Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 1893–1904
Expert Systems with Applications
www.elsevier.com/locate/eswa

Web usage mining with intentional browsing data

Yu-Hui Tao [a,*], Tzung-Pei Hong [b,1], Yu-Ming Su [c,2]

[a] Department of Information Management, National University of Kaohsiung, 700 Kaohsiung University Road, Nan-Tzu District, Kaohsiung 811, Taiwan, ROC
[b] Department of Electrical Engineering, National University of Kaohsiung, 700 Kaohsiung University Road, Nan-Tzu District, Kaohsiung 811, Taiwan, ROC
[c] InfoChamp System Corp., 11F, No. 6, Ming-Chuan 2nd Road, Chien-Cheng District, Kaohsiung, Taiwan, ROC

Abstract

Many researches have developed Web usage mining (WUM) algorithms utilizing Web log records in order to discover useful knowledge to be used in supporting business applications and decision making. The quality of WUM in knowledge discovery, however, depends on the algorithm as well as on the data. This research explores a new data source called intentional browsing data (IBD) for potentially improving the effectiveness of WUM applications. IBD is a category of online browsing actions, such as "copy", "scroll", or "save as," and is not recorded in Web log files. Consequently, the research aims to build a basic understanding of IBD which will lead to its easy adoption in WUM research and practice. Specifically, this paper formally defines IBD and clarifies its relationships with other browsing data via a proposed taxonomy. In order to make IBD available like Web log files, an online data collection mechanism for capturing IBD is also proposed and discussed. The potential benefits of IBD can be justified in terms of its enhancing and complementary effectiveness, which are illustrated by the rule implications of Web transaction mining algorithm for an EC application. Introducing IBD opens up the scope of WUM research and applications in knowledge discovery.

Keywords: Web usage mining; Intentional browsing data; Web log files; Browsing behaviour; Knowledge discovery

1. Introduction

Data mining focuses on the techniques of non-trivial extraction of implicit, previously unknown, and potentially useful information from very large amounts of data (Agrawal & Mehta, 1996). In relation to this, the rise of Internet technology specifically promoted Web mining by applying data mining techniques to Internet data (Joshi, Joshi, Yeti, & Krishnapuram, 1999). Among Web mining categories, Web usage mining (WUM) addresses mining Web log records (Han & Kamber, 2001) which typically include the host name or IP address, remote user name, login name, date stamp, retrieval method, HTTP completion code, and number of bytes in the file retrieved. Several techniques exist in data mining: association rule (Agrawal & Srikant, 1994; Park, Chen, & Yu, 1997), classification (Mehta, Agrawal, & Rissanen, 1996; Yu, 1999), cluster (Perkowitz & Etzioni, 2000), sequential pattern (Agrawal & Srikant, 1995), and time series (Mannila & Ronkainen, 1997). However, the ultimate purpose of WUM is to discover useful knowledge from Web users' interactive data in order to fulfill business goals by means of strategies like marketing and customer relationship management or services. To our observations, this ultimate purpose is achieved mainly from the advancements in WUM algorithms, which are worthy of further exploration.

From the perspective of decision support systems (DSS), their quality relies on their components which are

* Corresponding author. Tel.: +886 7 5919220; fax: +886 7 5919328.
E-mail addresses: ytao@nuk.edu.tw (Y.-H. Tao), tphong@nuk.edu.tw (T.-P. Hong), swimming@icsc.com.tw (Y.-M. Su).
[1] Tel.: +886 7 5919191.
[2] Tel.: +886 7 5350101x543.

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.02.017

either the user interface, model base, or the database (Holsapple & Whinston, 1996). At most, algorithms represent the model-base side of contributions, while the database content is equally important but is underlooked in WUM. For example, conventional database marketing utilizes different perspectives of data sources such as demographics, psychographics, and technographics (Modhal, 1999) in the Internet era to understand consumers. Also, Mittal and Lassar (1987) differentiated personalization from customization based on the involvement of users' interactions, which implies the importance of each user's interaction. Similarly, traditional human-computer interaction also considers data as raw as keystroke-level actions (Newman & Lamming, 1995). As a result, WUM may probe into minor details such as key press, button click, or the scroll bar of browser interface components, which represent the Web user's cognitive process not captured in Web log files. These data elements, containing hidden user's intention, are valuable to WUM for broadening or enriching the mining base, and these eventually lead to better DSS performance even without inventing new algorithms.

Since no references related to online browsing data exist in WUM, how to apply online browsing data into WUM for potentially better effectiveness is an issue worth studying. This research addresses the limited sources of data which originate from Web log files for WUM by providing a cornerstone research of intentional browsing data (IBD). Theoretically conjecturing, the IBD may become an effective new component of WUM research and applications.

2. Literature review

In order to build up a common ground for discussing mining sources, an in-depth review of WUM is presented in Section 2.1 for appropriately inferring WUM data sources in Section 2.2, followed by existing online data collection methods in Section 2.3.

2.1. Web usage mining

According to Han and Kamber, applying data mining techniques to Internet data comprises a new area of Web mining. However, the Web also poses new challenges to effective data sourcing and knowledge discovery due to its size, the complexity of Web pages, its dynamic nature, the broad diversity of user communities, and the low relevance of useful information (Han & Kamber, 2001). Consequently, Web mining has been developed into three categories, including Web structure mining that identifies authoritative Web pages, Web content mining that classifies Web documents automatically or constructs a multilayered Web information base, and WUM that discovers users' access patterns of Web pages (Cooley, Mobasher, & Srivastava, 1999). From the data-source perspective, both Web structure and Web content mining target the Web content, while WUM targets the Web access logs.

WUM includes three major processes: pre-processing, data mining, and pattern analyzing (Cooley et al., 1999). Pre-processing performs a series of processing on Web log files, covering data cleaning, user identification, session, session identification, path completion, and transaction identification. Then mining algorithms are adopted for generating association rules, which is also called basket analysis. An association rule represents the relationships among items such that the presence of some items in a transaction tends to imply the presence of some other items. Sequential patterns sort out regularly occurring rules that are similar to association rules, but with item sets ordered by time stamps.

How can WUM be useful in practice? Cooley et al. (1999) proposed a user browsing behavior model which assumes that a given user's treatment of each page is either for the purpose of 'navigation' or 'actual content,' and this is determined by the page references and associated time obtained in Web server logs. Another example is that in the work of Cunha and Jaccoud (1997) who attempted to determine a Web user's next access by defining two types of users based on the navigation strategy: the "net surfer" who is interested in exploring the cyberspace and the "conservative" user who is concerned with the contents of a certain site. These two examples illustrate that raw Web log records are used by WUM algorithms to infer the browsing behavior or usage implications in an application domain. Aggregately speaking, studies utilizing Web log records have been conducted in analyzing system performance (Iyengar, MacNair, & Nguyen, 1997), improving system design by Web caching (Bonchi et al., 2001) and Web page prefetching (Cunha & Jaccoud, 1997), identifying best places for Web advertisement (Chen & Shen, 2000), predicting best browsing paths (Chen, Park, & Yu, 1998; Hong, Lin, & Wang, 2002; Hsieh & Chang, 2001; Kitsuregawa, Shintani, & Pramudiono, 2001; Yun & Chen, 2000), forming novel approaches for efficient search engines (Zhang & Dong, 2002), and building adaptive Web sites (Perkowitz & Etzioni, 2000).

