of logs being used in different countries, query information in different languages, or URIs which are interpreted differently in different contexts, for example, the suffix *.com* might represent commercial organisations in one situation, whereas in another context it might include private users as well (for instance, customers who are logged in through a service provider). Similar scenarios exist in marketing

as well as web meta data, especially when data is distributed across different locations.

There exist two general approaches of how to tackle semantic heterogeneity. One is a priori, which requires that the context in which the information will be used is known; the other one is a posteriori, which allows the usage of information in multiple contexts ([BBH98, KS98]). In the context of web mining it is sufficient to follow the first direction, since the usage of the data — discovering marketing intelligence — is well defined, as opposed to running dynamic queries from multiple users against the collected data. Consequently, semantic heterogeneity can be resolved at the data pre-processing stage.

3.2 Pre-processing Activities

In addition to generic data warehousing-like pre-processing activities, some Internet- and electronic commerce-specific operations have to be carried out. First group consists of a set of techniques comprising cleaning, transforming, and aggregating. It is referred to in some of the standard literature (for instance, [BS97, CD97]) for more detailed information. For the purpose of this article, it is focused on the second group.

[CMS99] have created an interesting web mining architecture, which contains a battery of sophisticated pre-processing tasks, each tailored towards a specific knowledge discovery goal (see also Section 6). From server logs, meta data files and some optional usage statistics, a user session file, a transaction file, the site topology, and page classifications are derived. Some activities are based on artificial intelligence techniques, for instance, the page classification is based on rule induction, whereas others use lookup tables, filtering (of multimedia information), and so on.

Many of the outlined pre-processing steps have been adopted, since they are inevitable for successful web usage mining. In order to handle all data sources as outlined in the previous section, essential operations were added. Some of the techniques were expanded in order to handle extended log file entries and customer data were linked to a user session via cookies. Furthermore, some initial work has been carried out in considering ontology-based content information, expressed in XML. Also, a new data type *URL* has been added to the data mining system MKS, which has been developed at the authors' laboratory ([AST+97]). For a complete description of Internet- and electronic commerce-specific data pre-processing activities, the reader is referred to [BMAH98].

4 Creating a Materialised View

After a rather informal description of data sources in electronic commerce scenarios and an arsenal of heterogeneity resolution and pre-processing techniques, the physical and logical design of the data warehouse is presented more formally.

4.1 Web Log Data Hypercube

In order to create a materialised view which is used as repository for further analytical activities, an *n*-dimensional web log data cube is defined.

Definition 1. A web log data hypercube H represents an n-dimensional information space, such that $H = [D_1, D_2, D_3, \ldots, D_n]$, where each D_x represents a dimension of H. □

This hypercube represents an intentionally denormalised materialised view of the pre-processed data.

Definition 2. A dimension D_x represents m attributes such that $D_x = [a_{x_1}, a_{x_2}, a_{x_3}, \ldots, a_{x_m}]$. □

Some attributes may be derived fields in order to hold summarisation information, like resource hits, sales aggregation, or customer purchase sums.

The total number of cells is calculated as $|C| = \prod_{i=1}^{n} |H_i|$, where $|H_i|$ is the cardinality of H (number of attributes). This number leads to an exponential growth of the number of cells, many of which are naturally sparse, which is handled by internal compression mechanism. Also, efficient computation of the web hypercube has to be guaranteed. [HRU96] covers both aspects, which are based on dynamic sparse matrixes; further details are beyond the scope of this paper.

In order to simplify generalisation and aggregation operations as well as the modelling of dimensional hierarchies in snowflake schemas (see Section 4.2), each dimension is defined on a multi-level concept hierarchy ([HF95]).

