Market segmentation via structured click stream analysis

Kuang-Wei Wen
Department of Information Systems, University of Wisconsin-La Crosse, La Crosse, Wisconsin, USA
Kuo-Fang Peng
Department of Information Management, National Central University, Chung-Li, Taiwan, Republic of China

Keywords
Market segmentation, Personality, Analysis, Neural networks

Abstract
Accurate market segmentation has been the basis for successful customization of products and services. To date, however, the marketing management literature has focused mainly on the exploration of segmentation variables, but lagged behind in the development of practical means for segmentation mechanisms using contemporary information technology. Motivated by this shortcoming, the current study attempts to devise an effective method that allows for systematic collection and analysis of online customers' click stream data to facilitate market segmentation. Cohen's CAD theory was employed in conjunction with artificial neural network models to provide the analytical foundation of this research. To test the effectiveness of the proposed method, a sizable online field experiment utilizing a disguised 7-ELEVEN Website was conducted, and 912 useful click streams collected. The results from the subsequent data analysis supported the feasibility of the current work, but also identified the needs for further study.

Industrial Management & Data Systems
102/9 [2002] 493–502
© MCB UP Limited
[ISSN 0263-5577]
[DOI 10.1108/02635570210450172]

Introduction

Given the dramatic rise and fall of the dot-com companies in the past two years, the topic of Internet marketing has been gaining more than its fair share of attention from academics. Current market downturn notwithstanding, the importance of joining the digital economy, as a competitive strategy, has also been recognized by both the click-and-click and brick-and-click corporations. As most companies will continue to invest in the Internet technology in the future, any such investment is also likely to be subject to harsher scrutiny for its worthiness. This new trend calls for innovative ways to leverage the Internet technology, and market segmentation using Web browsing data is one of the most promising applications that interest both academia and industry.

Internet marketing is a multidimensional activity that faces many challenges. The prevalent use of Web sites to maximize the reach of static information to potential customers improves the efficiency of advertisement, but is not a novel marketing function by itself. No matter how rich and splendid the content of the Web site is, the company is at best appealing to a faceless public. While the true effectiveness of this mode of digital dissemination of marketing information or messages is still hard to assess at the current stage, more promising Internet applications based on active server technologies have been attempted. However, in order to gain a deeper understanding of the potentials of these new technologies, we need to revisit the principles of modern marketing management.

Kotler *et al.* (1999) set three essential steps for marketing management: market segmentation, market segment targeting, and design and implementation of marketing mix. While it is not difficult to see that each of these steps requires special handling under the new marketing vehicle of the Internet, the crucial importance of market segmentation as the foundation for the follow-up steps remains unchanged. We must know who our customers are, what special characteristics they exhibit and what shopping preferences they have before we can find effective ways to reach them and design the right products or services for them.

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0263-5577.htm

The issue of market segmentation presents a unique challenge to electronic commerce (e-commerce hereafter). Traditionally, we could analyze customer data collected from purchase transactions, credit applications, membership history, and market tests to identify market segments and consumer groups accurately. But this is essentially an *ex post* approach, meaning that some marketing activities must have taken place and that we have been in contact with our customers. There is very little we can do about prospective customers under this approach other than performing a crude analysis on publicly available demographic data. (The unethical practice of buying or exchanging customers' private data without explicit permissions is excluded from discussion in this work.) The same dilemma exits, and may be more acute, in Internet marketing if we continue to follow the traditional approach. Given the fact that companies could achieve worldwide exposure to prospective customers through the Internet, it is imperative for them to be able to understand any voluntary business lead that comes in the form of Web site visits.

So far, segmenting online customers as a means for mass customization remains a challenging task (Chang, 1998). The difficulty of it stems from two main sources:

1 customers' reluctance to provide personal information upon initial contact(s) with the companies' Web sites; and

2 scarcity of repeat Web site visits by surfers (Huberman *et al.*, 1998).

The well-known techniques of micro marketing (e.g. cookies and Web bugs) and collective profiling/collaborative filtering (Deitel *et al.*, 2001) employed by many e-companies such as Amazon.com cannot overcome these problems, due to their reliance on credible customer records and longitudinal transaction data. What we really are lacking in our arsenal is an invisible and non-intrusive method for collecting useful and analyzable data from every one-time visit by Web surfers – an observational and behavioral approach.

