Available online at www.sciencedirect.com

SCIENCE DIRECT®

Decision Support Systems 41 (2005) 189–204

Decision Support Systems

www.elsevier.com/locate/dsw

An e-customer behavior model with online analytical mining for internet marketing planning

Irene S.Y. Kwan[a,*], Joseph Fong[b], H.K. Wong[b]

[a]Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, Hong Kong, China
[b]Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

Available online 7 April 2005

Abstract

In the digital market, attracting sufficient online traffic in a business to customer Web site is vital to an online business's success. The changing patterns of Internet surfer access to e-commerce sites pose challenges for the Internet marketing teams of online companies. For e-business to grow, a system must be devised to provide customers' preferred traversal patterns from product awareness and exploration to purchase commitment. Such knowledge can be discovered by synthesizing a large volume of Web access data through information compression to produce a view of the frequent access patterns of e-customers. This paper develops constructs for measuring the online movement of e-customers, and uses a mental cognitive model to identify the four important dimensions of e-customer behavior, abstract their behavioral changes by developing a three-phase e-customer behavioral graph, and tests the instrument via a prototype that uses an online analytical mining (OLAM) methodology. The knowledge discovered is expected to foster the development of a marketing plan for B2C Web sites. A prototype with an empirical Web server log file is used to verify the feasibility of the methodology.

Keywords: Customer behavior; Internet marketing; OLAM; Knowledge discovery; Traversal pattern

1. Introduction

In e-commerce, the current challenge is determining how to design responsive Web site infrastructure that provides a sustainable competitive advantage through a better understanding of target customers. The quality of an e-commerce site depends on interrelated factors such as site architecture, network capacity, Web services, and the unpredictability of e-customer behavior. These characteristics imply the need to measure the behavior of the Web-based system and its users. Knowledge management is the key to business learning. The technologies that support knowledge management in e-business are data warehousing, data mining, the Internet, and document management systems [21,25,26].

* Corresponding author.
E-mail addresses: drikwan@ln.edu.hk (I.S.Y. Kwan),
csjfong@cityu.edu.hk (J. Fong),
hkwong@cs.cityu.edu.hk (H.K. Wong).

0167-9236/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.11.012

Online marketing aims to produce online revenue by understanding customer needs. Meeting this objective requires knowledge of how e-customers' online movements change from awareness of products to the exploration of options and further to purchase commitment. An online analytical mining (OLAM) system using an underlying cognitive model and e-customer behavioral graph can be used to articulate the online activities of e-customers on a particular Web site. This can provide the framework of an e-customer behavior (eCB) model that can be used to discover e-customer profiles which identify the significant dimensions of online behavior and identify Web pages that trigger behavior changes. The knowledge thereby obtained will foster informed Internet marketing decision making, and allow Web content and infrastructure refinement to support Internet marketing.

2. Current background, theoretical underpinnings and hypotheses

Electronic commerce (EC) is growing rapidly, and offers a diversity of related issues to investigate. Ngai [18] presents a literature review and classification scheme for EC research. Over 78% of EC research has been focused on applications, implementation and technical issues, and only 9% has touched the topic of e-customers, with very few studies directly addressing the issue of e-customer preferences and their effects on Web site acceptability. Because e-customers learn fast and want Web sites that are driven by their needs, the historical analysis of customer behavior will help to identify current preferences. This paper aims to establish an understanding of e-customers' online behavior, and uses an eCB model to discover this knowledge via OLAM. OLAM has been used in many applications of knowledge management and decision support in e-business. For instance, Menczer [15] proposes an adaptive population of intelligent agents that mine the Web when a query is made. The performance of the system is evaluated by comparing its effectiveness in locating recent and relevant documents with that of search engines. Menczer and Belew [16] propose a multi-agent model for online, dynamic information mining on the Web. Each agent navigates from page to page following hypertext links, trying to locate new documents that are relevant to the user's query, with only limited interaction with other agents. Eirinaki and Vazirgiannix [6] analyze the collected data from content-based filtering, collaborative filtering, rule-based filtering and Web usage mining. Huang et al. applies online analytical processing (OLAP) technology in combination with data mining techniques for the prediction, classification and time-series analysis of Web log data [9]. Ohura et al. [19] clusters user requests from the access logs using an enhanced K-means clustering algorithm, and then applies them for query expansion by recommending categories that are similar to the request and suggesting related categories. Our target is to discover knowledge about the customers of an e-commerce Web site to foster the development of a feasible Internet marketing plan. The elements of our instrument include a three-phase e-customer behavior graph, a mental cognitive model, a set of OLAM algorithms and a webmaster.

2.1. The mental cognitive model

Internet marketing is the process of building and maintaining customer relationships through online activities to facilitate the exchange of ideas, products, and services that satisfy the goals of both parties [17]. It is concerned with using the Internet to create intense and profitable relationships with their customers. Three primary forces that are generated by the Internet effect e-marketing-individualization, information, and interactivity. Marketing research has begun to investigate how these forces can be utilized to create long lasting relationships with customers. The need to understand the target customers of Internet marketing has become obvious recently. Much recent research has investigated the human aspect of computing, with attempts to explore the meaning and effects of computer interface design and its interactive use [3,20]. Moreover, Devaraj et al. [5] examine the determinants of EC channel satisfaction and preference using survey data, and Koufaris [12] examines how emotional and cognitive responses to visiting a Web-based store for the first time can influence revisit intention. Kalakota and Whinston [10] discuss a six-step interactive marketing process on the Internet, in which the initial step emphasizes the study of e-customers' behavioral approaches. Kiang et al. [11] propose a scheme for determining e-product characteristics for Internet marketing. Dalgleish [4] proposes a hypothetical customer theme of common activities that customers want to complete when they visit a particular Web site. Sang and Young [24] examine the relationship between consumers' perceived importance of and satisfaction with Internet shopping. Indeed, Internet marketing has recognized that e-customer advocacy, reactions to stimulated transactions, and sensory, cognitive, and emotional experiences are all crucial in building an understanding of customer experience in the design of appropriate marketing programs for securing customer relationships [17]. We have developed a mental cognitive model to articulate our research hypothesis and significant dimensions to measure e-customers' behavior. Fig. 1 explains the research framework and presents a mental cognitive model that makes use of the four rules of association to provide a basis for quantifying the behavior of e-customers in four dimension—path length by session (PL), access frequency (PF), revisit recency (PR) and duration (PD)—to articulate e-customer behavior. The hypothetical construct is induced from Fact 1, which deduces the associations in Facts 2, 3, 4, and 5 that relate to the four dimensions identified to measure e-customer online behavior.