WUM embeds a relationship among browsing data, browsing behavior, and WUM application, which is similar to the relationship of data, information, and knowledge in knowledge management. In other words, the Web log data are processed by WUM algorithms which will become useful information like browsing behaviors or patterns, and these are then applied in practice by knowledgeable workers.

2.2. Data sources for WUM

The major advantage of Web server logs is their availability and convenience, but they cannot accurately lock individual or recognize interactive Web pages with dynamic contents, and thus may lead to biases in analyses. Other mining data are rare but do exist, such as adding items purchased into the traditional browsing paths in Web transaction mining (WTM) (Yun & Chen, 2000).

WTM has pushed WUM applications closer to practical usage by predicting a customer's purchasing pattern with the Web traversal path. For more effective personalization, Modhal (1999) combined usage and content mining, since "usage-based personalization can be problematic when little usage data is available pertaining to some objects or when the site content changes regularly." Therefore, WUM can be more effective with extra data sources other than Web log records for personalized applications.

Cooley et al. (1999) proposed a general architecture for WUM, in which the data come from Web server logs, referral logs, registration files, index server logs, and document and usage attributes. Although these references shed light on valuable data as a key to the effective practice of WUM, raw browsing data are still incomplete. Catledge and Pitkow (1995) then described Web browsing strategies, in which all possible interface events were listed. The actions in function menus and tool bars, such as BACK, Home Document, Print, Hot list, and even Close Window were recorded. Statistics will then show the most favorite actions users did through these recorded events. Maglio, Campbell, Barrett, and Selker (2001) proposed a framework for developing an attentive information system, in which users' behaviors on local machines were all collected by integrating keyboard input and eyeball direction for judging users' intention and making appropriate suggestions to meet the attentive goal. These two references support a preferable direction of browsing data in suggesting users' intention in effective WUM.

2.3. Online data collection methods

Cooley et al. (1999) considered data quality as a research direction, including techniques of data collection, data integration, and data grouping. In addition to Web log files, there are two methods for collecting browsing data, namely, page conversion and agent systems. Page conversion switches execution privilege to the recording server before entering the clicked page, and switches back when the recording is done (Lin, 1997). The advantage of this method is that parameters can be passed between servers, dynamic contents can be recognized, and individual users can be locked. However, the downside is that the screen may freeze during switching and cannot return to the original Web page if the recording server encounters problems. An agent system, which automatically records each user's browsing history and informs the recording server, is a better approach (Chan, 1997; Fann, 1999) since the users will not be affected during recording. However, the bandwidth is occupied in this way. Thus, the adopted agent system must be simplified to avoid screen delay. Nevertheless, these two approaches still capture similar data as in the Web log files, except for the clean and accurate data corresponding to the needs for analyses.

Both works from Catledge and Pitkow (1995) and Maglio et al. (2001) adopted a collection mechanism as a specific program installed and running on client machines. This is, however, not a norm in the Web mining domain due to the difficulties in making an ad-hoc netter to install such programs for privacy concerns.

3. Research methods

The objective of this research is to extend the applicable WUM scope by changing the unanimous usage of Web log records. IBD, though not formally defined but certainly a part of the browsing data, is the new component adopted in this research to address this issue of WUM in three ways:

1. Define IBD and propose a taxonomy of browsing data as a cornerstone work for future improvements on WUM algorithms and applications.
2. Discuss a feasible mechanism for collecting online browsing data as a supporting evidence for bridging the availability gap of IBD to be easily adopted in WUM applications.
3. Demonstrate the benefits of IBD on WUM using the WTM algorithm (Yun & Chen, 2000), which combines the effects of purchasing behavior with traditional travel patterns of customers. There are two practical issues in applications of WTM. First is fine-tuning its capability in predicting the purchasing behavior of a potential customer for more effective promotion strategies and performance; and second is addressing the pages without items purchased in the travel paths for further segmenting the potential merchandise interesting to the customers.

4. Intentional browsing data

To provide a basic understanding of IBD, we define IBD in Section 4.1, propose a taxonomy of browsing data in Section 4.2, describe an online data collection mechanism in Section 4.3, and justify the potential benefits of IBD in Section 4.4.

4.1. Definition of intentional browsing data

Assuming that browsing data refer to a user's actions on a Web browser while surfing the Internet, then IBD can be defined as a user's Web browsing actions whose goal or motivation is not obvious or certain to a Web administrator, but from which useful information or implications hidden could be derived and interpreted. For example, "copy," "scroll," or "save-as" can be recorded online and analyzed together with Web log data for probing the real intention of users, which may be related to future Web site strategies and interactions with users.

IBD is compared to Web log files from two perspectives. First, their items are different. A Web log file is a predefined format of file, containing data such as time, page name, and so on, and is automatically obtained within the Web server. In contrast, IBD includes items not

available in Web log files, such as "copy" and "print," which require additional programming efforts and usually reveal no direct mining effects. Nevertheless, if IBD is combined with other mining data, different or more precise results may be obtained.

Second, an inclusion relationship exists between browser events and IBD items. Browser events contain IBD items, but not all browsing events are IBD items. For those browsing events whose focuses are not on the body of the browser window, the intention may be out of concern or capability to capture. For example, "open file" is a precise browsing event, but reveals unclear users' intention if without the filename or even the contents of the file, which is out of the focal scope of the browser window. Therefore, a parameter of file names needs to be included in mining algorithms in order to really grasp the purpose of viewing an existing file, which is not considered an IBD item in our context.

4.2. Taxonomy of browsing data

Browsing data present a flexible classification for their taxonomy. In this research, browsing data are classified into three categories: precise browsing data, IBD, and other browsing data, as shown in Fig. 1. Precise browsing data include most browsing data used in current WUM, which are further divided by the data items that are or are not included in Web log files. The items not in Web log files include users' basic data and transaction data. An example would be data filled out by a user in a browser form, which are precise but not recorded in Web log files.

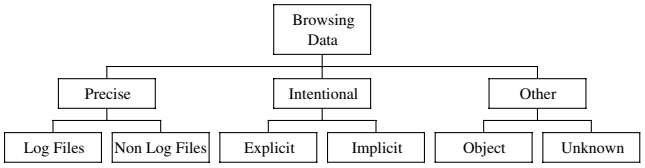


Fig. 1. Taxonomy of browsing data.

IBD includes menus, tool bars, and short-cut commands. Similar to precise browsing data, IBD is divided into two subcategories. One is explicit IBD, such as "add to my favorite," which clearly shows the preference of a user on this Web page; the other is implicit IBD, such as "copy," which requires additional information like "page name" or "page subject" in order to clearly judge the intention of a user. Strictly speaking, explicit intentional data are determined by the target of the browsing action, which must be the Web page itself and not involving another non-browser object, such as another user, partial content of the Web page, uncertain-purpose text input, and browsing direction or destination.

A browsing data item falls into the "other" category if it is not classified into the above two categories, which are also divided into two subcategories. Object browsing data refer to data not directly linked to Web browsing information and whose intention is known only if provided with other parameters or data, such as the "open file" described in Section 4.1. Unknown data refer to one that cannot be classified into any of the previous subcategories.