The main contribution of this research is in the development and validation of an integrated scheme that:

- systematically identifies a market segmentation variable;
- derives its behavioral characteristics;
- designs discriminating choice factors for the Web pages;
- dynamically collects and records relevant click streams; and
- uses a high power analytical tool (i.e. neural networks in the current study) to effect segmentation.

The validated scheme has the ability to employ specially designed Web pages to capture click streams from the first visit of a potential customer and fairly accurately determine his/her personality type. We dubbed this scheme "structured click stream analysis" in order to differentiate it from the more simplistic and general approach based on the analysis of server log files (Chang, 1999; Wen and Liu, 1998). Also notice that our scheme does not pertain to the now celebrated data mining techniques (e.g. Ha and Park, 1998; Petersohn, 1998; Vellido *et al.*, 1999) that apply solely to already collected data and do not intend to address the data collection methods.

The remainder of this paper is organized in the following fashion. The next section presents a review of relevant literature on market segmentation, personality assessment, Web browsing behavior analysis, and artificial neural networks. The following section outlines our research methods and procedure design, along with a description of the construction of the experimental Web site. Our data analysis based on 912 complete observations collected from Taiwan's Internet users in 1999 is presented in the penultimate section, whereas the final section concludes the paper and suggests the most fruitful directions for future research to follow.

Literature review

Market segmentation

As the first step leading to sound determination of marketing mix and strategies, effective market segmentation should be measurable, sustainable, accessible, differentiable, and actionable (Kotler *et al.*, 1999). To this end, a good market segmentation scheme always requires a careful identification of segmentation variables that are geographical, demographical, psychological, or behavioral. Of the psychological variables, personality type and life style are the explicit ones, whereas many factors such as product usage, product knowledge and attitude towards products constitute the behavioral variable. Although each type of segmentation variable has its distinct relevance to different segmentation schemes, we chose to focus on personality type in this study for several important reasons. First, since we were restricted by the sensing technologies available for observing surfing behavior, demographical data proved difficult to collect. Second, the fact that the vast majority of subjects of our experiment were drawn from the Internet market in Taiwan (a Chinese-speaking population) nullified the significance of geographical variables. And third, most behavioral variables require long-term observation and record keeping on the same subject, thereby defeating the purpose of being able to segment the market based on data of single Web visits. Theses research constraints left us with the psychological factors for consideration, of which we selected personality type, due to its strong support from existing psychology and marketing literature.

Personality types and classification instruments

There exist many theories on personality trait/type in the literature of psychology; however, our choice has leaned towards Horney's (1945) theory because:

- this theory has seen many applications in marketing research (Noerager, 1979); and
- local and cross-cultural validations of the theory have been performed in the host nation – Taiwan (Hsu, 1993; Wen and Liu, 1998).

It should be understood that careful validation of the theory is still needed if it is to be used in a substantially different market.

Basically, Horney classified personality into three generic types: compliant, aggressive, and detached (CAD). The compliant type expect to become a part of others' lives, to be loved, appreciated, and needed, and to look to others for answers to their personal problems. People of this type tend to overemphasize friendship and love, resulting in becoming oversensitive, over generous, and over compassionate. The typical traits of this type include being non-selfish, affectionate, compassionate, charitable and polite. On the other hand, the aggressive type tend to pursue excellence, achievements, power and compliments. They see others as competitors, strive to control personal emotions and fears, and hope to become superior strategists. As they consider dominance, power, and practicality as necessities for success, they hold a utilitarian or an egalitarian view towards the world. These people enjoy controlling other people. Finally, the detached type tend to maintain a distance from others; they are cold, independent, and they dislike responsibilities. Most of them value intelligence and logical reasoning, but not emotion, in solving problems. In addition, they do not like to argue with others, nor do they trust people easily.

Having adopted Horney's personality categories, we still need to find a robust and feasible instrument for classifying customers into the correct categories, and Burger's (1993) work could shed some light on this task. Burger described three basic techniques for assessing personality types:

1. *Survey instruments.* This is the most widely used technique for assessing personality using a set of survey questions. Scores are assigned to the subject's answers to the questionnaire, and a classifying scheme is employed to translate scores into personality types. This technique is simple, consistent, easy to implement, and objective in terms of assessment. However, it suffers from all the common drawbacks of a typical perception-based questionnaire survey, in that the effects of sentiment, knowledge, and self-selection of the subject on the answers have long been questioned.
2. *Projection techniques.* The basic idea behind these methods is to apply certain structured or unstructured external stimuli (such as a picture or a photo) to a subject and let him/her verbally describe the resulting free association. The experimenter would then determine the personality type of the subject from the descriptions based on expert opinions. This technique has the advantage of penetrating the subject's self-defense and probing into the inner psychology to pinpoint personality type. However, the technique is subjective, requires the use of highly-trained experts, and can only handle a limited number of subjects due to cost and consistency considerations.
3. *Behavioral observation.* This method employs concealed recording devices (e.g. movie cameras, voice recorders, or manual forms) to document subjects' explicit behavior. The necessary recording period varies in length according to survey design, and the actual determination of personality type also relies on expert judgments of the recorded data. This is the most non-intrusive technique for collecting behavioral data, but it is also subject to judgmental biases by the human expert(s).