This mental cognitive model is based upon our study of Thomas Brown's secondary laws of association. Brown (1778–1820) [22,23] proposed nine secondary laws that predict which sensations are most likely to be associated with each other. We select his first, third, fourth, and fifth rules to define the four dimensions of measuring customers' online behavior, and use them to form the four indicators of e-customer behavior: duration (PD), access frequency (PF), revisit recency (PR), and path length (number of associated Web pages) by session (PL). These factors allow us to show the behavioral association of two point and click online movement, and allow quantitative measurement for comparison. The four laws of association are extracted as follows.

1. (1st rule) Association between sensations is modified by the length of time during which the original sensations endured. (PD)
2. (3rd rule) Association between sensations is modified by the frequency of their pairing. (PF)
3. (4th rule) Association between sensations is modified by the recency of their pairing. (PR)
4. (5th rule) Association between sensations is modified by the number of other associations in which the sensations to be paired are involved. (PL)

2.2. The e-customer behavior graph

Human behavior analysis and learning has long experimented with the analysis of choice and preference, and stimulus generalization and respondent conditioning [22,23]. However, Internet marketing has recently led the field of marketing to undertake a fundamental re-examination of its core principles and doctrine [17]. Human behavior analysis in Internet marketing enables the understanding of online experiences: from the initial entry to the homepage and the exploration of related Web pages to the final decision to submit or abandon a shopping cart. It is important to articulate the possible routes that e-customers can travel on a Web site when designing an Internet marketing plan. As a mental model enables individuals to make inference and predictions, to understand phenomena, to decide what action to take, to control execution of that action and to experience events by proxy [14], we model the movement of generic surfers on a Web site by an e-customer behavior graph. Fig. 2 shows how surfers can interact with an e-commerce site through a series of consecutive and related requests made during a single session. Typical requests can be Login, Home Page, Browse, Search, Register, Select, Add to Shopping Cart, or Pay. Different surfers exhibit different navigation paths and request different Web pages in various ways and with various frequencies. Some customers are frequent buyers while others are occasional buyers who browse extensively but seldom commit to buying from the site. We have divided a generic travel path into three phases, e-customer awareness, e-customer exploration, and e-customer commitment, to keep track of changes in online behavior.

We emphasize Internet marketing by categorizing surfer requests into three phases, Awareness, Exploration, and Commitment. Any corresponding requested Web pages reflect the e-customer's behavior and intention.

e-Customer behavior: URL request

Phase 1: Awareness: request entry, home page, browse page.

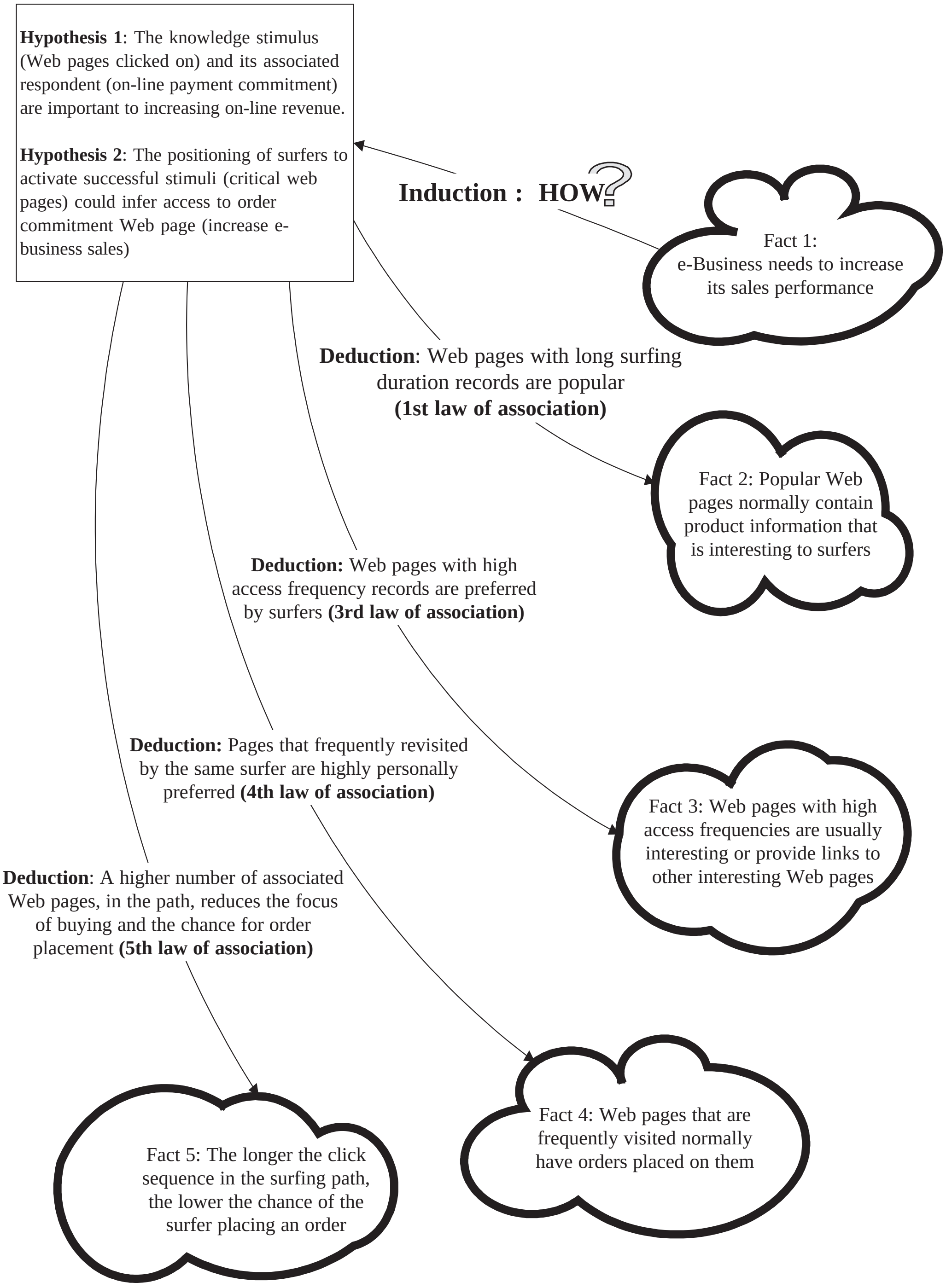

Hypothesis 1: The knowledge stimulus (Web pages clicked on) and its associated respondent (on-line payment commitment) are important to increasing on-line revenue.
Hypothesis 2: The positioning of surfers to activate successful stimuli (critical web pages) could infer access to order commitment Web page (increase e-business sales)
Induction : HOW?
Fact 1: e-Business needs to increase its sales performance
Deduction: Web pages with long surfing duration records are popular **(1st law of association)**
Fact 2: Popular Web pages normally contain product information that is interesting to surfers
Deduction: Web pages with high access frequency records are preferred by surfers **(3rd law of association)**
Fact 3: Web pages with high access frequencies are usually interesting or provide links to other interesting Web pages
Deduction: Pages that frequently revisited by the same surfer are highly personally preferred **(4th law of association)**
Fact 4: Web pages that are frequently visited normally have orders placed on them
Deduction: A higher number of associated Web pages, in the path, reduces the focus of buying and the chance for order placement **(5th law of association)**
Fact 5: The longer the click sequence in the surfing path, the lower the chance of the surfer placing an order

Fig. 1. The mental cognitive model.

