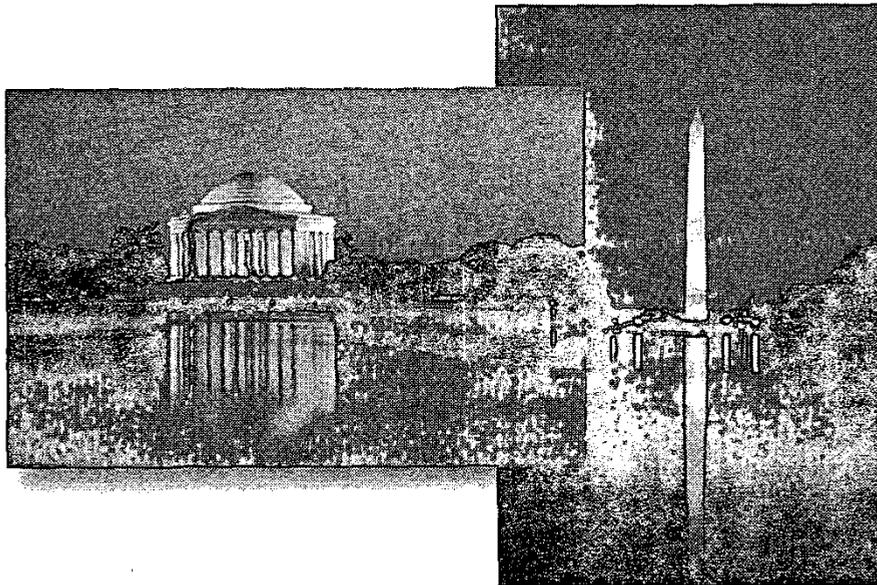# SMC '03 Conference Proceedings

2003 IEEE International Conference on Systems, Man and Cybernetics

Conference Theme
System Security and Assurance

Sponsored by the Systems, Man and Cybernetics Society of the Institute of Electrical and Electronics Engineers, Inc.

TC5: Systems and Control V
Room: Fairfax Time: 2:00pm - 3:30pm
Chair: Roberto Santiago

TC6: Intrusion Detection and Security II
Room: Potomac 4 Time: 2:00pm - 3:30pm
Chair: Marlin Mickle

TC7: Signal Processing I
Room: Kennedy Time: 2:00pm - 3:30pm
Chair: Irene Gu

Designing a Mechanism to Assure the Reasonableness and Fairness of Information Stored in Databases on WWW

Ko-Kang Chu
Dept. of Info. and Comp. Eng.,
Chung-Yuan Christian Univ.
Chung-Li, 320, Taiwan
kirk@ms2.hinet.net

Maiga Chang
Dept. of Special Education,
Chung-Yuan Christian Univ.
Chung-Li, 320, Taiwan
maiga@ms2.hinet.net

Yen-Teh Hsia
Dept. of Info. and Comp. Eng.,
Chung-Yuan Christian Univ.
Chung-Li, 320, Taiwan
hsia@ice.cycu.edu.tw

Abstract - *Web counters are usually used for recording the number of times a web page has been visited. Apart from the recording of pageviews, advertisement publishers also use Web counters to count the number of clicks made by potential customers on the Internet ads. As performance-based Internet advertisement pricing models are becoming popular and generally acceptable, there is a need from the perspective of the advertisers to ensure the accuracy of the values recorded by Web counters. This paper describes the idea of a virtual IP, and proposes a mechanism for assuring the reasonableness and fairness of the values of Web counters. The result is an accuracy counter (AC). The AC is a real system. It has been successfully tested in a related class project.*

Keywords: Web counter, Internet advertisement, performance-based pricing model.

1 Introduction

Web counters are usually used for recording the number of times a Web page has been visited. The value of a Web counter could represent the importance of the page on which the counter is located. There are many companies that provide counter services. If we use the counter service of such a company, we can put a counter on our homepage, and the value of the counter will be stored in the database of the company. When the value of a counter is explicitly shown in a Web page, we call it a "visible type" counter.