Very often, a single instance of IBD cannot fully represent the interest of a user. Thus, aggregated data items which consist of multiple instances of the first three categories are important in analyzing users' interests or behaviors. For example, a user's interest is often represented by either the occurrences or the duration of viewing a particular Web page as Aggregated Sample #1 in Table 1. Yet, the combination of both can better represent the user's interest on that Web page. Moreover, adding both "scrolling" or "copying" IBD items to browsing time as Aggregated Sample #2 in Table 1 may further strengthen the accuracy of the result. Examples of the browsing data are provided in Table 1.

Table 1
Examples of browsing data in taxonomy

Classification			Examples of browsing data
Precise browsing data	Non-log file items	User data	Member ID, Name, Sex, Phone number, Birth date, Address, Social security number
		Transaction data	Transaction ID, Transaction time, Merchandise, Quantity, Cost, Payment type
Intentional browsing data	Log file items explicit intention		User IP, User's Browser Info., Web page name, Duration, Entering time, Leaving time message board, Historical records, Email, Homepage, FTP, Add to Favorite, Hyperlink, Refresh, Editing
	Implicit intention		Chat, Copy, Select, Scroll Bar, Search, News Group, Key-in Subjective Content, Back, Forward
Other browsing data	Object		Open File, Save As, Online Help, View original code
	Unknown		Internet Option
Aggregated browsing data		Transaction	Transaction Volume, Number of transactions, Transaction value
		Action	Key word, Occurrences of copy, Content of copy
		Aggregated Sample #1	Occurrences of Web Page, Browsing time
		Aggregated Sample #2	Browsing time, Scroll Bar, Copy

One objective of this research is to provide a formal differentiation of IBD and other browsing data so that this new data source can be clearly understood and used in WUM-related research or applications.

4.3. Online data collection mechanism

With the defined concepts of browsing data and IBD, an immediate issue prior to developing related WUM techniques is how these IBDs can be collected online. In responding to the above question, this research proposes an online data collection mechanism with a goal to support Web site management in a flexible and friendly way for realizing such an operation.

Fig. 2 demonstrates the proposed online data collection mechanism via four blocks *A*–*D*. Web site management starts with block *A* for Web log file setup and with block *B* for browsing data setup. Normally, the Web log file in Block *A* is the default in the Web server. Thus, unless changes are needed, no action is required. However, Web site management has to act on Block *B* for configuring additional browsing data. Block *C* contains the metadata of Web structure and internal database to be used in configuring the desired browsing data. Built-in programs can thus be automatically inserted into designated Web pages for capturing data into browsing data or internal databases during an execution period. Block *D* includes the final online data (in browsing data and internal databases) and log files collected from Blocks *A* and *B*.

The online data collection mechanism depicted in Fig. 2 is an ideal environment, but practical issues exist in its implementation which are as follows:

1. Member vs. non-member: If a user's identity is known, the data collected from this member user can be stored in a database with a login user ID. Otherwise, the collected data can only be stored with the available attributes of IP or session id, which does not guarantee belonging to the same non-member user. In Fig. 2, the input and output of the internal database only make perfect sense for a member user, since otherwise, no internal data are available for any non-member user.
2. Database redundancy: Most Web sites have their own databases, which presents an overlapping and compatibility issue to the browsing-data database. Therefore, how to effectively integrate existing internal databases with a newly added browsing-data database deserves an insightful study.
3. External hyperlink: The data collection module only collects browsing data on the authorized Web site, and stops once the user clicks on a hyperlink leading to external unauthorized Web sites.
4. Client-side program installation: Some client-site programs need to be installed (Bonchi et al., 2001) on local computes in order to collect all possible browsing data, especially intentional browsing data. Users may not accept it.

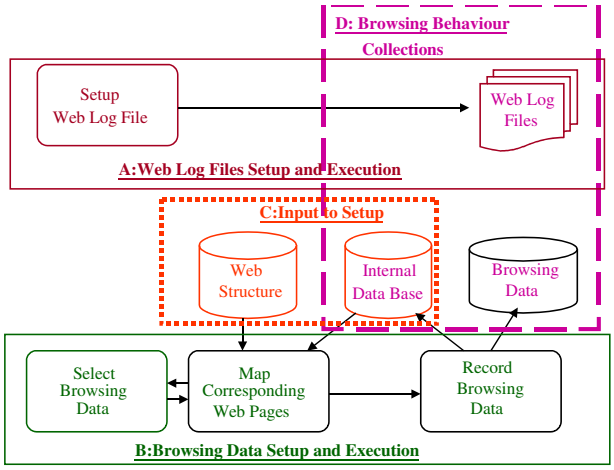


Fig. 2. Online data collection mechanism.

Among the four abovementioned issues, client-site program installation may be the most critical one that affects the feasibility of such an online data collection mechanism, while the rest are more likely technical issues that can be resolved by techniques themselves. Due to privacy and security concerns, majority of users may not agree to install client-side programs that are designed to capture privacy information, especially to those who are just visitors browsing through a Web site. Two alternatives to this problem are proposed below from the perspectives of the client-server side:

1. Client-side module: An authorized client side module can be locally installed for those users who will cooperate with the Web site. The odds are more likely to be on membership-like Web sites with strong customer bonding and trust. This type of module is completely feasible with a typical representative of WebLogger written in Visual Basic working well with Internet Explorer (Reeder, Pirolli, & Card, 2000).
2. Server-side module: Under most circumstances, the data collection module needs to be installed on the server side, which can either be a partial-functional or a full-functional module depending on the Web site management.
   a. Partial-functional module: This module collects only browsing data that can be recorded by existing server-side (back-end) programs, such as ASP and JSP, and client-side (front-end) programs, such as JavaScript and VBScript, together with HTML form components such as menus, buttons, text, text area, and selection.
   b. Full-functional module: In addition to the partial-functional module, other browser events that can only be collected from the client-side browser, such as the "other browsing data" in Table 1, needed to be resolved. A creative alternative is to replace the client-site browser by a server-generated simulated

browser for the user to use. Then the users may be convinced to be operating normally over the simulated server-controlled browser, in which the browsing data are recorded in the background by the Web server. The full-function module is technically feasible, although some events still cannot be captured due to the limitation of client-site event components. The shortcomings are that the transmission may be slower due to simulating the browser screen for recording the events, and approval from the user is preferred.

We had successfully implemented the partial-functional module with some additional features indicated in the full-functional module by replacing some of the client's functional menus on Internet Explorer (Tao, Chung, Chung, Kao, & Yang, 2004).

4.4. Benefits of IBD

Theoretically speaking, IBD can potentially enhance or complement the effectiveness of WUM. Its actual benefit level, however, depends on the application domain, the applied WUM algorithm, and the way WUM is implemented. Justifications of the two potential benefits are described as follows.

Benefit 1: IBD may enhance an existing WUM algorithm by providing incremental effectiveness over Web log records.

If a WUM algorithm can generate a maximum effectiveness E_w via Web log records, or E $(=E_w + E_i)$ with additional IBDs, then E may be manipulated to equal E_w by making $E_i = 0$ in the worst case scenario, or to be greater than E_w otherwise.