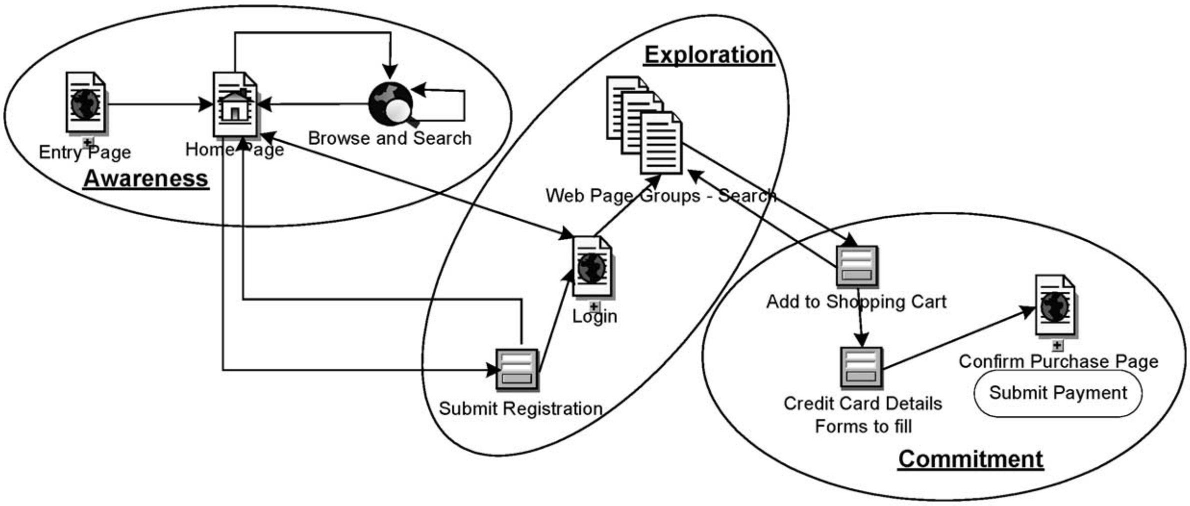


Fig. 2. The e-customer behavior graph.

Phase 2: Exploration: login page, registration page, search page.
Phase 3: Commitment: select page, add to shopping cart page, payment page.

An e-marketing scheme motivates target customers in a process that moves from awareness and exploration to commitment. The eCB model discovers knowledge about contiguous associations between the awareness and exploration, and exploration and commitment phases for identifying the critical Web pages that allow e-customers to progress from phase to phase until the payment Web page.

2.3. Research data and methodology

e-Customer behavior changes as requests move through the awareness and exploration phases to the commitment phase of a surfing an e-commerce Web site. The clicking stream of Web pages represents e-customer behavior within a time frame. The Web pages that trigger e-customer behavioral changes are the critical pages that affect online revenue. The eCB model applies induction and deduction, using association rules and OLAM techniques for continuous discovery of e-customer click and tick sequences, and recognizes individual e-customers by cookie files and IP addresses from the Web server log. The two important of sources of data-cookies and the server transfer log file known as the access log-about Web site visitors are used as our research data. A cookie is a small text file that is placed on user's hard drive by a Web page server. The user identification card that is read by the server gives the cookie to a surfer. The server log file consists of every transaction between the server and browser recorded with a date and time, the IP address of the server making the request for each page on the site, the status of the request, and the number of bytes that were transferred to the requester. We analyze surfers' activities on a Web site using the cookie file and the server log file. As Web servers generate cookies and server log records with valuable identification and continuous data streams for online analysis and querying, these source data are selected for our mining process to discover the association semantics of e-customer tick sequences on Web pages. The association semantic of click sequence patterns on Web pages provides the top referring Web pages that trigger changes of an e-customer's behavior. Hence, our eCB model discovers (1) customer segmentation into customer and very important customer groups, (2) critical Web pages that affect the click direction and sequence, and (3) individual customer's profiles in terms of their preference for particular online characteristics. This knowledge offers us a better understanding of customer profiles, and will allow the redesign of Web site infrastructure to incur positive voluntary clicks from e-customers by directing them to the purchase commitment phase.

This paper proposes an eCB model, which includes the hybrid use of four measuring dimensions of online behavior derived from our mental cognitive model, the three-phase e-customer behavior graph, and the OLAM system, to discover e-customer profiles, behavior, and preferences. It aims to collect each e-customer's online movement by keeping track of their click sequences using cookies and Web log records, and to identify the linkage between an order commitment and the previous associated Web pages. The correlation pattern that is traced provides a valuable base for the analysis of the relationships between stimuli and the resulting clicks.

3. The OLAM methodology

Our OLAM system for path traversal patterns includes incremental Web usage mining updates [8,13,27,28]. It stores the derived Web user access paths in a data warehouse. The system updates the user access path pattern in the data warehouse by data operation functions that are automated by webmaster. The result is an OLAM that uses the underlying e-customer behavior graph, which is capable of discovering association semantics between tick sequences, e-customer profiles for customer segmentation, and a set of preferred and referring Web pages for analysis that will allow the development of effective Internet marketing plans. In addition, the eCB model has a self-learning capability. It allows webmaster to input confidence and support level values to adjust the knowledge that is generated. These revised factors provide adjusted values to the Web mining algorithm to regulate its mining process according to e-customers' behavioral changes. The association rules that are discovered are constantly reviewed to reflect the actual online travel situation. Fig. 3 depicts an overall conceptual architecture of our OLAM methodology.

To extract transaction activities for the discovery of association rules, we load desired data fields from the server log file as a text file into a relational table for further processing. Log data are updated to generate the association rules to discover knowledge for marketing decision support in Web site design. We capture the e-customer experience via the Web server created log records, and present them in continuous data streams. We then extract a useful data pattern so that we can discover stimuli factors that infer e-customer behavioral changes. The patterns of click-