For the purpose of our work here, since we intend to develop an efficient and automated way of performing personality classification, the most suitable approach would appear to be using a survey instrument in conjunction with behavioral observation. Consequently, Cohen's (1967) 35-question (ten for compliance, 15 for aggressive, and ten for detached) personality measuring instrument was identified and adopted.

Artificial neural networks

Since our final goal for this research is to classify customers accurately and quickly into their correct personality types, an efficient classification scheme is needed. To date, the parametric approach of multiple discriminant analysis (MDA) and the learning-based artificial neural networks (ANNs) are the top contenders for the task. We chose the ANN over the MDA for its flexibility in model building and ability to incorporate new information. A brief discussion of the ANNs is presented below. For an in-depth review please refer to Widrow *et al.* (1994).

ANNs are an information processing technology pertaining to the area of machine learning in artificial intelligence. Unlike the traditional algorithmic approach to computing which requires a fixed structure for knowledge representation and predetermined processing steps, a neural network employs an adaptive structure that can be trained with application data to capture complex relationships between input and output variables. Although most neural networks mimic the fundamental structure of the human brain, their processing power, as well as learning ability, are still limited by today's computer hardware and software technologies in the area of distributed parallel processing. But even so, the neural network technique has been successfully used to solve difficult practical problems and perform important commercial tasks.

An ANN is composed of two or multiple layers of artificial neurons called processing elements (PEs) that are linked by variable-strength connections. Each of the PEs receives inputs, processes them, and delivers a single output. Whereas the inputs can be raw data or outputs of other PEs, the output can be the final product or an input to another PE.

Depending upon the intended application, the network structure can be any one of the following: feed-forward (associative), feed-back (autoassociative), or the hybrid of the two (Aleksander and Morton, 1990). A feed-forward network typically directs data flow in a single direction from the input layer to the final output and also disallows inter- and intra-layer feedback, but these restrictions are fully or partially relaxed in the other two types of networks.

The first learning algorithm, known as the delta rule, for a two-layer feed-forward ANN (termed Perceptron) was provided by Widrow (1962). However, this technique was insufficient for training nets with hidden layers that do not have direct exposure to the inputs and outputs. Although Hopfield's (1982) energy model and its associated simulated annealing approach did overcome the learning difficulty in autoassociative networks, it was the back propagation method suggested by Rumelhart *et al.* (1986) that enabled effective learning for multi-layer feed-forward networks. In essence, the back propagation algorithm consists of three general steps:

1. compute outputs;
2. compare outputs with desired targets; and
3. adjust connection weights and parameters of the activation function(s) to remove as much output errors as possible.

This procedure is performed iteratively with a sufficiently large training data set until the performance of the network reaches a predetermined level (a process called convergence). In each iteration of training, the computed errors are distributed to all the layers in a backward fashion, while the adjustments to each PE and its connections are performed using a gradient method. It is important to note that the back propagation method for training alone does not always guarantee satisfactory convergence. Successful network construction also hinges on other crucial factors including correct identification of input variables, adequate design of network structure, and appropriate fine-tuning of the training algorithm.

Web browsing behavior analysis

In its simplest form, Web browsing behavior analysis degenerates to the familiar click stream analysis. Early click stream analyses such as Tom's (1996) only provided a mechanism to extract up to nine data items from a Web server's log file: number of hits, Web page number, session starting time, URL of destination, page title, MIME type, referring page, page duration, and session time. Today, it is common practice for data mining software imbedded in large commercial CRM packages (for example, 3Com's *Contact Advantage*) to analyze a subset of these data items to either assess the performance of a specific Web site, or to build customer profiles from longitudinal databases. However, these conventional uses of log files do not offer any insights to help companies better understand their potential customers.