A similar situation can be found in normal database-marketing practice like in the credit-card industry in which no methods can provide any valid customer segmentation for effective marketing results ($E = E_w$). Under this worst scenario, any new source of customer related data becomes a potential solution (E_i). Through an analysis, the potential benefit can be determined by some small market tests before implementing a formal marketing campaign. The worst case is that the test results are insignificant, no campaign will be conducted ($E = E_w$), and the market does not get a significant negative effect due to this new source of data ($E_i = 0$).

The above example can fit into WUM applications assuming that the marketing activities are shifted to the Internet environment. The example in Section 5.3.1 demonstrates this benefit.

Benefit 2: IBD may complement an existing WUM algorithm on an application to which Web log records cannot contribute.

If a WUM algorithm cannot contribute to an application at all due to the natural limitation of the Web log records (i.e., $E = E_w = 0$), then IBD may help create new effectiveness ($E = E_i > 0$) by associating IBD with the Web log records in any non-worst scenario, or create no extra effectiveness ($E = E_i = 0$) otherwise.

Current WUM algorithms usually use browsing occurrences and/or durations on Web pages as the key input data. A commonly known problem is that if a user leaves the computer for a period of time while browsing a Web page, the Web server still records the amount of time on that specific Web page, which may cause inaccuracy in applications such as e-Learning that requires accurate learning history ($E = E_w = 0$). If some IBD, such as copying the page content, selecting an option, or scrolling the window, is recorded, the analytical accuracy may be raised ($E = E_i >= 0$).

The above instance shows that an insignificant IBD may significantly contribute to real-world WUM practice. The example in Section 5.3.2 demonstrates this benefit.

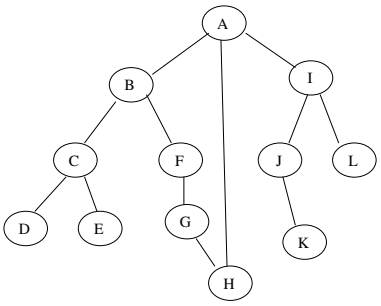
5. Demonstration of IBD benefits with the WTM algorithm

As explained, IBD should be able to contribute to Web-based applications, such as EC, e-Learning, or CRM (Customer Relationship Management). An EC example of the WTM algorithm introduced in Section 5.1 is modified to accommodate IBD, which is described in Section 5.2, for illustrating the beneficial implications of IBD in Section 5.3.

5.1. An EC example of WTM

The complete notations, definitions, and implication rules of WTM can be referred to in the original paper (Yun & Chen, 2000). Here, only what is necessary for understanding the descriptions and discussions in the remaining sections is covered. The basic assumptions are that one Web page can have only a single merchandise item for sale, and each Web page can record only one of the many IBD types acted by users. Situations without these assumptions can also be handled but are more complex. Let $N = \{n_1, n_2, \ldots, n_{p-1}, n_p\}$ be a set of Web pages on a Web site, $I = \{i_1, i_2, \ldots, i_{m-1}, i_m\}$ be the merchandise items sold on the Web site, and $B = \{b_1, b_2, \ldots, b_{p-1}, b_p\}$ be the set of IBDs corresponding to N, where p and m are none-zero positive integers and need not be the same value. These notations will be used to represent the basic elements in the algorithm in the remaining paper.

Fig. 3 illustrates a Web page structure and related browsing data, where $A, B, \ldots, L$ represent the Web page names, and A is the root page without any merchandise



Node	Item	Browsing Type
A	-	-
B	i_1	b_1
C	i_2	b_2
D	i_3	b_3
E	i_4	b_4
F	i_5	b_5
G	i_6	b_6
H	i_7	b_7
I	i_8	b_8
J	i_9	b_9
K	i_{10}	b_{10}
L	i_{11}	b_{11}

Fig. 3. An illustration of Web page structure and corresponding transaction data.

item. Each page is assigned an IBD type to aid in data analysis. The numbers of browsing each page and their associated IBD-type action are used in the mining process. Therefore, $C\{i_2^7, b_2^2\}$ represents all users browsing Web page C who purchased totally 7 items of i_2 with 2 occurrences of b_2-type IBD, and $E\{i_4^0, b_4^3\}$ represents all users who purchased nothing at page E but had three occurrences of b_4-type IBD.

The WTM algorithm mines transaction patterns without IBD as shown below.

Step 1: Sort the records by user numbers, either IP or membership number, in ascending order.
Step 2: Generate a candidate set of 1-transaction patterns C1, and calculate the occurrences of purchased items in each Web page without repetition.
Step 3: Set the hurdle of support values, and save all the C1 items whose occurrences are greater than or equal to the hurdle value into large 1-transaction patterns T1, which represent possible browsing paths for purchasing one item over a preferred hurdle value.
Step 4: Generate a candidate set of 2-transaction patterns C2 by joining two items in T1.
Step 5: Save all C2 items whose occurrences are greater than or equal to the hurdle value into large 2-transaction patterns T2.
Step 6: Repeat Steps 4 and 5 until no large k-transaction sets can be generated.

The WTM algorithm is mainly for predicting possible merchandise purchases with discovered transaction patterns. For instance, a rule of $\langle ABFG: B\{i_1\}, G\{i_6\}\rangle$ implies that a user who browses through ABFG path may purchase item i_1 on Web page B and item i_6 on Web page G. However, WTM may generate many unqualified candidate transaction patterns, thus degrading its performance. Therefore, Yun and Chen (2000) proposed two more algorithms, MTS_{PJ} and MTS_{PC}, to modify WTM. MTS_{PJ} used the concept of path trimming to reduce computational overhead, while MTS_{PC} further added large-count filtering into purchase combination for reducing unqualified patterns. Since effectiveness and not efficiency is the main concern in this research, the original WTM is thus adopted for illustration in the following sections.

5.2. Intention-based WTM (IWTM)

As indicated in Section 4.4, there exist two potential benefits of IBD on WTM. The first direction is to enhance the prediction power of the transaction patterns with users' interest levels on Web pages and merchandise items by the occurrences of certain IBDs. The corresponding algorithm $IWTM_p$ is described in Section 5.2.1. The second direction is to complement WTM by covering no-purchase Web pages in WUM scope via IBD. The corresponding algorithm $IWTM_{np}$ is described in Section 5.2.2.

5.2.1. $IWTM_p$

$IWTM_p$ differs from WTM introduced in Section 5.1 only in Step 2. It generates a candidate set of 1-transaction patterns C1 from Step 1. For each user, Step 2 counts only one for repeated purchases of the same items, but records the exact occurrences of IBDs. In cases when one user has the same IBD in different paths, it takes the minimum value. It then takes the maximum among all users for the IBD on each Web page (Theoretically speaking, a minimum value can be taken instead under a very conservative requirement, which is not considered in this research). The following example based on the Web structure in Fig. 3 is given to illustrate the $IWTM_p$ algorithm:

Step 1: Order the transaction records in ascending order of ID. Assume that the sorted data are listed in Table 2, which includes the browsing path, purchased items, and IBD.
Step 2: Generate the pattern candidate set C1 from the pages with purchases. For example, four users in Table 3 purchased item i_1 on page B. Even though users 2 and 4 both had multiple purchases on page B, the purchase count is still 1 for each of them.