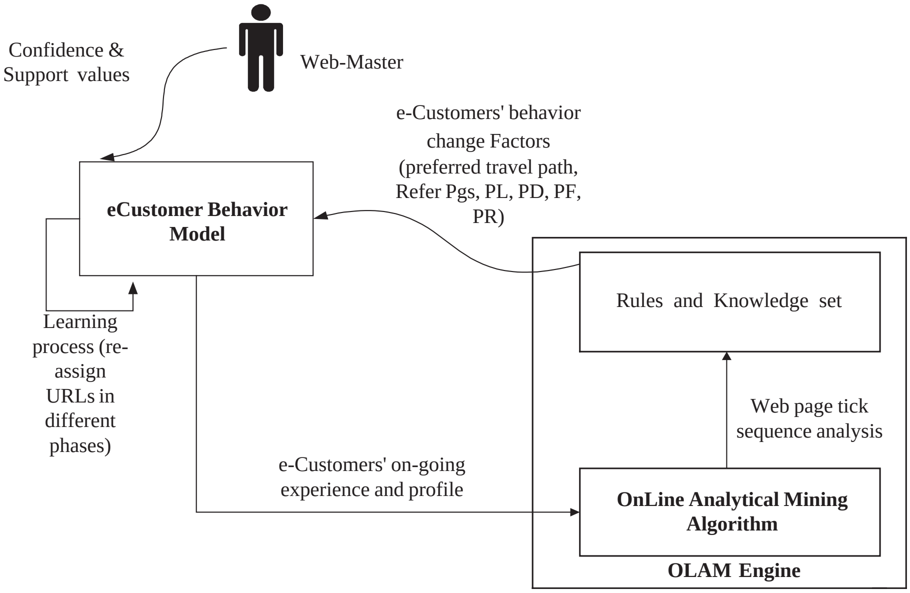


Fig. 3. The OLAM methodology for the internet marketing support.

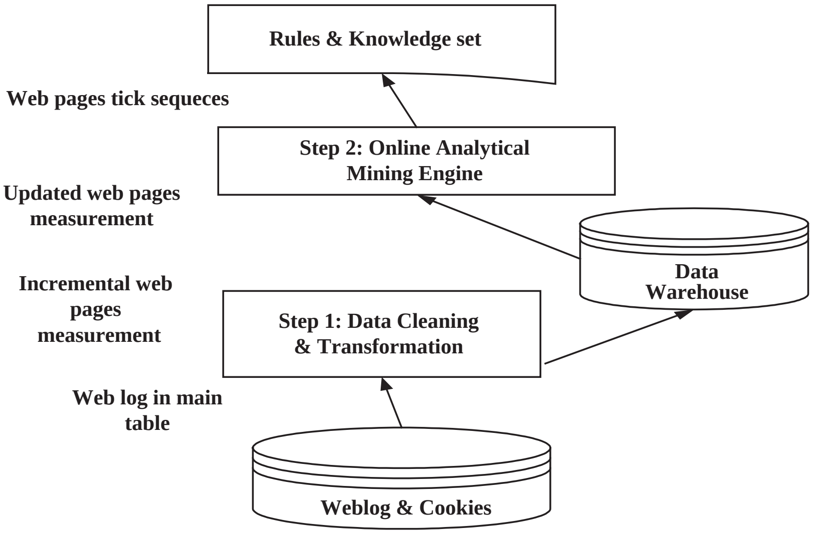

Rules & Knowledge set
Web pages tick sequeces
Step 2: Online Analytical Mining Engine
Updated web pages measurement
Data Warehouse
Incremental web pages measurement
Step 1: Data Cleaning & Transformation
Web log in main table
Weblog & Cookies

Fig. 4. The OLAM engine.

stream data that are extracted are analyzed to clarify how users traverse the site from page to page, and identify the items that they select, the patterns of repeated visits, and the end-result of visits. This pattern analysis identifies trends in consumer browsing and purchase behavior that allows the comprehensive profiling of Web site visitors.

We discover five types of data patterns, including (1) the path content: the sequence of pages by which the user traverses the site; (2) the length of path (PL): the number of pages in a complete surf, a successful surf with purchase and an unsuccessful surf; (3) the click actions (PF): the frequency of clicking on each Web page; (4) time-related information (PD): the log-in time and the stay duration on each page; and (5) repeat visits (PR): whether the user comes back several times during a short period. This is accomplished by the eCB model, which keeps a record of e-customer log data patterns. For instance, when an e-customer spends much time looking at item descriptions, a preference for the item is indicated. When a customer leaves a page instantly, a lack of preference for the item is indicated. These patterns reveal knowledge that is useful in anticipating e-customer behavioral changes.

3.1. The OLAM procedure

In the OLAM engine, the data that is collected from the Web log goes through two steps: data cleaning and transformation, and the OLAM process. Fig. 4 shows the architecture of our overall OLAM procedure.

The data is filtered to remove irrelevant information in Step 1. All entries of the log files are mapped into a relational database. After the first step, the Web log is loaded into a relational database and new implicit data such as the time spent by each visitor on each page is computed. The database facilitates information extraction and data summarization based on individual attributes such as domain name, user location, and date. In Step 2, the data mining engine extracts interesting correlations. After the Web log has been loaded into the data warehouse, whenever Web page access is recorded in the Web log file, a

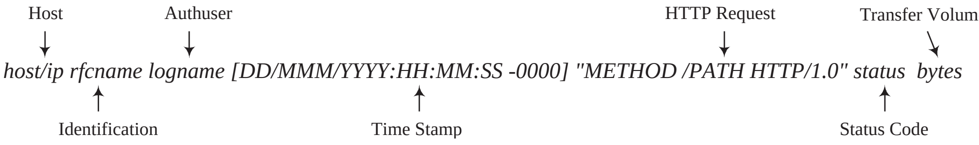

Host
Authuser
HTTP Request
Transfer Volume
host/ip rfcname logname [DD/MMM/YYYY:HH:MM:SS -0000] "METHOD /PATH HTTP/1.0" status bytes
Identification
Time Stamp
Status Code

Fig. 5. Common log format.

```
begin
  open access log;
  do until end of file
    read a line from the log file;
    if ((access_method = “get” or “GET”) and (status = 200))
        if (path = “htm” or path = “html”)  // exclude the graphics, pictures
            use space as a separator to identify the field and determine array size
            do {
                  for (int i = 0; i <= array size; i ++) {
                        if (it is scanning the last field)
                              store it in the array’s last element;
                        else{
                              if (the field starts with “[“ ){ // handling the time format
                                    get the length of this field;
                                    store the content after “[” up to “:” character to array;
                                    i = i + 1;
                                    store the content from “:” to the end of this field to array;
                                    reset the colon index value to 0; }
                              else if (the field end with “]”){ // handling the case of “+0800]
                                    storing content exclude the character “]” to array; }
                              else if (the field starts with ““” ){ // handling the “GET” method
                                    storing content exclude the character of “ to array; }
                              else if (the field ends with “”” ){ // handling the case of HTTP/1.0"
                                    storing content exclude the character of “ to array; }
                              else
                                    storing field content fully to the array; }
                          update field position value;
                  } // end for
            } // end do
            write record to main table; }
     else
       skip a line; }
 end do
end
```

Fig. 6. Algorithm for data preprocessing.

corresponding update is made in our path log by user, thus triggering the update of the Web page tick sequence, and generating Web page association rules.