Definition 3. A concept hierarchy T is an undirected, connected, acyclic graph which is defined as the tuple $T = (L, E)$, where $L = \{l_0, l_{1_1}, l_{1_2}, \ldots, l_{1_{|l_1|}}, l_{2_1}, l_{2_2}, \ldots, l_{2_{|l_2|}}, \ldots, l_{n_1}, l_{n_2}, \ldots, l_{n_{|l_n|}}\}$ and $E = \{e_1, e_2, e_3, \ldots, e_m\}$. Each l_x represents a value of a domain D_x of a hypercube H, such that the granularity $g(l_n) < g(l_{n-1})$, $n > 0$. Each e_k has the form $e_k = \{l_i, l_j\}$; $l_i, l_j \in L$, l_0 has indegree 0, $l_1 \ldots l_n$ have indegree 1. l_i is subconcept of l_j iff $l_i \subset l_j$; l_i is superconcept of l_j iff $l_j \subset l_i$. □

Thus, a hypercube can also be represented as an *n*-tuple $H=(T_1, T_2, T_3, \ldots, T_n,)$. An example of a three-dimensional web log data cube as well as a drilled-down version with the dimensions customer, location and product is depicted in Figure 2.

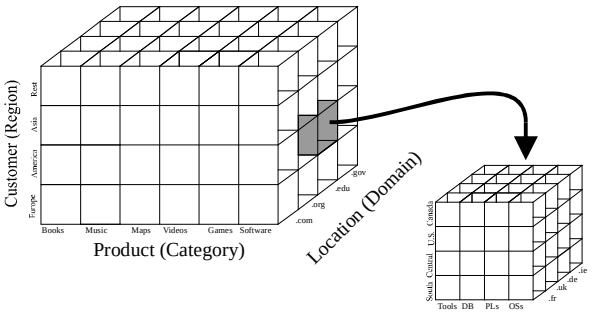

Figure 2. A three-dimensional Web Log Data Cube

4.2 Schematic Design

To construct a cube as depicted above and supporting further analysis activities such as OLAP and data mining, a schema based on the relational calculus is modelled. Each dimension is represented as a relation, which is connected to a fact table. Fact tables act as connecting element in a data model representing keys and summarisation information. This star schema is sufficient for one given set of scenarios, which uses data input of the same granularity ([BS97]). For more advanced operations, as necessary in web mining, a snowflake schema is required, which supports multiple granularities. Snowflake schemas provide a refinement of star schema where the dimensional hierarchy is explicitly represented by normalising dimension tables ([CD97]).

The following figure shows a fact table which is typical for analysis activities in electronic commerce scenarios. It contains key fields (CustomerKey, ProductKey, LocationKey, DateKey, SessionKey) as well as some statistical summarisation information (Quantity, TotalPrice, ClickThroughRate).

CustomerKey
ProductKey
LocationKey
DateKey
SessionKey
Quantity
TotalPrice
ClickThroughRate

Figure 3. A Web Log Fact Table

Around this fact table it is now possible to create a star schema, which is used for simple analysis activities, as well as a basis for the more complex and powerful snowflake schema below.

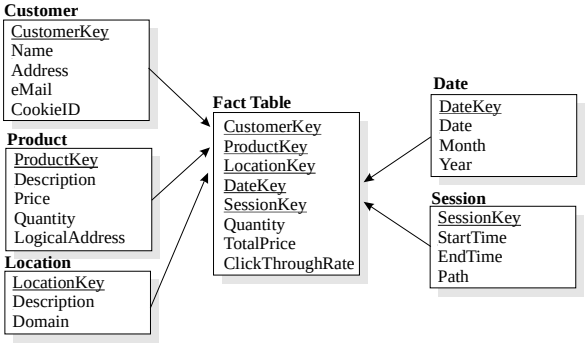

Figure 4. A Web Log Star Schema

Although expressive enough for simple data analysis activities, the star schema is insufficient for complex web usage mining. A more flexible schema is provided below.

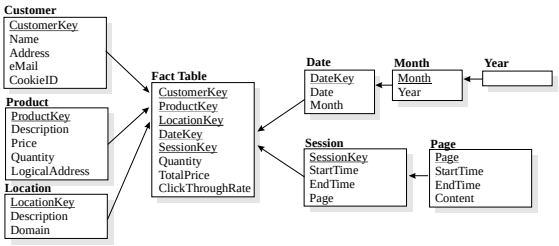

Figure 5. A Web Log Snowflake Schema

The snowflake schema depicted above is just one possible example, which can be used to handle data from Internet sources. Other alternatives are possible, depending on the problem to be solved.

5 Discovering Marketing Intelligence

In order to facilitate the steps carried out above, a web-enabled knowledge discovery process has been developed (see Figure 6), which is an adoption of a generic process defined in earlier work ([AB97]).