Table 2
An example of transaction data

ID	Path	Purchase and browsing data
1	$ABCE$	$B\{0, b_1^5\}, C\{i_2, b_2^3\}, E\{i_4, b_4^1\}$
	$ABFGH$	$B\{i_1, b_1^2\}, F\{0, b_5^4\}, G\{0, b_6^3\}, H\{i_7, b_7^1\}$
	$AIJK$	$I\{0, b_8^2\}, J\{i_9, b_9^1\}, K\{0, b_{10}^4\}$
2	$ABCE$	$B\{i_1, b_1^2\}, C\{i_2, b_2^5\}, E\{0, b_4^3\}$
	$ABFGH$	$B\{i_1, b_1^3\}, F\{i_5, b_5^2\}, G\{0, b_6^4\}, H\{0, b_7^5\}$
3	$ABCE$	$B\{0, b_1^4\}, C\{0, b_2^2\}, E\{i_4, b_4^1\}$
	$ABCD$	$B\{i_1, b_1^2\}, C\{0, b_2^5\}, D\{i_3, b_3^2\}$
	AIL	$I\{0, b_8^2\}, L\{0, b_{11}^6\}$
4	$ABCE$	$B\{i_1, b_1^5\}, C\{i_2, b_2^5\}, E\{i_4, b_4^3\}$
	$ABFGH$	$B\{i_1, b_1^3\}, F\{i_5, b_5^2\}, G\{i_6, b_6^1\}, H\{0, b_7^3\}$
	$AIJK$	$I\{i_8, b_8^4\}, J\{i_9, b_9^1\}, K\{0, b_{10}^2\}$

Table 3
1-Transaction pattern candidate set (C1)

Path	Pattern	Sup	Path	Pattern	Sup
AB	$B\{i_1, b_1^3\}$	4	$ABFG$	$G\{i_6, b_6^1\}$	1
ABC	$C\{i_2, b_2^5\}$	3	$ABFGH$	$H\{i_7, b_7^1\}$	1
$ABCD$	$D\{i_3, b_3^2\}$	1	AI	$I\{i_8, b_8^4\}$	1
$ABCE$	$E\{i_4, b_4^3\}$	3	AIJ	$J\{i_9, b_9^1\}$	2
ABF	$F\{i_5, b_5^2\}$	2			

Table 4
Large 1-Transaction pattern set (T1)

Path	Pattern	Sup
AB	$B\{i_1, b_1^3\}$	4
ABC	$C\{i_2, b_2^5\}$	3
$ABCE$	$E\{i_4, b_4^3\}$	3
ABF	$F\{i_5, b_5^2\}$	2
AIJ	$J\{i_9, b_9^1\}$	2

Accordingly, the support value for page B is 4. However, the exact occurrences of IBD are calculated as the minimum values for each user. In the cases of user 2, $\min(b_1^2, b_1^3)$ leads to b_1^2, and of user 4, $\min(b_1^3, b_1^3, b_1^4)$ leads to b_1^3. Then the maximum value is calculated from the users with the same IBD. For b_1 in this example, the result is $\max\{b_1^2, b_1^2, b_1^2, b_1^3\}$, which equals to b_1^3. The first entry in Table 3 displays the results of the above processing. The other entries are derived in the same way.

Step 3: Assume that the support hurdle is set to 2. Only those patterns whose support values greater than or equal to 2 are kept in the large 1-transaction pattern set T1, as seen in Table 4.

Step 4: Generate the 2-transaction pattern candidate set from T1 by joining items in T1. The results are shown in Table 5.

Step 5: Only those patterns whose support values greater than or equal to 2 are kept in the large 2-transaction pattern set T2. The results are shown in Table 6.

Step 6: Generate the 3-transaction pattern candidate set C3 by joining 2-itemsets from T2. The results are shown in Table 7. Because the support values are all less than the support hurdle, the algorithm stops here.

The following three Web transaction rules are derived in this example: $\langle ABC\text{: } B\{i_1, b_2^3\} \Rightarrow C\{i_2, b_2^5\}\rangle$ with support = 2 and confidence = 50%, $\langle ABF\text{: } B\{i_1, b_1^3\} \Rightarrow F\{i_5, b_5^2\}\rangle$ with support = 2 and confidence = 50%, and $\langle ABCE\text{: } C\{i_2, b_2^5\} \Rightarrow E\{i_4, b_4^3\}\rangle$ with support = 2 and confidence = 67%. In addition to the rule implications of the original WTM, these rules also reveal the corresponding occurrences of selected IBDs. In practical applications, the Web host can predict more accurately by observing these IBD clues for potential buyers, and can interact with customers online with appropriate promotion strategies.

Table 5
2-Transaction pattern set (C2)

Path	Pattern	Sup	Path	Pattern	Sup
ABC	$B\{i_1, b_1^3\}C\{i_2, b_2^5\}$	2	ABF	$C\{i_2, b_2^5\}F\{i_5, b_5^2\}$	0
$ABCE$	$B\{i_1, b_1^3\}E\{i_4, b_4^3\}$	1	AIJ	$C\{i_2, b_2^5\}J\{i_9, b_9^1\}$	0
ABF	$B\{i_1, b_1^3\}F\{i_5, b_5^2\}$	2	ABF	$E\{i_4, b_4^3\}F\{i_5, b_5^2\}$	0
AIJ	$B\{i_1, b_1^3\}J\{i_9, b_9^1\}$	0	AIJ	$E\{i_4, b_4^3\}J\{i_9, b_9^1\}$	0
$ABCE$	$C\{i_2, b_2^5\}E\{i_4, b_4^3\}$	2	AIJ	$F\{i_5, b_5^2\}J\{i_9, b_9^1\}$	0

Table 6
Large 2-Transaction pattern set (T2)

Path	Pattern	Sup
ABC	$B\{i_1, b_1^3\}C\{i_2, b_2^5\}$	2
ABF	$B\{i_1, b_1^3\}F\{i_5, b_5^2\}$	2
$ABCE$	$C\{i_2, b_2^5\}E\{i_4, b_4^3\}$	2

Table 7
3-Transaction pattern candidate set (C3)

Path	Pattern	Sup
$ABCF$	$B\{i_1, b_1^3\}C\{i_2, b_2^5\}F\{i_5, b_5^2\}$	0
$ABCE$	$B\{i_1, b_1^3\}C\{i_2, b_2^5\}E\{i_4, b_4^3\}$	1
$ABCEF$	$B\{i_1, b_1^3\}C\{i_2, b_2^5\}E\{i_4, b_4^3\}F\{i_5, b_5^2\}$	0

5.2.2. $IWTM_{np}$

WTM only addresses the Web pages with purchases because there is no data available for the Web pages without purchases. With IBD available, the Web pages without purchases can be addressed directly. Therefore, $IWTM_{np}$ differs from $IWTM_p$ only in the targeted Web pages, which generate no sale at all. The $IWTM_{np}$ algorithm is thus the same as the $IWTM_p$ algorithm described in Section 5.2.1, except replacing the records with the no-purchase Web pages. Only Table 8 listing C1 and Table 9 listing T2 are shown for illustration. A complete step-by-step process of $IWTM_{np}$ can be seen in Section 5.3.2.