3.2. The pre-processing of data sources

Data cleaning is an important step of knowledge discovery in data preprocessing [1,7]. As not all materials within the log file are relevant to the mining, a data preparation process is performed first. We focus on preprocessing the two server-level Web access log files, namely the common log format access log and the cookies file. The common log format is presented in Fig. 5. The content of the cookie record varies in length and format, and acts as a user identification card. The log entries must be partitioned into logical clusters using one or a series of transaction identification modules, including user and session identifications.

3.2.1. Step 1.1: data loading and cleaning

The Web access log is a plain text file. Each field is separated by a space. Only successful file-retrieval log

IP Address	Date	Time	URL Request
144.214.36.91	07/May/2001	22:42:04	A.htm
144.214.36.91	07/May/2001	22:45:06	B.htm
144.214.36.91	07/May/2001	22:49:15	D.htm

Fig. 7. Cleaned web log data.

```
begin
    open cleaned log;
    open cookies file;
    do until end of file
            read first line from cookies file
            extract the IP address , userID and timestamps
            replace the IP address(cleaned log) with userID (cookies) in cleaned log
            go to next line
    end do
end
```

Fig. 8. Algorithm for user identification.

records with special characters removed are stored as clean data sources for mining. Fig. 6 presents the algorithm for this data preprocessing.

After the removal of irrelevant records from the Web log file, valid records are stored in the main table, as shown in Fig. 7.

3.2.2. Step 1.2: user identification and session identification

After preprocessing the Web server log, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. In an ideal scenario, each user is allocated a unique IP address when accessing a Web site. Some users access the Web from a different machine or Web browser each time. In the absence of such information, server cookies are the only choice that is available to identify a unique user. Cookies capture the IP address, user name and the timestamps when a user surfs a Web site. Consequently, we use the user name from cookies instead of the IP address from the server log to identify a user in a navigation session, which is a session that includes all page references made by a user during a single visit to a Web site. Identifying user sessions is similar to the problem of identifying individual users. User interaction within a Web site acts as a collection of user navigation sessions, the information of which is recorded in the Web server log. A user session is defined as a sequence of requests from the same IP address such that no two consecutive requests are separated by more than 30 min [2]. Fig. 8 shows the pseudo code for user identification.

3.2.3. Step 1.3: data warehousing

After the removal of irrelevant records from the Web log, we derive the user and session identifications that are stored in the main table. The user access path records are stored in the fact table of the data warehouse, and the corresponding dates are stored in the dimension table. The algorithm for recording user access paths into the data warehouse is shown in Fig. 9.

```
Given: view V, data to be updated into data warehouse view $\delta R$, and data warehouse view V' after update

begin
        for record added in log (cleaned log)
                extract desired data fields (user id, path, timestamps, duration);
                if access path exists
                        increment the frequency pattern by 1;
                else
                        add the new user access path into fact table;
                end if
        end for

        /* V' = V + Applied Group by on $\delta R'$ with aggregate, re-computing total count and aggregate count */
        if $\delta R$ comes from updates to fact table destination relation
                $V' = V \cup \delta R'$;
        end if
end
```

Fig. 9. Algorithm for recording user access paths into the data warehouse.

3.3. The OLAM process

This section presents the second step of the OLAM engine, as depicted in Fig. 4.

Step 2: Data mining

OLAM is used to discover e-customer preferences and popular Web pages based on the traversal patterns that are kept in the data warehouse. To ensure the accuracy of our knowledge, we apply support and confidence levels, which are two measures of rule applicability. Support decides how much the knowledge discovered is supported by the source data. Confidence describes the degree believability of the rules or knowledge which is discovered after the mining process. Together they reflect the usefulness and certainty of discovered rules. We calculate four different aspects of the confidence level: duration, which indicates preference; access frequency, which indicates popularity; revisit frequency, which implies personal interest; and associated page count per path, which allows us to deduce the normal length of a single session that leads to the targeted page. The following examples provide the support level of whether the candidate is a frequent visitor and confidence level of whether the candidate is an important customer.

$$\text{Support Level} = \frac{\text{no. of times that a particular customer has logged in}}{\text{total no. of successful login customers}} \tag{1}$$

$$\text{Confidence Level} = \frac{\text{no. of payments of a particular customer}}{\text{no. of times that the customer has logged in}} \tag{2}$$

We store the record count of the target attribute and its associated attributes as fact relations arranged in the order of date, session ID and user ID sequence in the following.

Fact table: destination relation R_{FACT}

Date	Session ID	User ID	Page visited	Web-page count (Access Freq.)	Duration
Date1	Session-id	UID	P1	C1	D1

The four stimuli factors (PD, PF, PR, and PL) are derived and measured from the above cleaned fact relations, and organized in sequence by date and user ID as follows.

3.3.1. Duration of web page visit requested

$$\text{Duration } (A_1) = \text{Time of request } (A_2 - A_1),$$

where A_2 is the Web page that is requested after A_1 by the same user ID in a single session.

$$\text{PD}(A_1) = \frac{\text{Duration}(A_1)}{(\sum \text{Duration}(A_1 \cap A_2 \cap \ldots A_n)/\text{No. of web} - \text{pgs in the path} A)} \tag{3}$$

```
If Support level >1 then
    page $A$ is preferred and put Time($A_1$)and $A_1$ in UID(preferred page) table;
else
    $A$ is not preferred and put Time($A_1$)and $A_1$ in UID(not-preferred page) table;
endif
```

The preferred and non-preferred Web pages are identified in the following table format. The duration for each page is arranged in descending sequence, with

```
begin
        select date
        for each record of the specific date
                compare each URL in access path to the URL template of the Web site
                extract the desired URL and perform mapping;
                if Web page exists in the access path match the URL in the Web site template
                        then increment the frequency count of the $URL_n$ by 1 into Frequency count table
                        and add 1 to the associated Web-pg Count(path-id);
                         compare the next URL in the same path until end of path;
                        read next access path in fact table until end of file;
                end if
        end for
end
```

Fig. 10. Algorithm for web page access frequency and associated web page count /in a session.

its corresponding URL, and both fields are repeating fields, where Date, UID, and URL are composite keys.

Date	User ID	Web-Page (URL)	Total time (duration)
Date1	UID	P1	T1

3.3.2. Web page access frequency and associated web page count in a session

Fig. 10 shows the algorithm that keeps track of the frequency count of each accessed Web page and the number of associated Web pages in a single session within a specified time frame regardless of the user ID value.

The frequency count table format is presented below, and date and URL are composite keys. A URL with high access frequency must be compared with the above tables on total duration. High access frequency and long visit duration reconfirm it as a valuable URL with high interest to e-customers. In contrast, high access frequency with a short visit duration reconfirms it as an unpopular URL that needs further investigation into its design and resources allocation.