Only a few attempts in literature have tried to link the order, number, and duration of page visits to customer preference. Wen *et al.* (1997) demonstrated an exploratory study that employed regression models to establish such linkage, and subsequently validated the results against preference data obtained via the analytical hierarchy process (AHP) analysis. An alternative way of establishing such linkage using ANNs was later explored by Wen and Young (1998) and the results compared favorably with those of regression analysis. More recently, a third alternative based on logistic regression was proposed by Chang (1999), and improved results were obtained.

It is important to point out that all the aforementioned click stream analyses treated log file data types as given, and strived to make the best sense out of the server records. Although this stream of research work has paved the foundation for more sophisticated analysis, its assumption of data sources presents a conceptual hurdle that might limit methodological progress. What we try to achieve in this study is to remove the very hurdle by developing an effective and robust data collection framework at the front end.

Research method and design

Design of research procedure

The objective of the research design is to develop a framework and a mechanism to collect pre-specified click stream data and build a neural network to correctly classify customers into three personality types according to a single visit to the Web site. To increase the external validity of our work we employed the field observation method (Cooper and Schindler, 1998). We chose and obtained permission to imitate and modify the corporate Web site of the largest convenience store chain, 7-ELEVEN, to collect click stream data from unsuspicious visitors (subjects). For validation purpose, we also encouraged each Web site visitor to fill out Cohen's CAD questionnaire by using a sweepstake scheme as incentive. The CAD data established the correct personality types on which a neural network model was trained, and against which the predictions by the model were compared. The two-layer framework is presented in Figure 1.

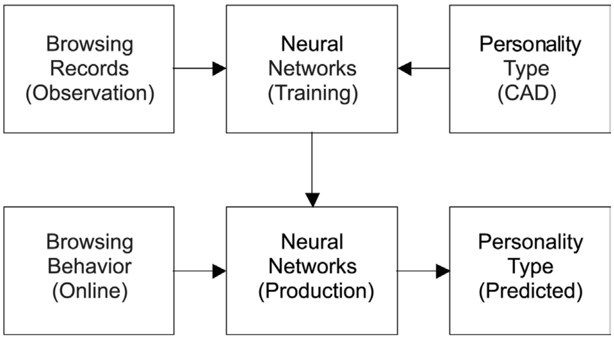
The complete research procedure for the current study includes: discovery of personality discriminating factors (PDFs), developing methods for operationalizing the PDFs in Web site design, pilot study, data collection through the main experiment, and construction and validation of the ANN model. This procedure is depicted in Figure 2.

Discovery of the PDFs

The three personality types referred to in this research in fact were "interpersonal orientations" defined by Cohen (1967). Since most of later works (e.g. Noerager, 1979; Schiffman and Kanuk, 1997) equated this term with personality types, we also follow this convention for consistency of comparison.

Based on a review of personality literature, we found two categories of possible discriminating factors that are manageable in Web page design: non-textual and personality trait related. The non-textual factors included color (Rogers *et al.*, 1983) and shape (Fu, 1995) preferences by different personality types, whereas the personality traits related factors could be the descriptive tendencies of each type summarized by Horney (1945). Even with these findings, operationalizing these PDFs in the context of a Web page still was a difficult task. To the extent that the quality of the PDFs could directly affect the power of the ANN to be built, more theories are needed for guiding the exploration of robust PDFs.

Figure 1
Framework of research design


Browsing Records (Observation)
Neural Networks (Training)
Personality Type (CAD)
Browsing Behavior (Online)
Neural Networks (Production)
Personality Type (Predicted)

Web page design with embedded PDFs

Based on generalizability consideration of our research results, we determined that the corporate Web site was the optimal choice, given the fact that it has been the most visible type of commercial application of Internet technology. In general, the content of a typical corporate Web site encompasses company profile, current news, products and services, hot links, correspondence section and e-mail as the main menu items. The detailed second-layer options could vary across companies. This knowledge allowed us to design a real-looking corporate Web site while still being able to embed many PDFs in each page in a natural way.

After a careful examination of a large number of corporate Web sites in Taiwan, we found the site of 7-ELEVEN fitted our needs perfectly. This particular site was well known locally, and its content was rich enough for the inclusion of our intended PDFs. Using such a real site for data collection would help avoid the subjects' self-selection problem that could arise when they doubt the validity of the site being visited. To the credit of corporate 7-ELEVEN in Taiwan, permission was granted to us for the set up of a disguised site that was linked to the real Web site. During the field experiment period, many 7-ELEVEN site visitors were unknowingly directed to our site and eventually volunteered to participate in our research.