Table 8
C1 table

Path	Pattern	Sup
AB	$B\{0, b_1^5\}$	2
ABC	$C\{0, b_2^2\}$	1
$ABCE$	$E\{0, b_4^3\}$	1
ABF	$F\{0, b_5^4\}$	1
$ABFG$	$G\{0, b_6^4\}$	2
$ABFGH$	$H\{0, b_7^5\}$	2
AI	$I\{0, b_8^2\}$	2
$AIJK$	$K\{0, b_{10}^4\}$	2
AIL	$L\{0, b_{11}^6\}$	1

Table 9
T2 table

Path	Pattern	Sup
$ABFGH$	$G\{0, b_6^4\}H\{0, b_7^5\}$	1
$AIJK$	$I\{0, b_8^2\}K\{0, b_{10}^4\}$	1

The Web transaction rules derived in this exam- ple include $\langle ABFGH : G\{0, b_6^4\} \Rightarrow H\{0, b_7^5\}\rangle$ with support = 1and confidence = $\langle ABFGH : G\{0, b_6^4\} \Rightarrow H\{0, b_7^5\}\rangle / \langle ABFGH : H\{0, b_7^5\}\rangle = 1/2 = 50\%$, and $\langle AIJK : I\{0, b_8^2\} \Rightarrow K\{0, b_{10}^4\}\rangle$ with support = 1 and confidence = $\langle AIJK : I\{0, b_8^2\} \Rightarrow K\{0, b_{10}^4\}\rangle / \langle AIJK : K\{0, b_{10}^4\}\rangle = 1/2 = 50\%$. The original WTM is not applicable to data with no-purchased merchandise items, but the two derived rules are made possible by operating on IBD, which can be associated back to the merchandise items. In practical applications, the Web host can make decisions on handling those no-purchase merchandise items with these IBD clues representing the interest levels of potential buyers.

5.3. Implications on EC Strategies

The example in Section 5.2 is very simple and is used only to illustrate how the algorithm works. A larger data set is needed to illustrate the implications in practice. We thus conducted a data-collection experiment of 20 subjects who browsed through a Web shopping site. Each subject was asked to hypothetically shop the Web shopping site twice, with a very limited purchasing budget at the first time and an unlimited budget at the second time as their purchase-decision constraint. To meet the assumption of WTM, the shopping Web site has been modified to contain only one merchandise item in each page and no external link. Assume that only scroll-bar browsing data are collected on each page, which was automatically collected by the program as described in Yun and Chen (2000).

5.3.1. $IWTM_p$ with purchases

$IWTM_p$ used the data set generated by the subjects with unlimited purchasing budget. Due to the large size of the experimental data, only the partial 1-transaction pattern candidate set C1 and the final large 3-transaction pattern set T3 are shown here. They are in Tables 10 and 11, respectively.

The Web mining results generated these clues: $\langle ABHI: B\{i_1, b^3\}H\{i_7, b^4\}I\{i_8, b^3\}\rangle$ with Sup = 5, $\langle AB: B\{i_1, b^3\}\rangle$ with Sup value = 14 and $\langle ABH: B\{i_1, b^3\}H\{i_7, b^4\}\rangle$ with Sup = 7. Accordingly, two rules were derived as follows:

Rule 1: $\langle ABHI: B\{i_1, b^3\} \rightarrow H\{i_7, b^4\}I\{i_8, b^3\}\rangle$.
Confidence = $\langle ABHI: B\{i_1, b^3\}\mathrm{H}\{i_7, b^4\}I\{i_8, b^3\}\rangle / \langle AB: B\{i_1, b^3\}\rangle = 0.36$.

Rule 2: $\langle ABHI: B\{i_1, b^3\}H\{i_7, b^4\} \rightarrow I\{i_8, b^3\}\rangle$.
Confidence = $\langle ABHI: B\{i_1, b^3\}H\{i_7, b^4\}I\{i_8, b^3\}\rangle / \langle ABH: B\{i_1, b^3\}H\{i_7, b^4\}\rangle = 0.7$.

With the path ABHI, if a user has purchased at Web page *B*, the next possible Web pages with purchases are either *H* or *I*, or when a user has purchased at both Web pages *B* and *H*, the most likely Web page to purchase again is *I*. Consequently, the implied EC strategies are discussed as follows.

1. Enhancement within one rule
   The occurrences of IBD, i.e., scroll-bar action, can be used to enhance the derived rules by providing extra information for judging whether or not an online user's intention leads to next actual purchasing. For example, if a user has purchased on Web pages *B* and *H*, rule 1 infers that the user may also purchase the merchandise on Web page *I*, which is what the WTM algorithm would have provided. However, with the extra clue of the scroll-bar IBD, there are two more situations to be considered. First, if the user also had some occurrences of the scroll-bar IBD on Web page *I*, then the user was more likely interested in the merchandise on page *I*. Otherwise, zero occurrence of the scroll-bar IBD on page *I* indicates no immediate interest so far on the merchandise of page *I*, and more monitoring is desired.
   The other situation happens when users, for example Tom and Jim, both have purchased on Web pages *B* and *H*. Assume Tom also had four occurrences, while Jim only had one occurrence of the scroll-bar IBD, which may be an important indicator of the relative interests between different users. From the perspectives of online marketing, Tom has the higher possibility of purchase on Web page *I* at this time. If any marketing promotion to motivate potential buyers is triggered, Tom should be the one with a more preferable discount or free gifts. This implication illustrates how a Web site can better deploy the strategies and resources to its browsers.
2. Enhancement between rules
   The implication can also be judging the relative suitability of the two derived rules by the occurrences of the scroll-bar IBD. If a user has purchased on Web page *B*, then the next likely Web page with purchase can be

Table 10
Partial 1-Transaction pattern candidate set

Path	Pattern	Sup
AB	$B\{i_1, b^3\}$	14
ABC	$C\{i_2, b^3\}$	16
ABCD	$D\{i_3, b^2\}$	4
ABCE	$E\{i_4, b^3\}$	8
ABCEF	$F\{i_5, b^3\}$	4
ABCEFG	$G\{i_6, b^4\}$	4
ABH	$H\{i_7, b^4\}$	9
ABHI	$I\{i_8, b^3\}$	8
ABHIJ	$J\{i_{19}, b^3\}$	5
AK	$K\{i_{10}, b^5\}$	14
AKL	$L\{i_{11}, b^3\}$	15
AKLM	$M\{i_{12}, b^5\}$	7

Table 11
Large 3-Transaction Pattern Set

Path	Pattern	Sup
ABHI	$\mathrm{B}\{i_1, b^3\}\mathrm{H}\{i_7, b^4\}\mathrm{I}\{i_8, b^3\}$	5

judged by the Sup values. In the above case, Web page *H* has a higher Sup value than Web page *I* and should thus take a better promotion. Accordingly, the overall promotion cost can be lowered for better potential results. For instance, assume Tom has purchased on Web pages *B* and *H*, and had 5 occurrences of the scroll-bar IBD on Web page *I*, while Jim has purchased on Web page *B* and had 2 occurrences of the scroll-bar IBD on Web page *H*. Then the Web site can allocate more resources in promoting to Tom than to Jim by judging the occurrences of the scroll-bar IBD.