Date	Web-Page (URL)	Access Frequency Cnt.
Date1	P1	C1

The associated Web page per access path count has the following table format, where date, user ID and session ID are the composite keys. The URL is not of concern here, as we focus on session length only and involve the calculation of the average number of associated Web pages per session by single e-customer. Knowledge of the average session length of successful paths (sessions in which online purchases are made) by each important e-customer provides valuable indicators for marketing.

Date	User-id	Session ID	No. of Assoc.-pg. Cnt.
Date1	UID	Session-id	C2

3.3.3. Revisited web pages

The revisit pattern of a particular URL by each important e-customer is significant knowledge that can support customer personalization and preferences.

Date	User-id	Web-Page (URL)	Revisit-Cnt.
Date1	UID	P1	C3

Fig. 11 is our Web mining algorithm that discovers the four aspects of e-customer online behavior.

4. Prototype

We verify the eCB model by using a prototype. Section 4.1 provides an outline of our OLAM process with data flow diagrams. Section 4.2 explains the prototype system. The Web log file was collected from the Computer Science Laboratory Web site of the City University of Hong Kong. The site hosts a variety of information, ranging from departmental information and courses to individual Web sites. We identified seven pages for our empirical study.

Page 1 (P1):	Department history, facilities, and message from department head
Page 2 (P2):	News, events, and seminar notifications
Page 3 (P3):	Listing of academic staff
Page 4 (P4):	Listing of programmes available by department
Page 5 (P5):	Research groups, research projects, publications, etc.
Page 6 (P6):	Academic staff information
Page 7 (P7):	Research student information.

More details of this prototype, its theoretical and technical underpinnings can be found at www.ln.edu.hk/is/staff/drikwan/eCB.

4.1. Outline of the methodology

Our prototype stores the derived Web user access paths in a data warehouse. The system updates the user access path pattern continuously. The data warehouse is analyzed in terms of the time, frequency, recency and number of associated Web page factors in the Web site within a set period. Fig. 12 shows the data flow diagram of the prototype.

4.2. Running guide and system tests

Fig. 13 shows the main menu of the prototype. It consists of three constructs: initialization, analysis, and online update scheduling.

```
begin
    webmaster input date range, support and confidence thresholds that they desired to analyze
    open source relation, read each line subject to select criteria (date range) until EOF;
    aggregate and summaries all data retrieved in group of userID
    if computed support level ≥ minimum requested support threshold & computed confidence level ≥ minimum
            requested confidence threshold
      then aggregate and store the result in the temporary table
    end if

    // compute trigger pages (awareness, exploration, commitment) by taking most frequently visited trigger pages
    use array(1 to 7) to store each trigger pages count
    select all navigation path where  userID = 'input value'
    do while not EOF
        identify all trigger pages in a navigation path
        increment the count of the trigger page in the corresponding array index
        move next path
    end do

    identify all trigger pages by comparing the count in the array set

    // compute PL: taking average number of page on a buying path using SQL statement
    select page, frequency_count from temporary table  userID = 'input value'
    do while not EOF
        total_page_count = page * frequency_count
        move next
    end do

    average = total_page_count / no of payment_count

    // compute PF: no log login path that lead to purchase commitment
    select purchase_count from temporary table where userID = 'input value'

    // compute PR: higher revisit frequency of particular page
    select frequency_count from temporary table where userID = 'input value' order by frequency_count Des
    get the greatest count of pages

    // compute PD: longest surfing duration of a particular page
    select page, duration from fact_page where userID = 'input value' order by duration Des
    get the longest surfing duration page

    // classify the user type
    if count(purchase) / count(login ) >= 80% then
        user classify as VVIP type
    else if count(purchase) / count(login ) >= 60% then
        user classify as VIP type
    end if

   List the result on the screen. (User Name    Trigger Page    PL,  PF,  PR,  PD,    Class Type)
end
```

Fig. 11. The knowledge discovery algorithm.

The initialization construct consists of three major functions: Initial Setup of Open Log File and Cookies; Step 1: Data Loading and Cleaning; Step 2: Extracting and Rule Generation. In Step 1, we place the access log and cookie files in text format into the main table. This step includes data loading, cleaning, and user and session identifications. In Step 2, the main table that is created in Step 1 is transformed into a fact table. At the same time, a summary of the page count, the user access path count, and a statistical summary are calculated and stored. After executing the initialization phase, a user access statistical summary is generated for analysis.

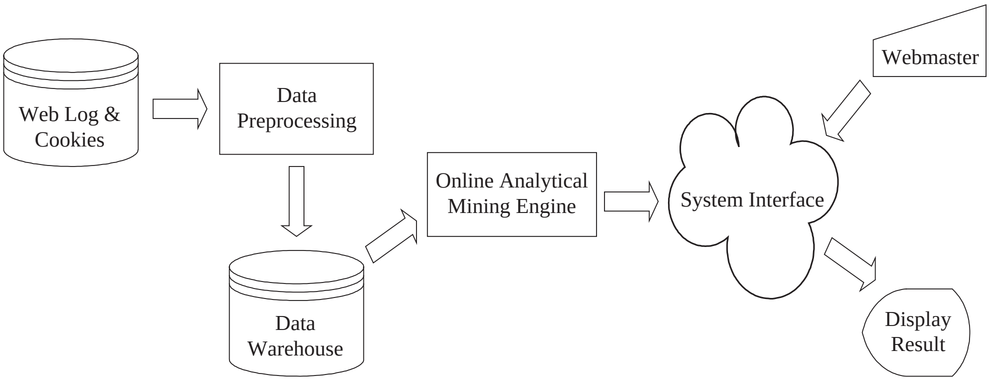


Fig. 12. Data flow of the prototype.

In the second constructthe analysis of the prototypethe program asks the user to specify several parameters before building a knowledge set and statistical summary. Firstly, the user selects the time range and the two thresholds values: i.e., the support and confidence levels. Clicking on the GO button results in a set of potential e-customer names under the set conditions. Analysts obtain their detailed summaries by selecting the user name. Other than the trigger URL of each phase by e-customer, the four behavior indicators of each important customer (VIP) or more important customer (VVIP)Path length (PL), Log on Path Frequency (PF), Page Revisit Frequency (PR), and highest Page Duration (PD) during the specified periodare generated to foster the development of an Internet marketing plan. All relevant user statistics are displayed in order of VVIP and VIP. Fig. 14 displays the results of one query. The customers whose login sessions counts in terms of purchase commitment are greater than 80% are classified as VVIP, while those between 60% and 80% are classified as VIP.