Through careful selection we identified five PDFs for each personality type; they are displayed in Table I.

It was our belief that these PDFs could create discernible patterns in the click streams left by subjects of different personality types, which would then be captured by a properly designed ANN.

Construction of the experimental Web site

The initial content of our experimental Web site was downloaded with permission from 7-ELEVEN's corporate Web site (www.7-11.com.tw). After filtering out pages that were irrelevant to our study we retained over 40 pages for further modification. However, we divided these pages and arranged them into the original six topics in the corporate Web site:

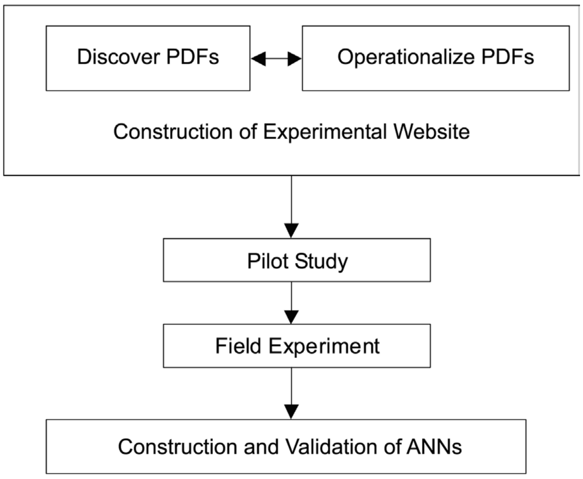
1 Enterprise Observation Tower.
2 Super Popular Products.
3 Social Concerns.
4 Hope Bank.
5 Current Events.
6 Communication Board.

These topics constituted the main menu on the left frame of the first page (as shown in Figure 3 with Enterprise Observation Tower being selected and displayed in the main window) and more detailed options showed up in the right window upon clicking the main items.

This three-frame design has the advantage of allowing fast browsing by the subject while minimizing the danger of disorientation in the site (Wen *et al.*, 1997). In addition, the division of windows and use of menus could help balance the chances of the PDFs being clicked, and contribute to the creation of better browsing data. We also designed and constructed the online questionnaire pages that implemented Cohen's CAD instrument.

Microsoft Windows NT and its Internet information server (IIS) were employed as the Web site platform, whereas both active server pages (ASP) and JavaScript were used to interactively and dynamically extract, as well as process, input data and click streams. The back-end database management system was ACCESS 97, which stored the experimental data in several databases.

Figure 2
Research procedure


Discover PDFs
Operationalize PDFs
Construction of Experimental Website
Pilot Study
Field Experiment
Construction and Validation of ANNs

Pilot study and the main experiment

The pilot study was conducted on the campus of a large national university in southern Taiwan, with subjects drawn from the student population. Volunteers were informed of this study as the opening of a real Web site and attracted to the experiment by small prizes. They were given ten minutes to browse the pages freely after entering the site and then asked to fill out personal information and the CAD questionnaire online. Upon completion of the experiment, the subject received a prize by mail. During the experiment, the system recorded the click stream data at two-minute intervals, thus providing up to five data streams from each subject for subsequent data analysis. The small-scale pilot study validated the experimental site and all data collection, transmission, and storage functions.

The main experiment was conducted on the national Internet of Taiwan. We posted the announcement of a new 7-ELEVEN Web site on all academic BBS, commercial portals, and many popular private Web sites in order to attract a representative sample of subjects. Since we needed a large number of observations for training our ANN, we

Table I
PDFs for all personality types

C1 Green entry	**C2** Social issues	**C3** Warm love story	**C4** Brand new lotto	**C5** Express opinions
A1 Red entry	**A2** Breaking news	**A3** Brand-named products	**A4** Competitive marketing strategies	**A5** Entrepreneurship
D1 Blue entry	**D2** Management concept analysis	**D3** Analysis of operations data	**D4** Clean up of special development zones	**D5** Work environment and benefits analysis