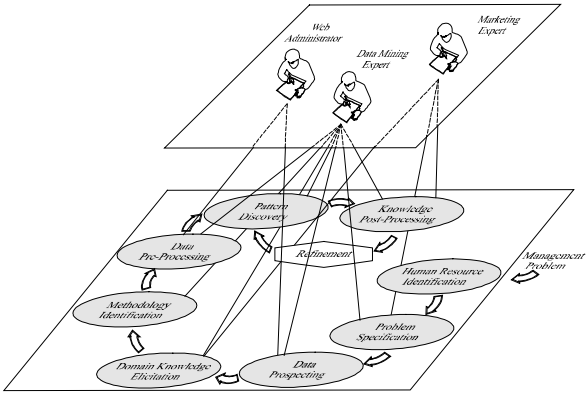

Figure 6. Web-enabled Knowledge Discovery Process

The process contains the entire life-cycle of a knowledge discovery project and involves required

expertise, personified as a web administrator, a marketing expert, and a data mining expert, all of whom are involved in various steps of the process. The entire web-enabled process is described in more detail in [BMAH98]; here it is only focused on the pattern discovery task, which shows the applicability of the proposed concepts.

Marketing experts divide the customer relationship life-cycle into three distinct steps, which cover attraction, retention, and cross-sales. A pattern discovery scenario is presented for each period, each of which covers the discovery goal, marketing strategy, and data mining approach[1].

5.1 Customer Attraction

The two essential parts of attraction are the selection of new prospective customers and the acquisition of the selected potential candidates. One marketing strategy to perform this exercise, among others, is to find common characteristics in already existing visitors' information and behaviour for the classes of profitable and non-profitable customers. These groups are then used as labels for a classifier to discover Internet marketing rules, which are applied online on new site visitors. Depending on the outcome, a dynamically created page is displayed, whose contents depends on found associations between browser information and offered products / services.

The three classification labels used were 'no customer', that is browsers who have logged in, but did not purchase, 'visitor once' and 'visitor regular'. An example rule is as follows.

```
if Region = IRL and
   Domain1 IN [uk, ie] and
   Session > 320 Seconds
then VisitorRegular
Support = 6,4%; Confidence = 67,2%
```

This type of rule can then be used for further marketing actions such as displaying special offers to first time browsers from the two mentioned domains after they have spent a certain period of time on the shopping site.

5.2 Customer Retention

Customer retention or attrition is the step of attempting to keep the online shopper as loyal as possible. Due to the non-existence of distances between providers, this is an extremely challenging task in electronic commerce scenarios. One strategy is similar to that of acquisition, that is dynamically creating web offers based on associations. However, it has been proven more successful to consider associations across time, also known as sequential patterns ([MBNG97]). Typical sequences in electronic commerce data are representing navigational behaviour of shoppers in the forms of page visit series ([CPY96]).

We have extended [AS95]'s a priori algorithm so it can handle duplicates in sequences, which is relevant to discover navigational behaviour. A found sequence looks as following.

```
{
ecom.infm.ulst.ac.uk/,
ecom.infm.ulst.ac.uk/News_Resources.html,
ecom.infm.ulst.ac.uk/Journals.html,
ecom.infm.ulst.ac.uk/,
ecom.infm.ulst.ac.uk/search.htm,
}
Support = 3.8%; Confidence = 31.0%
```

The discovered sequence can then be used to display special offers dynamically to keep a customer interested in the site, after a certain page sequence with a threshold support and / or confidence value has been visited.

5.3 Cross-Sales

The objective of cross-sales is to horizontally and / or vertically diversify selling activities to an existing customer base. We have adopted a traditional cross-sales methodology ([APHB98]), in order to perform the given task in an electronic commerce environment.

In order to discover potential customers, characteristic rules of existing cross-sellers had to be discovered, which was performed through the application of attribute-orientated induction. For a scenario in which the product CD is being cross-sold to book sellers, an example rule is

```
if Product = book then
   Domain1 = uk and
   Domain2 = ac and
   Category = Tools
Support = 16.4%; Interest = 0.34
```

Deviation detection is used to calculate the interest measure and to filter out the less interesting rules. The entire set of discovered interesting rules can then be used as the model to be applied at run-time on incoming actions and requests from existing customers.