An "invisible type" counter, on the other hand, is one that does not have its value shown in the Web page on which the counter is located. In the past, researches found that CPC (Cost Per Click) and CTR (Click Through Rates) are both fair and the easiest Internet advertisement pricing model [2][3][5][6]. Since then, whenever an advertiser rents a place to put its advertisement, the advertisement publisher would install a counter for each advertisement and charge the company

according to the value of the counter. As the name suggests, such counters are invisible, and users do not know their values.

The three elements - a counter, a Web page (on which the counter is located) and users (who visit the Web page) - constitute a space. We call this space a *counting environment*, and we divide all counting environments into two categories: *conflicting environments* and *non-conflicting environments*. In a non-conflicting environment, we generally do not pay too much attention to the value of the counter, be it high or low, or even reset to zero. This is not the case in a conflicting environment. When the CPC technique is used, we have a typical example of a conflicting environment. In this environment, the ad publisher would like the value of the counter to be as high as possible. The advertiser, on the other hand, would like the counter value to be not only high but also reasonable. Most likely, the advertiser will be unhappy if the counter value is in some way "faked".

A straightforwardly implemented counter system cannot prevent its value from being either unreasonably or malicious increased [1]. For example, by using a special program that reloads the Web page automatically, we can easily increase the value of its associated counter. Phenomena like this often happen in a conflicting environment. When the underlying counter technology cannot handle unreasonable connections in a satisfactory way, the counter used in a conflicting environment has to be made invisible in order to avoid unnecessary disputes.

To improve the accuracy of counters used in conflicting environments, we propose a counter system that ensures the reasonableness and correctness of the counter's value. In addition to providing the usual functionality, we add to the system a mechanism that detects unreasonable connections. Theoretically, there is no way we can detect all forms of unreasonable connections. Nevertheless, the method that we use is

capable of detecting most kinds of unreasonable connections that are seen on today's Internet. As a result, the counting service that we offer is arguably fairer than its traditional counterpart. We call this proposed counter service *accuracy counter* (AC for short).

The remaining part of this paper is organized as follows. Section 2 motivates our research. Section 3 gives the detail of the AC system. Section 4 describes an experiment and its result. Section 5 gives a discussion and concludes.

2 Internet Ad, Pricing Models and Web Counters

As the Internet is becoming more and more popular and the browser technology is also matured, many businesses are using the Internet as a place for advertising. According to *the Year 2002 Internet Ad Revenue Report* [12] published by the Interactive Advertising Bureau (IAB) and Pricewaterhouse Coopers (PwC), the revenue of the online advertisement market in U.S. was 5.95 billions in 2002. Although the ad revenue of U.S. in 2002 was 17% less than its 2001 counterpart, the Internet advertisement market in Year 2002 still grew steadily and reached a high record of 200 millions in U.K., covering 1.4% of its advertisement market [11]. Beside the revenue in ad, IAB and PwC also discovered that businesses of most advertisement publishers were growing year by year and were also profitable. Therefore, the standard of Internet advertisement pricing models for advertisers is a big issue. It is obvious that such a standard should be well designed.

One of the major issues of a pricing model is the measurement of the size of an advertisement. The Ad Sizes Task Force (ASTF) in IAB focuses on developing a standard format of Internet advertisement that advertisers and publishers really need. ASTF proposed Banner in 1996 and conducted 3 categories (involving 15 sizes) for Internet advertisement, called Interactive Marketing Units [7]. They review and update the standard every two years [17]. In August 2002, ASTF started a series of jobs in conducting a whole new standard of ad format in order to reduce the cost spent and the planning time took by advertisers.