Fig. 15 presents the third construct of our prototypethe time scheduling menuwhere the user sets the time condition that triggers the update in the data warehouse whenever a user navigation path is recorded in the Web log file. Consequently, an up to date knowledge set is maintained. The system provides four options for the webmaster to set the time

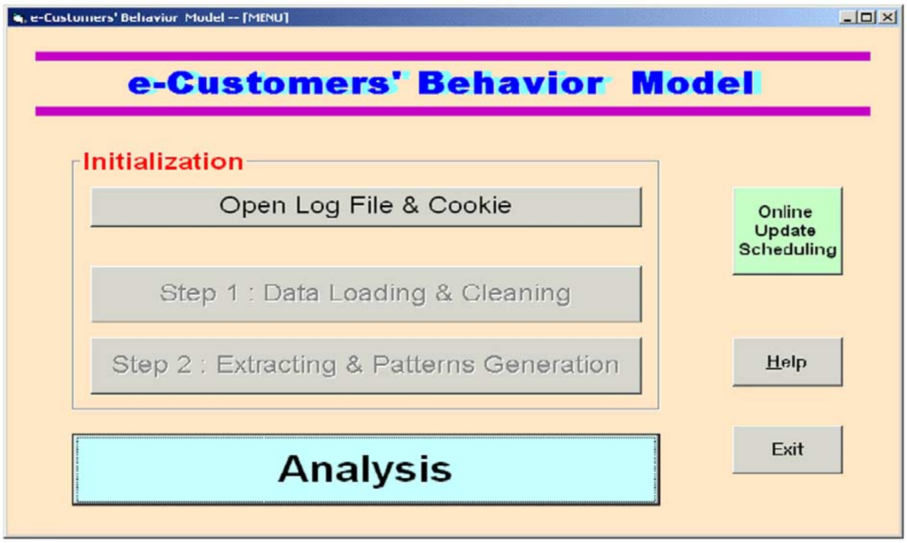
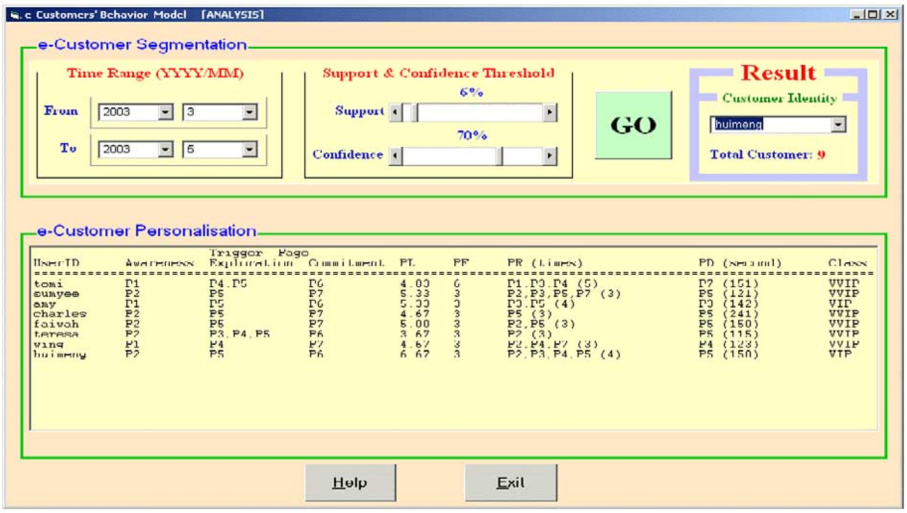

e-Customers' Behavior Model -- [MENU]
e-Customers' Behavior Model
Initialization
Open Log File & Cookie
Step 1 : Data Loading & Cleaning
Step 2 : Extracting & Patterns Generation
Analysis
Online Update Scheduling
Help
Exit

Fig. 13. Main menu.


e Customers' Behavior Model [ANALYSIS]
e-Customer Segmentation
Time Range (YYYY/MM)
Support & Confidence Threshold
GO
Result
e-Customer Personalisation
Help
Exit

From: 2003 / 3
To: 2003 / 5
Support: 6%
Confidence: 70%
Customer Identity: huimeng
Total Customer: 9

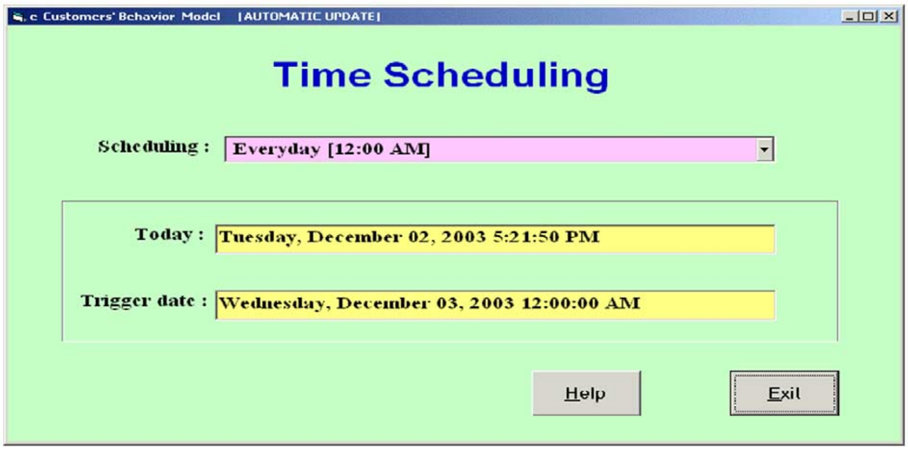
```
                       Trigger   Page
UserID      Awareness  Exploration  Commitment  PL     PF   PR (times)            PD (second)   Class
=====================================================================================================
tomi        P1         P4.P5        P6          4.00   6    P1.P3.P4 (5)          P7 (151)      VVIP
sumyee      P2         P5           P7          5.33   3    P2,P3,P5,P7 (3)       P5 (121)      VVIP
amy         P1         P5           P6          5.00   3    P3.P5 (4)             P3 (142)      VIP
charles     P2         P5           P7          4.67   3    P5 (3)                P5 (241)      VVIP
faivoh      P2         P5           P7          5.00   3    P2,P5 (3)             P5 (150)      VVIP
teresa      P2         P3,P4,P5     P6          3.67   3    P2 (3)                P5 (115)      VVIP
wing        P1         P4           P7          4.67   3    P2,P4,P7 (3)          P4 (123)      VVIP
huimeng     P2         P5           P6          6.67   3    P2,P3,P4,P5 (4)       P5 (150)      VIP
```

Fig. 14. Result set.

scheduling: every hour, every day (12:00 noon), the first day of every week (Sunday, 12:00 noon), and the first day of every month (12:00 noon).

5. Conclusions and future research

This paper proposes an eCB model that uses an OLAM methodology to discover e-customer behavioral changes on a Web site to support Internet marketing. We undertake an empirical study using a prototype built upon the eCB model to verify its feasibility. Our prototype identifies different customer segments according to their successful path frequencies, counting sessions that result in purchases, the referring Web pages that determine targeted e-customer behavioral changes, and, based on the correlation semantic discovered between clicks in sessions from e-customers' click histories, the most popular paths taken by target e-customers. Moreover, typical trigger URLs that lead to positive progression toward purchase commitment are recovered. These patterns represent e-customer behavioral changes over the three phases of our e-customers behavior graph. In summary, we have proposed an instrument to measure the four dimensions of e-customer behavior: average path length (PL), log on path



Scheduling : Everyday [12:00 AM]
Today : Tuesday, December 02, 2003 5:21:50 PM
Trigger date : Wednesday, December 03, 2003 12:00:00 AM

Fig. 15. Time scheduling (Everyday [12:00AM]).

frequency (PF), page revisit frequency (PR) and page duration (PD). This knowledge is valuable in fostering informed decision making for effective Internet marketing plans and understanding the e-customers.