6 Related Work

[Etz96] has suggested three types of web mining activities, viz. *resource discovery*, usually carried out by intelligent agents, *information extraction* from newly discovered pages, and *generalisation*. For the purpose of the discussion of related work only the

[1] No methodology-related information (neural networks, rule induction, bayesian belief networks, statistics, etc.) is provided, since this choice is of generic nature and neither Internet- nor electronic commerce-specific.

latter category is considered, since it has the most important impact on electronic commerce research. The two research directions which our approach has successfully combined and extended are web usage mining (driven by Cooley, Mobasher & Srivastava) and online analytical mining (coined by Zaïane, Xin & Han). These endeavours are the major focus of the review. Earlier and less advanced approaches for the discovery of navigational sequences ([MT96, CPY96]) or for visitor clustering ([YJGD96]) are described and evaluated in great detail in [CMS99].

[ZXH98] have applied various traditional OLAP and data mining techniques to Internet log files in order to find different types of patterns, which can be harnessed as electronic commerce decision support knowledge. The process involves a data cleansing and filtering stage (manipulation of date and time related fields, removal of futile entries, etc.) which is followed by a transformation step that reorganises log entries supported by meta data. The pre-processed data is then loaded into a data warehouse which has an *n*-dimensional web log cube as basis. From this cube, various standard OLAP techniques are applied, such as drill-down, roll-up, slicing, and dicing. Additionally, artificial intelligence and statistically-based data mining techniques are applied on the collected data which include characterisation, discrimination, association, regression, classification, and sequential patterns. The system is similar to ours in that it follows the same process of data pre-processing, data warehousing, and the application of online analytical mining. However, the approach is limited in several ways. Firstly, it only supports one data source — static log files —, which has proven insufficient for real-world electronic commerce exploitation. Secondly, no domain knowledge (marketing expertise) has been incorporated in the web mining exercise, which we see as an essential feature. Thirdly, the discovery of web access patterns is designed as a one-way flow, rather than a feedback process, that is the discovered knowledge does not feed back in the data mining process. And lastly, the approach is very data mining-biased, in that it re-uses existing techniques which have not been tailored towards electronic commerce purposes.

[CMS97] have built a similar, but more powerful architecture. It includes an intelligent cleansing (outlier elimination and removal of irrelevant values) and pre-processing (user and session identification, path completion, reverse DNA lookups, et cetera) task of Internet log files, as well as the creation of data warehousing-like views ([CMS99]). In addition to [ZXH98]'s approach, registration data, as well as transaction information is integrated in the materialised view. From this view, various data mining techniques can be applied; named are path analysis, associations, sequences, clustering and classification. These patterns can then be analysed using OLAP tools, visualisation mechanisms or knowledge engineering techniques. Although more electronic commerce-orientated, the approach shares some obstacles of [ZXH98]'s endeavour, which are the non-consideration of all data sources as described in Section 2 the non-incorporation of marketing expertise, and the one-way operation of the architecture.

7 Conclusions and Further Work

An environment has been proposed which combines existing online analytical mining as well as web usage mining approaches, and incorporates marketing expertise. This marketing knowledge is essential in order to perform electronic commerce activities such as personalisation, adaptation, customisation, profiling, and recommendation

We are currently implementing the entire web-enabled knowledge discovery process as described in [BMAH98], which is based on the work proposed in here. First results of trial runs on real-world data, have shown promising results. We are also enhancing some existing pattern discovery algorithms, in order to discover actionable marketing knowledge, which consider multi-level concept hierarchies from the defined hypercubes more explicitely ([HF95, ZXH98]).

Future work includes the incorporation of more sophisticated domain knowledge and syntactic constraints, better transaction support, similar to [CMS99], and the scalability improvement of developed web-enabled data mining algorithms. The work proposed in here exclusively focuses on customer-to-business relationships on the Internet. A (financially) more attractive, but currently less mature discipline, are business-to-business models, which are currently evaluated towards knowledge discovery support.

8 References

[AB98] S.S. Anand, A.G. Büchner. Decision Support through Data Mining, FT Pitman Publishers, 1998.