With the help of American Association of Advertising Agencies, ASTF proposed in December 2002 the Universal Ad Package (UAP) in which there are only four ad sizes [8]. The UAP standard was provided to advertisers and publishers, and comments were requested. Then, in April of 2003, UAP was officially announced as the standard of ad format. Most advertisement publishers, including AOL, MarketWatch.com, USAToday.com, Wall Street Journal Online, New York Times Digital, MSN, CBS and Yahoo!, will have to comply with the standard (UAP) before either April 2004 or October 2004 [16].

Once the ad format has been standardized, publishers can then charge advertisers according to ad size. Several Internet advertisement pricing models exist. Of these pricing models are flat fee, impression-based and performance-based. The flat fee model, for example, lets advertisers choose the position and webpage he/she wishes to place his/her ad. Different positions and webpages would mean different prices. Other pricing considerations include issues such as publishing duration and whether the space is shared with other advertizers.

Take Hinet, the largest Internet Service Provider (ISP) in Taiwan, as an example. To publish a standard Button-size ad (150×50 pixels) on the homepage of Hinet, the advertiser would have to pay 70,000 NTD (2,000 USD) for each week, while sharing the same space with another 5 advertisers [9]. Placing an ad with a less prominent publishers is only a littl bit less expensive. Publishing a Button-size ad (100×32 pixels) on the homepage of Sony Network Taiwan Ltd., for example, requires 50,000 NTD (1,430 USD) for each week [18]. Looking at these example prices, we can understand how expensive it can be should an advertiser decides to take the flat-fee rate.

The impression-based pricing model offers an alternative for the advertisers. With this pricing model, the advertisers can more or less ensure that potential customers really saw the ad that they have paid for. Here, the number of impressions is defined to be the number of pageviews of the webpage in question. According to the statistics collected by IAB in 2002, Cost Per thousand iMpressions (CPM) is the predominant pricing model of choice in year 2002 [13][15]. Let us use the same ad size in the above examples to make a comparison. To publish a standard Button-size ad using the CPM pricing model, the fee charged by Hinet and So-net would be around 70 to 350 NTD/CPM (2 to 10 USD/CPM). To place the same ad in Hong Kong, on the other hand, the fee will be 0.8 to 15 USD/CPM [10].

The impression-based pricing model appears to be a fair model for the advertisers. However, the fact that there are one thousand impressions does not necessarily mean that there has been one thousand "visits" to the website/webpage by potential customers. On the extreme side, it may even be the case that there was not one single customer who really clicked on the ad. Because there are ways to view a page without having to click its corresponding ad. (Just witness the working of a robot.) Consequently, performance-based pricing models are becoming more popular and acceptable to both the

advertisers and publishers. Statistics shows that 21% of the Internet advertisement used performance-based pricing models in year 2002 (as compared to 12% in 2001).

There are different kinds of performance-based pricing models, including Cost Per Click (CPC), Click Through Rate (CTR), Click Per Lead (CPL), Click Per Sale (CPS) and hybrid models [14][4]. The most commonly used performance-based pricing model is CTR. CTR is charged by the ratio between the number of clicks on the ad and the number of impressions of the page on which the ad is located. Put it another way, the publishers charge the advertisers according to the percentage that an ad was clicked by customers when the page on which the ad is located is viewed (presumably by the customers). Though both CTR and CPC appear to be fair to the advertisers, it is nevertheless possible that something to the disadvantage of the advertisers may occur. For example, some one may hire an intern or a part-time worker to repeatedly click on an ad. This would have a direct and "faked" effect on a CPC-based environment. It could also have an indirect but still "faked" effect on a CTR-based environment.

In general, it is hard for an advertiser to give 100% trust to the performance report furnished by a publisher. What is in need is to further enhance the accuracy of the number of clicks as recorded in a Web counter, so as to increase the degree of trust on the part of the advertisers in the publishing mechanism. An accuracy counter (described in Section 3) is designed to do just that. It provides a more robust and accurate data to the advertisers.