5.3.2. $IWTM_{np}$ with no purchases

$IWTM_{np}$ used the data set generated by the subjects who hypothetically shopped the experimental Web site with very limited purchasing budgets in order to obtain more no-purchase Web pages. Because Section 5.2 did not list the steps with the $IWTM_{np}$ algorithm, the step-by-step process is also illustrated below.

Step 1: Calculate the Sup values of both the IBD and Web pages for those pages without any purchase. The partial results for the 1-transaction pattern C1 data generated are shown in Table 12.

Step 2: Set the Sup hurdle value to 2, and retain only those large 1-transaction patterns with their support values greater than or equal to 2. The results are seen in Table 13.

Step 3: Continue from T1 by joining items for generating 2-transaction pattern candidate set C2. Since many patterns do not really exist, such as path *BK* which does not have any connection in the Web structural chart, their Sup values will be 0. The paths with Sup = 0 are not listed in Table 14.

Step 4: Keep only those large 2-transaction patterns with their support values greater than or equal to 2. The results are shown in Table 15.

Table 12
1-Transaction pattern candidate set C1

Path	Pattern	Sup
AB	$B\{i_1, b^3\}$	13
ABC	$C\{i_2, b^5\}$	8
ABCD	$D\{i_3, b^3\}$	4
ABCE	$E\{i_4, b^3\}$	3
ABCEF	$F\{i_5, b^4\}$	4
ABH	$H\{i_7, b^3\}$	3
ABHI	$I\{i_8, b^4\}$	4
ABHIJ	$J\{i_9, b^3\}$	3
AK	$K\{i_{10}, b^3\}$	3
AKL	$L\{i_{11}, b^4\}$	4
AKLM	$M\{i_{12}, b^5\}$	5
AKLMN	$N\{i_{13}, b^4\}$	4
AKLMNO	$O\{i_{14}, b^2\}$	2
AKLMNOP	$P\{i_{15}, b^1\}$	1
AQ	$Q\{i_{16}, b^6\}$	6
AQR	$R\{i_{17}, b^3\}$	3
AQRS	$S\{i_{18}, b^4\}$	4
AQT	$T\{i_{19}, b^3\}$	3

Table 13
Large 1-transaction pattern set T1

Path	Pattern	Sup
AB	$B\{i_1, b^3\}$	13
ABC	$C\{i_2, b^5\}$	8
ABCD	$D\{i_3, b^3\}$	4
ABCEF	$F\{i_5, b^3\}$	4
ABHI	$I\{i_8, b^3\}$	4
AKL	$L\{i_{11}, b^2\}$	4
AKLM	$M\{i_{12}, b^2\}$	5
AKLMN	$N\{i_{13}, b^2\}$	4
AQ	$Q\{i_{16}, b^5\}$	6
AQRS	$S\{i_{18}, b^4\}$	4

Table 14
2-Transaction pattern candidate set C2

Path	Page w/o Purchase	Sup
ABC	$B\{i_1, b^3\}$ $C\{i_2, b^5\}$	2
ABCEF	$B\{i_1, b^3\}$ $F\{i_4, b^3\}$	1
ABHI	$B\{i_1, b^3\}$ $I\{i_8, b^3\}$	1
ABCEF	$C\{i_2, b^5\}$ $F\{i_4, b^3\}$	1
AKLM	$L\{i_{11}, b^2\}$ $M\{i_{12}, b^2\}$	1
AKLMN	$M\{i_{12}, b^2\}$ $N\{i_{13}, b^2\}$	1
AQRS	$Q\{i_{16}, b^5\}$ $S\{i_{18}, b^4\}$	1

Table 15
Large 2-transaction pattern set T2

Path	Pattern	Sup
ABC	$B\{i_1, b^3\}$ $C\{i_2, b^5\}$	2

The algorithm then stops since only one pattern remains in Table 15. The final browsing path is thus *ABC*, where *BC* is the sub-path with Sup value = 2, and its total occurrences of browsing data are 3 and 5 for *B* and *C*, respectively. In other words, the Web mining results generate these clues: $\langle ABC\text{: } B\{i_1, b^3\}\ C\{i_2, b^5\}\rangle$ with Sup = 2, and $<AB\text{: } B\{i_1, b^3\}\rangle$ with Sup = 13. Accordingly, one rule is derived as $\langle ABC\text{: } B\{i_1, b^3\} \rightarrow C\{i_2, b^5\}\rangle$ with Confidence = $\langle ABC\text{: } B\{i_1, b^3\}C\{i_2, b^5\}\rangle / \langle AB\text{: } B\{i_1, b^3\}\rangle = 0.15$. Similar to the implications for $IWTM_p$, the implied EC strategies for $IWTM_{np}$ are discussed as follows:

1. *Screening out no-purchase and low-interest merchandise*
   Unpopular merchandise management is an important issue in Web site management. Any unpopular merchandise may have to be taken offline if it has no potential at all, or may need more allocated budgets for a stronger promoting if it still has good potential. In our example, Table 12 reveals that Web page *B* has a high Sup of 13 but with low scroll-bar IBD occurrences of 3. As compared to Web page *C* with Sup of 8 and 5 occurrences of scroll-bar IBD, Web page *B* may be considered taken off the catalog for more effective use of the space and marketing budget.
2. *Promoting no-purchase but high-interest merchandise*
   For effectively allocating budget on promoting merchandise sales, the IBD occurrence serves as an indicator of

the level of interests reflected by the users during their browsing processes. In Table 12, although Web pages *C* and *Q* have high Sup values, these values are still less than that of Web page *B*. The occurrences of the scroll-bar IBD in *C* and *Q* are, however, higher than that of *B*. Therefore, instead of replacing Web pages *C* and *Q*, more efforts should be spent on investigating why users did not purchase potentially interesting merchandise. Is it because of the price or the specification? With adequate research and investigation, appropriate marketing strategies may be applied to realizing actual sales out of Web pages *C* and *Q*.

3. *Promoting positively correlated merchandises*
From Table 14, we learn that when a user had no purchase on Web pages *B*, then his/her chance to purchase on Web page *C* was low. However, considering the scroll-bar IBD of these two pages together, Web page *B* has 3 and *C* has 5 occurrences, which indicates a higher level of interests on *C*. For instance, if Web page *B* lists a hard disk while Web page *C* lists a CPU, there are two possible situations. First, the user may not want the hard disk on Web page *B*, but raise the interests in CPU on Web page *C*. A marketing strategy of cross-selling or package promotion can then be applied to the user, such as buying a hard disk with 20% off the CPU price in order to increase the sales of both merchandise items. Second, the Web site management can use the hard disk on Web page B as a basis of comparison to the user by strongly promoting the CPU on Web page *C* while maintaining a stiff price for the merchandise on Web page *B*, so that the user may really feel the bargain price offers of the interested CPU, and make the deal.