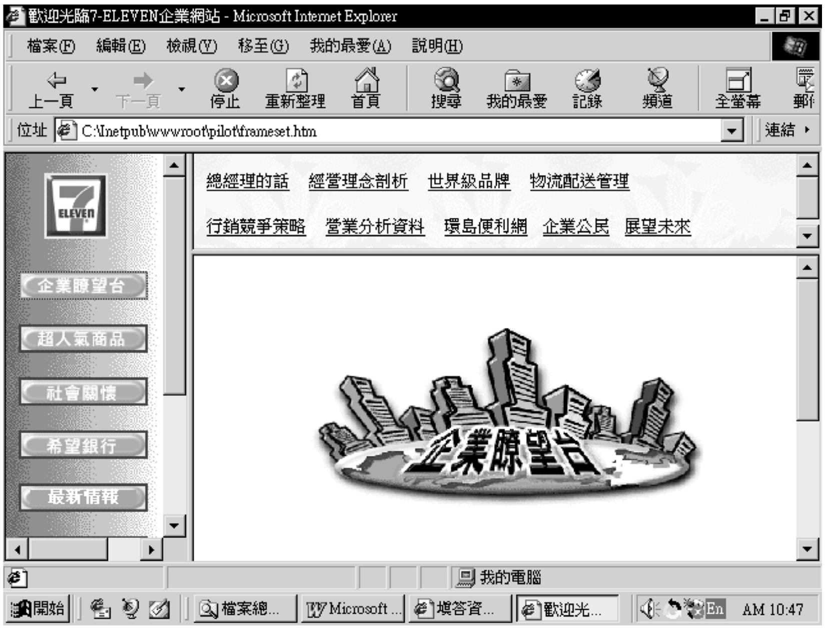
Notes: C: compliant; A: aggressive; D: detached resorted to a sweepstake scheme to motivate the public. Our condition for a subject to be eligible for prize draws was the completion of 10 minutes of browsing and the submission of the CAD questionnaire.

Construction and validation of the classification ANNs

After having carefully filtered out incomplete information and suspicious inaction, we retained 912 observations from the ACCESS databases for the construction, training, and validation of a set of classification ANNs. Two-thirds (608) of the click streams were used for model construction and training, whereas the remaining one-third (304) were used for validation.

For efficiency, we employed the NeuroSolution software provided by NeuroDimension in the development of the ANNs. Each of our models was a feed-forward Perceptron with two hidden layers. It contained 15 binary input variables corresponding to the PDFs and produced a single output – the personality type. Supervised back-propagation learning with the true personality type data derived from the CAD instrument was performed by the software. Since we collected the click streams at five different times, we had the opportunity to built five ANNs to explore the optimal timing of data collection, a novel and important subject that has not been studied so far. The results of our data analysis are presented next.

Figure 3
Sample Web site page



Results of data analysis

Distribution of personality types among Internet users (Taiwanese market)

Out of the 912 observations we used for data analysis, there were 731 (81 per cent) compliance type, 98 (10.5 per cent) aggressive type, and 77 (8.5 per cent) detached type. We believe that this distribution of personality type was fairly representative, because it did not significantly deviate from the distribution of the entire sample in the databases, which was characterized by an 8:1:1 ratio among the CAD groups.

The highly asymmetric distribution can readily be explained from a cultural perspective. The fact that more than two thousand years of Confucianism has effectively tamed the Chinese society and produced a predominately compliant population is quite self-evident here. An alternative explanation would come from the argument that compliant individuals have a stronger tendency to participate in academic research, given their charitable and compassionate nature. While this rather unsurprising phenomenon was not a problem by itself, it did present a difficulty in our ANN training: the very uneven training data set violated one of the basic principles for constructing a good ANN.

The trained ANNs

Traditionally the construction and training of ANNs has been heuristic at best, and trial-and-error at worst. Although there are powerful genetic algorithms that can automate this task, we preferred to take the traditional approach in order to gain deeper insights into possible structural problems. For this research we literally have tried hundreds of ANNs using a heuristic approach; the performance of the final ANNs are summarized in Table II.

Our best results demonstrated an overall accuracy of classification around 75 per cent, a level that is acceptable but can possibly be improved by a few extra steps. Although the models performed up to above 90 per cent accuracy for the majority type (compliant), the poor accuracies for the two minority types become alarming. Two plausible reasons could account for this weak outcome. First, we might not have fed the models with a sufficient number of observations for the two minority types, thereby allowing the compliant type to dominate learning. The net outcome was a set of nets that could only capture the characteristic patterns of the majority type. The second reason is more subtle; it relates to the time required for the

Table II
Accuracies of the ANNs

Duration	Accuracy-compliant (per cent)	Accuracy-aggressive (per cent)	Accuracy-detached (per cent)	Overall accuracy (per cent)
2 mins	94.4	3.6	7.4	77.5
4 mins	89.9	17.1	11.1	75.6
6 mins	86.1	7.1	7.4	71.1
8 mins	81.9	17.9	7.4	68.8
10 mins	73.6	21.4	22.2	63.8

minority types to reveal their distinct behavioral features. If we believe that the Chinese society is basically composed of the compliant population, then the personality traits of the other two types might also have been weakened. It follows that we will need more powerful PDFs for correctly identifying these few people from the large, homogeneous crowd of compliance type. This is a topic that deserves a separate study.