3 The Accuracy Counter System

In general, a user will tend not to repeatedly access the same Web page within a short time interval. Therefore, when the same computer accesses the same Web page again in a very short time, it is feasible to consider the connection attempt as "unreasonable". The AC system is designed to block increments of the counter's value under such circumstances.

Let w be a Web page on which an accuracy counter a is located. $|a|$ is the value of a. Let C be the set of all clients that were/are/will-be connected to w. c_m is an element of C. $c_m w$ represents a connection between c_m and w. $t_i^{c_m w}$, the i-th time when c_m "stops" its connection with w, is either the time when c_m closes the browser or the time when c_m switches the being browsed Web page from w to another page.

Definition (Unreasonable Connection)

An unreasonable connection occurs at $t_{now}^{c_m w}$ if and only if

$$t_{now}^{c_m w} - t_{now-1}^{c_m w} < \delta \qquad (1)$$

where δ is a pre-specified, fixed time interval.□

Note that $t_i^{c_m w}$ is defined not as the (i-th) time when a connection is requested, but as the (i-th) time when a "page viewing operation" is officially terminated. This may be quite different from what our intuitions may suggest. Below, when the situation permits, we will use t_i instead of $t_i^{c_m w}$.
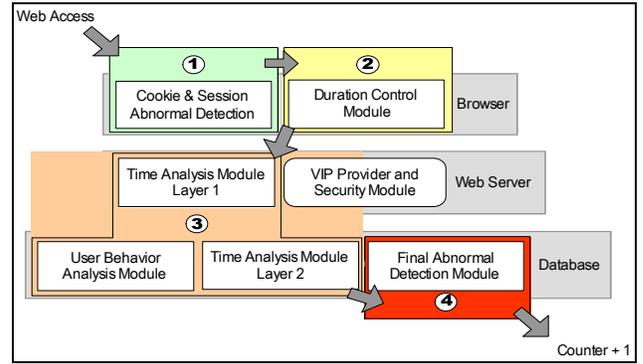
Figure 1. Three-tier architecture of the AC system.

Cookie detection and VIP assignment

What the definition of unreasonable connection says is as follows. If we could detect the differences among c_1, $c_2 \dots c_n$ and save for each c_i ($1 \leq i \leq n$) its corresponding t_{now-1}, we could then check whether t_{now} is reasonable. With today's Web architecture, we generally identify c_i by its IP address. However, there is a limitation in the use of IP addresses. If c_i is located behind a proxy, a NAT or a firewall, then the IP address that the Web server acquires is only the address of the proxy, the NAT or the firewall. In other words, it is not enough to just identify each c_i by its IP address. We need to find another useful source of information for identifying each c_i.

Although we cannot identify c_i just by acquiring its IP address, the idea that each computer has its own "identification number" is nevertheless worth pursuing. What we propose to do then is to give, at the first connection time, each computer a special identification number, called a Virtual IP (VIP for short). VIP is stored in a cookie located in c_i, and our design guarantees that no two VIPs will be the same. For convenience, we use Real IP (RIP for short) to refer to the original ISO OSI defined IP address.

In general, a cookie is only a text file and is not encrypted, making its security very weak. To take security problems into considerations, each data that we are to store in the cookie must be encrypted first. Note that we only store the assigned VIP and t_{now-1} in a cookie. t_{now-1}, on the other hand, is not only stored in the cookie but also stored in the database of the AC system.

Satisfaction of the threshold

Below, we use ct_{now-1} for t_{now-1} that is stored in the cookie, and we use dt_{now-1} for t_{now-1} that is stored in the database. As ct_{now-1} is stored in a cookie, a CGI program can retrieve it directly, thereby reducing the loading of the database. Normally, ct_{now-1} is equal to dt_{now-1}. However, a cookie file can be very easily modified or deleted. Therefore, we may have to resort to the use of dt_{now-1} when necessary.