[ABH95] S.S. Anand, D.A. Bell, J.G. Hughes. The Role of Domain Knowledge in Data Mining, in *Proc. 4th Int'l. ACM Conf. on Information and Knowledge Management*, pp. 37-43, 1995.

[APHB98] S.S. Anand, A.R. Patrick, J.G. Hughes, D.A. Bell. A Data Mining Methodology for Cross Sales, *Knowledge-based Systems Journal*, 10:449-461,1998.

[AS95] R. Agrawal, R. Srikant. Mining Sequential Patterns, in *Proc. 11th Int'l. Conf. on Data Engineering*, pp. 3-14, 1995.

[AST+97] S.S. Anand, B.W. Scotney, M.G. Tan, S.I. McClean, D.A. Bell, J.G. Hughes, I.C. Magill. Designing a Kernel for Data Mining, in *IEEE Expert*, 12(2):65-74, 1997.

[BBH98] A.G. Büchner, D.A. Bell, J.G. Hughes. A Contextualised Object Data Model based on Semantic Values, in *Proc. 11th Int'l. Conf. on Parallel and Distributed Computing Systems*, pp. 171-176, 1998.

[BMAH98] A.G. Büchner, M.D. Mulvenna, S.S. Anand, J.G. Hughes. An Internet-enabled Knowledge Discovery Process, submitted for publication, 1998.

[BS97] A. Berson, S.J. Smith. Data Warehousing, Data Mining and OLAP, McGraw Hill, 1997.

[CD97] S. Chaudhuri, U. Dayal. An Overview of Data Warehousing and OLAP Technology, Technical Report MSR-TR-97-14, Microsoft Research, 1997.

[CMS97] R. Cooley, B. Mobasher, J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web, in *Proc. 9th IEEE Int'l Conf. on Tools with Artificial Intelligence*, 1997.

[CMS99] R. Cooley, B. Mobasher, J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns, in *Knowledge and Information Systems*, 1(1), forthcoming, 1999.

[CPY96] M.S. Chen, J.S. Park, P.S. Yu. Data Mining for Traversal Patterns in a Web Environment, in *Proc. 16th Intl'l. Conf. on Distributed Computing Systems*, pp. 385-392, 1996.

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

[HF95] J. Han, Y. Fu. Discovery of multiple-level association rules in relational databases, in *Proc. 21st Int'l. Conf. on Very Large Databases*, pp. 420-431, 1995.

[HRU96] V. Harinarayan, A. Rajarman, J.D. Ullman. Implementing data cubes efficiently, in *Proc. ACM SIGMOD Int'l. Conf. on Management of Data*, pp. 205-216, 1996.

[KCGS93] W. Kim, I. Choi, S.K. Gala, M. Scheevel. On Resolving Schematic Heterogeneity in Multidatabase Systems, in *Distributed and Parallel Databases* 1(3): 251-279, 1993.

[KS98] V. Kashyap, A. Sheth. Semantic Heterogeneity in Global Information Systems: the Role of Metadata, Context and Ontology, in *M.P. Papazoglou, G. Schlageter (eds). Cooperative Information Systems*, pp. 139-178, 1998.

[MBNG97] M.D. Mulvenna, A.G. Büchner, M.T. Norwood, C. Grant. The Soft-Push: Mining Internet Data for Marketing Intelligence, in *Proc. Working Conf. on Electronic Commerce in the Framework of Mediterranean Countries Development*, pp. 333-349, 1997.

[MNB98] M.D. Mulvenna, M.T. Norwood, A.G. Büchner. Data-driven Marketing, in *Int'l. Journal of Electronic Markets*, 8(3) 32-35, 1998.

[MT96] H. Manilla, H. Toivonen. Discovering generalized episodes using minimal occurences, in *Proc. 2nd Int'l. Conf. on Knowledge Discovery and Data Mining*, pp. 146-151, 1996.

[W3C98a] World Wide Web Consortium. http://www.w3.org /RDF/, 1998.

[W3C98b] World Wide Web Consortium. http://www.w3.org /XML/, 1998.

[YJGD96] T. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal, From User Access Patterns to dynamic Hypertext Linking, in *5th Int'l. WWW Conf.*, 1996.

[ZXH98] O. R. Zaïane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, in *Proc. Advances in Digital Libraries Conf.*, pp. 19-29, 1998.