Given the above results, several observations can be made and discussed. First of all, the duration of Web browsing affects the classification accuracy of ANNs. While the accuracy for the compliant type degraded with time, the best performance for the other two types occurred at the end of the experiment period. However, if consistency of performance for all types is of concern, then the four-minute data yielded the best result. This is a rather interesting result; it implies that longer observation might be subject to fatigue problem, and that a minimum time is required for the revelation of any personality type. Since we used two minutes as the time unit, only limited observations can be recorded within a ten-minute long experiment session. A finer grid experimental design will be needed for future studies on this timing issue, if we are to find out the true optimal length for click stream data.

Second, prior site visiting experience reduces the accuracy of our classification scheme. As we compared the results from those who had visited the 7-ELEVEN site before to those of the first timers, a lower level of performance by the ANNs was observed. In fact this finding is encouraging. Because we stress the importance of knowing potential customers based on a one-time visit of the Web site, our scheme does serve the purpose well. As to the second or third timers, using their longitudinal records might be the best way to classify them. If customers were previously classified by the current scheme, some kind of Bayesian updating on the data might greatly increase the classification accuracy.

Third, the suitability of the CAD theory to the Internet environment needs further validation. Our study adopted Horney's and Cohen's theories that were developed several decades ago. The underlying assumption of the social environment in which people interact might not hold true for the virtual communities we participate in today on the Internet. Since our PDFs were derived based on the particular traits described by the theories, their effectiveness might also have diminished as the Internet undermined the validity of the theories. Although this argument appears presumptive, our observation of the browsing behavior of the detached type does offer some anecdotal evidence. From the collected data, we saw a large percentage of subjects of the detached type actually clicked the PDFs designed for identifying the compliant type. This was most frequently observed on the social issues and express opinion factors (see Table I). We suspect that the cold, private, and reserved characteristics of the detached type are observable external ramifications of the personality, which could be molded differently by different social environments. To the extent that the Internet provides a completely different and highly anonymous social environment, the risks of confusing the detached type with the compliant type using the current PDFs could be quite high.

Conclusion and suggestions

As firms are trying to gain competitive advantages by participating in e-commerce, they must also justify the worthiness of corporate IT investments. On the marketing side, tens of millions of dollars have been poured into the development of Internet-based consumer relationship management (CRM) software, but to date, few of them could function well against the massive one-time Web surfers who are potential customers. Motivated by this shortcoming, we set out to develop a systematic method to increase drastically the usefulness of click stream analysis by harnessing the power of social psychology and Web design. Our artificial intelligence-powered new scheme, termed structured click stream analysis, has been delineated in this paper, and the satisfactory results reported. Although we did not attempt to stretch the envelope of our ANN models' performance to the maximum in the current study, we believe the now 75 per cent accuracy rate can be greatly improved by enhancing their design and training methods. For such efforts, a detailed guideline can be found in Walczak (2001).

Through data analysis we also discovered two limitations on the current research. First, the social psychology theories on personality type we adopted from the literature might have lost their observational validity in the context of the Internet, thereby, in part, weakening our scheme. The only remedy for this problem is to initiate new validation projects on the Internet, and revise the external traits of the specific personality types accordingly. The second issue stems from the fit between the sampling population and the theories. For a society like the one in which we conducted the field experiment, the practical value of differentiating personality types does not seem very high. Since there is a predominating type in the population, marketers could not commit too many mistakes by catering all products and services to these people. Having stated this, we still could not rule out a better fit between the theories and some other large societies.

One interesting finding of our work that deserves a future research agenda is the active role of click stream length in the classification of surfers. We found differential peaking effect for different personality types in the experiment, but could not explore this phenomenon further due to time and resource constraints. Our view is that not only does this specific issue need more study for the purpose of advancing online market segmentation techniques, the general issue of identifying the optimal timing for recording Web browsing data in the context of data mining also entails intensive attention.

We have just scratched the surface of a new, exciting, and important Web research question. More coordinated interdisciplinary efforts by the academia in the future will be needed, in order to bring this quest to fruition.

References

Aleksander, I. and Morton, H. (1990), *An Introduction to Neural Computing*, 1st ed., Chapman & Hall, London.