In Figure 1, the Time Analysis Module Layer 1 is in charge of checking ct_{now-1} at the Web server, while the Time Analysis Module Layer 2 is in charge of checking dt_{now-1} at the database layer. If we find ct_{now-1} in the cookie, then we use it to check whether t_{now} is reasonable. If t_{now} is found to be unreasonable, then the checking stops (and the counter will not be incremented). But if t_{now} is found to be reasonable, then we still need to double-check it using dt_{now-1} stored in the database. The double-checking is needed here, because there is always a chance that values stored in the cookie be maliciously modified. The algorithm is as follows:

```
if  tnow - ctnow-1 < δ {
     /* Time Analysis Module Layer 1 */
     UNREASONABLE;
} else if tnow - dtnow-1 < δ {
     /* Time Analysis Module Layer 2 */
     UNREASONABLE;
} else {  // both checks are passed
     /* other checks before possible increment of |a|  */
}
```

Computation of page duration time

To our knowledge, all counters of today work in essentially the same way, by adding a value of one at about the same time when the Web server receives a request for the being advertised page. But if we are to take the factor of effective time-span into consideration, current Web counters no longer satisfy our need. Essentially, the concept of effective time-span is that a potential customer should at least "stay" on the same page for a minimum amount of time (e.g., 20 seconds) in order to make a "least meaningful" sense out of the page. If the time he/she "stays" on a Web page is below the minimum threshold, then it is ineligible to increment the counter.

Consequently, the time of counter increment should only occur when the user "exits" a Web page (i.e., when the page viewing operation officially terminates), and not when the user "enters" a Web page (i.e., when the page viewing operation begins). This is so that the system is able to compute the length of time that the user has "stayed" on the page. If the time span is so short that it is below the threshold that the AC system manager specifies, then the AC system simply considers t_{now} as invalid, and the counter will get incremented. The following JavaScript code implements the idea of checking effective time span when the user exits *w*.

```
<script language="javascript">
var enter_time = new Date;  // the time when page viewing begins
function exit()  // called when the user exits this Web page
{
    exit_time = new Date;  // the time when page viewing ends
    k = Math.ceil((exit_time.getTime() - enter_time.getTime()) / 1000);
    if (k > 10) { // duration time check passed
      /* do the remaining checking about counter increment */
      window.open ('exit.asp?stay=' + k, '', '');
    }
}
</script>
<body onUnLoad="exit()">
```

Further user behavior analysis

If there is some one in the counting environment that wants the value of the Web counter to be as high as possible, it is entirely possible that the person may use a program that helps him access the target Web page automatically day and night. One popular and easy way is (for example) to put the following html code in a file called index.html.

```
<meta http-equiv="refresh" content="30; url=index.html">
```

And so, for every 30 minutes, the web page will refresh automatically, and |*a*| will be incremented continuously. When a situation like this happens, we need a countermeasure that can effectively deal with it.

Consider t_{now}, t_{now-1} and t_{now-2}. If the following rule is not satisfied, then t_{now} is considered invalid (and the counter will not be incremented).

$$\left|(t_{now} - t_{now-1}) - (t_{now-1} - t_{now-2})\right| > \varepsilon \qquad (2)$$

ε is the allowable tolerance. $\square$

Obviously, this countermeasure of ours can only handle the kind of html code that we just mentioned above. If the proposed AC system ever becomes popular

in the future, smarter programs will be designed to deceive this counting system. When that happens, countermeasures equipped with more powerful heuristics would have to be installed.

4   Testing and Evaluation

To better understand the survival rate and performance of the AC system, we established a conflicting environment to test it. The system was tested for about four months. The conflicting environment that we use originated from a class project of a Web page designing course in Chung Yuan Christian University. There were four such classes and forty teams altogether. Each of the teams must design a website at the end of semester. We told the teams that we will install an AC on their homepages and that the final score will be related to the final values of the counters as well as the average time-span among the page requests. In calculating the average time-span, we dropped the extreme values.