5.3.3. Discussions

The illustration in Section 5.3.2 has its pros and cons. IBD is meant to bring "potential" effectiveness into WUM. Therefore, the occurrence of the scroll-bar IBD does not guarantee the actual users' levels of interest. For example, there may be no scroll bars on Web pages whose contents are less than one page, or Web pages may be very long and thus more scrolling actions are always performed. The scroll-bar IBD in this example is, however, an additional indicator for strategy deployment from a conservative perspective, which is better than nothing to the decision support. Moreover, the effects of benefited implications depend heavily on the application domain, the data set, the selected IBD, and the persons who generate or implement the strategies. Theoretically speaking, if the two potential benefits defined in Section 4.4 can be carefully manipulated, positive benefits would eventually justify the value of IBD in practice.

WTM is only an exemplar algorithm adopted for the purpose of illustration in this paper. From the way IBD was incorporated into the existing WTM algorithm, we learn that the IBD treatment can be thought of as an adds-on process to an original algorithm on Web log data items. Accordingly, any existing WUM algorithms with similar data structures can undoubtedly accommodate IBD for potential effectiveness.

Although the above discussion only focuses on the potential benefits of IBD, it is clearly an enabling component in practical WUM applications and an innovative one in further WUM research and development.

6. Conclusions and future works

This paper provided the cornerstone research results on IBD, which help open up the scope of WUM applications or decision support for knowledge discovery by jumping out the existing frame of algorithm as well as the conventional Web log records. We have defined the IBD, proposed a taxonomy for browsing data, described the mechanism for collecting online browsing data, and justified two benefits of IBD. Moreover, an EC example of WTM was adopted for illustrating the potential effectiveness of IBD via two simple modifications of intention-based WTM (IWTM) algorithms.

The major implication of IBD in decision support is that by capturing more subtle and personal online browsing behavior in addition to the fixed Web log records, business strategies with greater personal touch that were never experienced can be distinguished and deployed for better online competitive advantages. Moreover, IBD can be a value-added component to the regular WUM algorithms at a lower cost. In other words, IBD can be used just like Web log records in WUM algorithms, except with potential extra values in decision making.

An immediate future work is to enhance the online data collection module (Tao et al., 2004) so that IBD can be easily adopted in WUM for fulfilling its potential benefits. Because IBD sheds light for an unlimited imaginary space for future WUM applications, another future work is to survey the importance and priority of research and development issues in WUM applications with IBD in context. With this, practical innovations on WUM algorithms can be pursued. Examples would be unifying IWTMp and IWTMnp into one IWTM algorithm for a more convenient and easier usage, or releasing the constraint of one merchandise item or one IBD in ITWM algorithms.

References

Agrawal, R. & Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining. In *The international conference on very large database, Bombay, India*, pp. 544–555.

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. In *The international conference on very large database*, Santiago, Chile, pp. 487–499.

Agrawal, R. & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan*, pp. 3–14 .

Bonchi, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D., et al. (2001). Web log data warehousing and mining for intelligent web caching. *Data and Knowledge Engineering, 39*(2), 165–189.

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems, 27*(6), 1065–1073.

Chan, C. (1997). The access log to Web site and query language on WWW, Unpublished Master Thesis, Information Engineering Graduate School, National Central University.

Chen, M. S., Park, J. S., & Yu, P. S. (1998). Efficient data mining for path traversal patterns. *IEEE Transaction on Knowledge and Data Engineering, 10*(2), 209–221.

Chen, Z., & Shen, H. (2000). A study of a new method of browsing path data mining. *The sixth international conference of information management research and practice*. HsingChu, Taiwan, ROC: TsingHua University.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems, 1*(1), 5–32.

Cunha, C. R. & Jaccoud, C. F. B. (1997). Determining WWW user's next access and its application to pre-fetching. In *The second IEEE symposium on computers and communications, Alexandria, Egypt*, pp. 6–11.

Fann, C. (1999). *Personalized interactive marketing mechanisms on WWW*, Unpublished master thesis, Information Management Graduate School, National Pingtung University of Science and Technology.

Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. Academic Press.

Holsapple, C. W., & Whinston, A. B. (1996). Decision support systems: A knowledge-based approach. West Publishing Company.

Hong, T. P., Lin, K. Y., & Wang, S. L. (2002). Mining linguistic browsing patterns in the World Wide Web. *Soft Computing, 6*(5), 329–336.

Hsieh, C. C., & Chang, C. T. (2001). An enhanced transaction identification module on Web usage mining. *Asia Pacific Management*, 241–252.

Iyengar, A., MacNair, E. & Nguyen, T. (1997). An analysis of Web server performance. In *The IEEE global telecommunications conference*, Vol. 3, Phoenix, AZ, USA, pp. 1943–1947.

Joshi, K. P., Joshi, A., Yeti, Y., & Krishnapuram, R. (1999). *Warehousing and mining Web logs*. WIDM, Kansas City, Mo, USA: ACM, pp. 63–68.

Kitsuregawa, M., Shintani, T., & Pramudiono, I. (2001). Web mining and its SQL based parallel execution. *Proceedings on Information Technology for Virtual Enterprises, 5*(5), 128–134.

Lin, Y. (1997). *A design and implementation of a data collection mechanism in mining WWW information*, Unpublished master thesis, National Taiwan University.

Maglio, P. P., Campbell, C. S., Barrett, R., & Selker, T. (2001). An architecture for developing attentive information systems. *Knowledge-Based Systems, 14*(1–2), 103–110.

Mannila, H. & Ronkainen, P. (1997). Similarity of event sequences. In *The fourth international workshop on temporal representation and reasoning*, pp. 136–139.

Mehta, M., Agrawal, R. & Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. In *Proceedings of the fifth international conference on extending database technology, France*, pp. 8–32.

Mittal, B., & Lassar, W. M. (1987). The role of personalization in service encounters. *Journal of Retailing, 72*(1), 95–109.

Modhal, M. (1999). *Now or never: How companies must change today to win the battle for internet consumers*. Harper Business.

Newman, W. M., & Lamming, M. G. (1995). *Interactive system design*. Addison-Wesley.

Park, J. S., Chen, M. S., & Yu, P. S. (1997). Using a hash-based method with transaction trimming for mining association rules. *The IEEE International Conference on Knowledge and Data Mining, 9*(5), 813–825.

Perkowitz, M., & Etzioni, O. (2000). Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligence, 118*(1–2), 245–275.

Reeder, R. W., Pirolli, P. & Card, S. K. (2000). *WebLogger: A data collection tool for Web-use studies*, UIR Technical report UIR-R-2000-06, Xerox PARC.

Tao, Y.-H., Chung, S., Chung, M., Kao, H., Yang, K. & Lin., I. (2004). *The data-collection mechanism of Web browsing behavior*, Association of Electronic Commerce in Taiwan, March 26–27.

Yu, P. (1999). Data mining and personalization technologies. In *The sixth IEEE international conference on database systems for advanced applications*, pp. 6–13.

Yun, C. & Chen, M. (2000). Mining Web transaction patterns in an electronic commerce environment. In *The fourth pacific-asia conference on knowledge discovery and data mining*, pp. 216–219.

Zhang, D., & Dong, Y. (2002). A novel Web usage mining approach for search engines. *Computer Networks, 39*(3), 303–310.