Burger, J. (1993), *Personality*, Brooks/Cole Publishing Company, Boston, MA.

Chang, S. (1999), "Preference elicitation on WWW: an improved track analysis methods", master's thesis, National Chung Cheng University, Taiwan.

Chang, S. (1998), "Internet segmentation: state-of-the-art marketing applications", *Journal of Segmentation in Marketing*, Vol. 2 No. 1, pp. 19-34.

Cohen, J. (1967), "An interpersonal orientation to the study of consumer behavior", *Journal of Marketing Research*, Vol. 4, August, pp. 270-77.

Cooper, D. and Shindler, P. (1998), *Business Research Methods*, McGraw-Hill Irwin, New York, NY.

Deitel, H., Deitel, P. and Steinbuhler, K. (2001), *E-Business and E-Commerce for Managers*, Prentice-Hall, Englewood Cliffs, NJ.

Fu, R. (1995), *Personality and Life*, Book Springs Co., Taipei.

Ha, S. and Park, S. (1998), "Application of data mining tools to hotel data mart on the intranet for database marketing", *Expert Systems with Applications*, Vol. 15 No. 1, pp. 1-31.

Hopfield, J. (1982) "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences*, Vol. 79, pp. 2554-8.

Horney, K. (1945), *Our Inner Conflicts: A Constructive Theory of Neurosis*, Norton, New York, NY.

Hsu, H. (1993), "Relationships among consumers' personality type, purchase involvement, marketing mix and purchase intent: a case of high quality stereo systems", master's thesis, National Chung Cheng University, Taiwan.

Huberman, B., Pirolli, P., Pitkow, J. and Lukose, R. (1998), "Strong regularities in world wide web surfing", *Science*, Vol. 280, April, pp. 95-7.

Kotler, P., Ang, S. and Tan, C. (1999), *Marketing Management – an Asian Perspective*, Prentice-Hall, Englewood Cliffs, NJ.

Noerager, J.P. (1979), "An assessment of CAD – a personality instrument developed specifically for marketing research", *Journal of Marketing Research*, Vol. 16, February, pp. 53-9.

Petersohn, H. (1998), "Assessment of cluster analysis and self-organizing maps", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 6 No. 2, pp. 139-49.

Rogers, J., Slama, M. and Williams, T. (1983), "An exploratory study of luscher color test predicted personality types and psychographic shopping profiles," *AMA Educator's Proceedings*, pp. 30-4.

The authors would like to thank the 7-ELEVEN Corporation of Taiwan for their generous support of our work; this research could not have been successful without the Company's permission to use and imitate its corporate Web site.

Rumelhart, D., Hinton, G. and Williams, R. (1986), "Learning internal representations by error propagation", *Parallel Distributed Processing*, Vol. I/II, MIT Press, Cambridge, MA.

Schiffman, L. and Kanuk, L. (1997), *Consumer Behavior*, Prentice-Hall, Englewood Cliffs, NJ.

Tom, K. (1996), "Listener: an AppleScript for client-side investigation of world wide web browsing behavior", working paper, Wayne State University, Detroit, MI.

Vellido, A., Lisboa, P. and Meehan, K. (1999), "Segmentation of the on-line shopping market using neural networks", *Expert Systems With Applications*, Vol. 17, pp. 303-14.

Walczak, S. (2001), "An empirical analysis of data requirements for financial forecasting with neural networks", *Journal of Management Information Systems*, Vol. 17, Spring, pp. 203-22.

Wen, K. and Liu, C. (1998), "Implementing online personalized advertisement on the Internet using neural networks", *Proceedings of the 9th International Conference on Information Management*, Taiwan, CD-ROM.

Wen, K. and Young, J. (1998), "A comparison of two Web-based preference elicitation methods through Internet advertisement", *Proceedings of the 9th International Conference on Information Management*, CD-ROM, Taiwan.

Wen, K., Cheng, T. and Han, H. (1997), "Preference elicitation using World Wide Web: an experimental study", *Proceedings of the Third Conference on MIS Research and Practice*, Taiwan, pp. 492-9.

Widrow, B. (1962), "Generalization and information storage in networks of ADALINE neurons", in Yovits, G. (Ed.), *Self-Organizing Systems*, Spartan Books, New York, NY.

Widrow, B., Rumelhart, D. and Lehr, M. (1994), "Neural networks: applications in industry, business, and science", *Communications of the ACM*, Vol. 37 No. 3, pp. 93-105.