In addition to an AC, we also put a traditional counter (TC for short) on each homepage. However, the TC was not visible. We provided enough information about the AC, and we said nothing about the TC. We made every effort to ensure that there was no way for any team to know the existence of the TC.

We set δ to be 15 minutes and time-span threshold to be 10 seconds. At the end of the test, the blocking rate of AC was 61%.[1]

One way to interpret our result is perhaps this. If it is required that an AC (instead of a TC) be used on a Web page, we may be able to help the advertising company save 61% advertising fee. On the other hand, it means that the advertising agent would lose 61% of its income. This, unfortunately, is the very nature of a conflicting environment.

5   Discussion and Conclusion

A traditional counter cannot satisfy the need of a conflicting environment, because a simple technique of refreshing can easily increase the value of a counter automatically and continuously. What we do in this paper is to develop a new counter technology that can effectively block unreasonable connections. The resulting counter is called an accuracy counter. The AC system is able to better satisfy the need of a conflicting environment.

The proposed AC system uses VIP for the purpose of identifying each c_i. If there is no VIP, the system would use RIP for identification purposes and also use RIP to check whether t_{now} is reasonable. In the meantime, the system would also assign a VIP to the computer without VIP, in order that the computer is "VIP identifiable" next time around. That is to say, even if a previously assigned VIP is deleted, whether deliberately or accidentally, the system can still use RIP for identification and checking purposes. Since the use of "RIP identification" is always temporary, we can tolerate its drawback.

As a matter of fact, it is not even advantageous for someone to deliberately delete the assigned VIP of his/her computer when the computer in question is located behind a proxy, a NAT or a firewall. The system uses RIP to identify each computer whenever it cannot find the VIP of the computer. Since the RIP of all computers located behind the same proxy, NAT or firewall will be (from the Web server's point of view) the same, the system will treat all page requests from this same cluster of computers as belonging to one specific computer, causing more page requests to be judged as unreasonable.

The main problem, however, is that if there are no VIP found and there is also no proxy, NAT or firewall. What is more, suppose that the computer changes its IP address all the time (e.g., via the working of ADSL). When all of these happen, the AC system indeed would view each page request of this computer as a first connection to the website and, as a result, would increase $|a|$. We have not encountered such situations yet, but we know it could exist. If necessary, the AC system can view all IP under class C even class B network as only one IP. Extreme as it may be, the problem is solved as soon as the "first connecting" computer is assigned a VIP by the AC system.

We tested the AC system for its effectiveness. The conflicting environment that we use originated from a class project of a Web page designing course in Chung Yuan Christian University. There were four classes and forty teams. At the end of the test, we acquired 61% blocking rate. We believe that if the manager of a conflicting environment is willing to use the AC system for counting, the resulting value of the counter would be fairer.

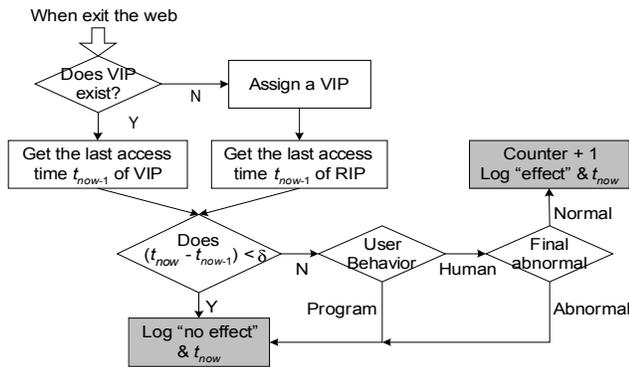[1] The formula is $(GC - AC)/GC \times 100\%$

Figure 2. The operation flow of the AC system.