Despite the promising result, our research findings could be further refined. More dimensions could be defined to measure the online movement of e-customers. The OLAM instrument could be enhanced to include the analysis of changes of access patterns over different durations, which will allow the analysis of e-customer behavioral evolution.

Acknowledgments

The authors acknowledge the valuable and constructive comments that were received from the editor and the three referees. We are indebted to them for their invaluable advice, which has greatly improved the quality of this paper.

This work was greatly supported by the University Grant Committee of Lingnan University of Hong Kong (grant code: DR03A5); Irene Kwan, Department of Computing and Decision Sciences, Lingnan University, Hong Kong.

References

[1] S. Anahory, D. Murray, Data Warehousing in the Real World, Addison Wesley, 1997, pp. 327–339.

[2] L. Catledge, J. Pitkow, Characterizing browsing behaviors on the world wide web, Computer Networks and ISDN Systems 27 (6) (1995) 1065–1073.

[3] C. Condon, M. Perry, R. O'Keefe, Denotation and connotation in humancomputer interface: the save as command, Behaviour and Information Technology 23 (1) (2004) 21–31.

[4] J. Dalgleish, Create customer-effective e-services, e-Business Advisor, Technology Strategies for Business Innovators (2000) 26–32.

[5] S. Devaraj, M. Fan, R. Kohli, Antecedents of B2C channel satisfaction and preference: validating e-commerce metrics, Information Systems Research 13 (3) (2002) 316–333.

[6] M. Eirinaki, M. Vazirgiannix, Web mining for web personalization, ACM Transactions on Internet Technology 3 (1) (2003) 1–27.

[7] J. Han, M. Kamber, Data Mining Concepts and Techniques, 2000.

[8] E. Han, G. Karypis, V. Kumar, Scalable parallel data mining for association rules, ACM (1997) 277–288.

[9] Z. Huang, J. Ng, D.W. Cheung, M.K. Ng, W.K. Ching, A Cube Model For Web Access Sessions And Cluster Analysis, Proceedings of the Mining Log Data Across All Customer Touch Points Workshop, Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001.

[10] R. Kalakota, A.B. Whinston, Frontiers of Electronic Commerce, Addison Wesley, 1996, pp. 215–292.

[11] M.Y. Kiang, T.S. Raghu, K.H.M. Shang, Marketing on the Internetwho can benefit from an online marketing approach?, Decision Support Systems and Electronic Commerce 27 (2000) 383–393.

[12] M. Koufaris, Applying the technology acceptance model and flow theory to online consumer behavior, Information System Research 13 (2) (2002) 205–223.

[13] F. Masseglia, P. Poncelet, M. Teisseire, Web Usage Mining: How to Efficiently Manage new Transactions and New Clients, Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'2000), September, 2000, pp. 530–535.

[14] R. McLeod, G. Schell, Management Information Systems, Prentice Hall, 2000, p. 333.

[15] F. Menczer, Complementing search engines with online web mining agents, Decision Support Systems 35 (2003) 195–212.

[16] F. Menczer, R. Belew, Adaptive retrieval agents: internalizing local context and scaling up to the web, Machine Learning 39 (23) (2000) 203–242.

[17] R.A. Mohammed, R.J. Fisher, B.J. Jaworshi, A.M. Cahill, Internet Marketing-Building Advantage in a Network Economy, International edition, McGraw-Hill/Irwin MarketspaceU, 2002, pp. 203–313.

[18] E.W.T. Ngai, F.K.T. Wat, A literature review and classification of electronic commerce research, Information and Management 39 (2002) 415–429.

[19] Y. Ohura, K. Takahashi, I. Pramudiono, M. Kitsuregawa, Experiments on Query Expansion for Internet Yellow Page Services Using Web Log Mining, Proceedings of the 28th Very Large Database (VLDB) Conference, 2002.

[20] A. Oulasvirta, P. Saariluoma, Long-term working memory and interrupting messages in humancomputer interaction, Behaviour and Information Technology 23 (1) (2004) 53–64.

[21] S.C. Park, S. Piramuthu, M.J. Shaw, Dynamic rule refinement in knowledge-based data mining systems, Decision Support Systems 31 (2001) 205–222.

[22] D. Piere, F. Epling, Behavior Analysis and Learning, Prentice Hall, 1995, pp. 262–323.

[23] H. Rachlin, Introduction to Modern Behaviorism, W.H. Freeman and Company, 1970, pp. 10–19.

[24] Y.K. Sang, J.L. Young, Consumers' perceived importance of and satisfaction with internet shopping, Electronic Markets 11 (3) (2001) 148–154.

[25] M.J. Shaw, C. Subramaniam, G.W. Tan, M.E. Welge, Knowledge management and data mining for marketing, Decision Support Systems 31 (2001) 127–137.

[26] D.J. Skyrme, Capitalizing On Knowledge From E-business To K-Business, Butterworth Heinemann, 2001, pp. 3–5.

[27] J. Srivastava, R. Cooley, M. Deshpande, P.N. Tan, Web Usage Mining: Discovery and Application of Usage Patterns from Web Data, SIGKDD Explorations 1 (2) (2000) 12–23.

[28] S. Sunita, T. Shiby, A. Rakesh, Integrating Association Rule Mining With Relational Database Systems: Alternatives and Implications, ACM (1998) 343–354.

Dr. Irene S.Y. Kwan is an associate professor of the Computing and Decisions Sciences Department at Lingnan University of Hong Kong. She is also the Assistant Director of business programmes in her University. She received her PhD degree from Brunel University of London in 1999. Kwan has published over 30 research papers in key journals and conferences. Her research interests are in intelligence business, knowledge management, knowledge discovery and data mining.

Dr. Fong is an associate professor of the Computer Science Department at City University of Hong Kong. He received his PhD degree from University of Sunderland in 1993. Fong has worked in the electronic data processing in US for more than 10 years. His industrial expertise and research interests are in database, data warehousing, data mining and XML. He has published more than 30 academic journal papers in these areas.

Hing Kwok Wong is a PhD student in the Computer Science Department of City University of Hong Kong. He received a BSc in Information Technology, with First Class Honors, in 1999 and a M.Phil degree in Computer Science, in 2001 at City University of Hong Kong. He has published several papers in journals and conferences. His current research interests are online analytical mining, web usage mining, and XML-enabled database.