References

[1] Anonymous, "Measuring click-through rate is unreliable," *Internet Business News*, Coventry, Aug 4, p. 1, 2000

[2] Michelle E. Bayles, "Designing Online Banner Advertisements: Should We Animate?" *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, Minneapolis, Minnesota, USA, 2002. ACM Press. NY: New York, USA. pp. 363 – 366, 2002

[3] Loren Fox, "The banner: unsung hero," Upside, Foster City, Vol. 12, Iss. 4, pp. 170-176, Apr. 2000

[4] Nicole Sperling, "Traditional rates clash with cost-per-click ad model," *Advertising Age*, Chicago, Vol. 69, Iss. 46, p. 50, Nov. 16, 1998

[5] Terry Sweeney, "Click-Through Rates on online ads are declining, but true believers say there's no place like the Web to build the brand," *InternetWeek*, Manhasset. Iss. 832, p. PG.57, Oct. 9, 2000

[6] Debra Aho Williams, "New ammo for click-rate debate," *Advertising Age*, Chicago, Vol. 67, Iss. 34, p. 24, Aug. 19, 1996

[7] Ad Sizes Task Force, "Interactive Marketing Units," *Interactive Advertisement Bureau*, Retrieved From http://www.iab.net/standards/adunits.asp, Retrieved by June 18, 2003

[8] Ad Sizes Task Force, "Universal Ad Package," *Interactive Advertisement Bureau*, Retrieved From http://www.iab.net/standards/uap/index.asp, Retrieved by June 18, 2003

[9] Hinet, "Advertisement Publishing Service - Fee List," *Hinet of Chunghwa Telecom Co.*, Retrieved From http://www.hinet.net/footer_index3.htm, Retrieved by June 19, 2003

[10] i818.com, "Advertisement Publishing Service - Fee List," *i818.com*, Retrieved From http://www.i818.com/sales, Retrieved by June 19, 2003

[11] Interactive Advertisement Bureau and PricewaterhouseCoopers, "Online ad spend jumps to record levels," *Interactive Advertising Bureau UK*, Retrieved From http://www.iabuk.net/index.php?class=news&view=707, Retrieved by June 18, 2003

[12] Interactive Advertisement Bureau and PricewaterhouseCoopers, (April 9, 20003), "IAB / PwC Release Top-Line Data From 2002 Internet Advertising Revenue Report," *Interactive Advertisement Bureau*, Retrieved from http://www.iab.net/news/pr_2003_04_09.asp, Retrieved by June 18, 2003

[13] Interactive Advertisement Bureau and PricewaterhouseCoopers, (June 12, 20003), "IAB / PwC Release Final Full Year 2002 Internet Ad Revenue Figures," *Interactive Advertisement Bureau*, Retrieved from http://www.iab.net/news/pr_2003_06_12.asp, Retrieved by June 18, 2003

[14] Interactive Advertisement Bureau, "Clickthrough Rate (CTR) (industry audit definition)," *Interactive Advertising Bureau UK*, Retrieved From http://www.interactivejargonguide.org/Glossary/Term/Clickthrough+Rate+(CTR)+(industry+audit+definition), Retrieved by Apr. 10, 2003

[15] Interactive Advertisement Bureau, "CPM (Cost-Per-Thousand)," *Interactive Advertising Bureau UK*, Retrieved From http://www.interactivejargonguide.org/Glossary/Term/CPM+(Cost-Per-Thousand), Retrieved by Apr. 10, 2003

[16] Interactive Advertisement Bureau, (April 28, 20003), "IAB Announces Final Interactive Universal Ad Package," *Interactive Advertisement Bureau*, Retrieved from http://www.iab.net/news/pr_2003_4_28.asp, Interactive Advertisement Bureau, Retrieved by June 18, 2003

[17] Interactive Advertising Bureau UK Ltd., "Ad formats - we want to hear your views," *Interactive Advertising Bureau UK*, Retrieved From http://www.iabuk.net/index.php?class=news&view=698, Retrieved by June 18, 2003

[18] So-net, "Advertisement Publishing Service - Fee List," *Sony Network Taiwan Limited*, Retrieved From http://www.so-net.net.tw/ad/description.html, Retrieved by June 19, 2003