2009 Ninth International Conference on Intelligent Systems Design and Applications

Agent-based Web Content Engagement Time (WCET) Analyzer on e-Publication System

Raymond S. T. Lee, James N. K. Liu, Karo S. Y. Yeung, Alan H. L. Sin and Dennis T. F. Shum

Abstract—**This paper focuses on the adoption of Agent Technology to calculate and evaluate the Web Content Engagement Time (WCET). Traditional Web traffic analysis metrics such as pageviews, unique browser, visitor loyalty, etc have been used to analyze the web traffic behaviour for a long time since the birth of World Wide Web, but the emersion of software robots and web crawlers trigger a huge impact on the integrity and correctness of these traditional Web statistics. For advertisers, these statistics are not enough for them to evaluate the actual return-on-investment (ROI). For instance, large amount of pageviews but extremely short session duration will not have much impact for the advertisers to promote their products and brands. Web Content Engagement Time (WCET) for the reading on interactive Web content such as e-magazines and e-publications, which focuses on the page duration between each "page-flipping", will give advertisers much more information and confidence on whether such eye-balls (i.e. attention) are actually focused on the Web content (and hence the eAds) or not, especially during the browsing of e-magazines and e-publications. But such indicator involves significant amount of calculation within the Web server, especially when over thousands of users are reading a popular e-publication at the same time. To tackle with this problem, a multi-agent based Web Content Engagement Time (WCET) Analyzer is proposed on e-publication system. From the experimental perspective, popular Chinese e-magazine "MingPaoWeekly", with over 0.5 million readership in Hong Kong and overseas Chinese communities are tested over the IAToLife.com Web Channel platform, promising Web Content Engagement Time (WCET) is recorded, which provides not only integrity and confidence for the publishers and advertisers, but also shines a new light for the future agent-based target marketing and e-reader profile and reading behavior analysis.**

I. INTRODUCTION

The popularity of E-publication (E-Pub) is growing rapidly for the past few years. Such new form of publication provides rooms for advertisers to actually interact with the readers, and hence obtain the readers' profile, but more importantly to analyze the readers' reading habit for the target-marketing and related promotion and activities.

Different from traditional Web system which relies solely on Page Views and Unique Visitors to evaluate the popularity of a Web site, with potential integrity threats caused by software robots and Web crawlers. E-Publications, with intrinsic human-like "page flipping" characteristic, can effectively solve this integrity problem which provides confidence both for the publisher to evaluate the "actual readership" of these e-publications, but also the advertisers to evaluate the "effectiveness" of their eAds.

To achieve this, one important metric introduced in e-Publication system is the Web Content Engagement Time (WCET), which records the duration of the e-reader between the flipping of every page within the e-publication (e.g. e-magazines or e-books). From the advertisers perspective, such figures can provide a solid metric of the degree of attention and popularity of a particular eAd within the e-publication, and hence the effectiveness of the eAds.

However, the traditional client-server technology used for web is not suitable for analyzing the WCET. The problems can be categorized into two folds. Firstly, it is the intrinsic limitation of client-server best-effort based Web technology which resulted in data lost and inaccurate measurement. Secondly, it is the inefficiency of the client-server technology which caused heavy network traffic for the calculation of WCET especially when over thousands (or million) concurrent users are reading a popular e-publication such as e-Newsweek or MingPaoWeekly at the same time over the world.

To tackle this problem, this paper introduces the multi-agent based Web Content Engagement Time (WCET) Analyzer on e-Publication System which adopts agent technology to solve the problems in traditional analytics technology and provide a more accurate measurement and analysis result.

Structure of paper is so follow: We first introduce some background knowledge and study, followed by proposed methodology. We will implement the method and provide some experimental results. Lastly, we conclude the paper and future work is suggested.

II. BACKGROUND STUDY

A. Web Analysis

According to Web Analytics Association (WAA), terms related to web analysis are defined in Table I.

In particular, "Page Views" and "Visit Duration" are the most frequently adopted metrics to evaluate the popularity of a particular Web site. However, these figures are not sufficient especially for advertisers because "Page Views" can only reflect the head-counts, but not how long the reader pay attention on a particular advertisement, while "Visit Duration" can only focus on the time within one session, but

This research work was supported in part by the CRG grant 1-ZV41 of Hong Kong Polytechnic University and RGC grant B-Q13J

Dr. Raymond S. T. Lee is the Founder and President of IATOPIA Research Centre, IATOPIA.com Ltd., Hong Kong.

Dr. James N. K. Liu is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.

Mr. Karo S. Y. Yeung, Mr. Alan H. L. Sin and Mr. Dennis T. F. Shum are with IATOPIA Research Centre, IATOPIA.com Ltd., Hong Kong.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.189

not the actual time spent on a particular page or interested area, for example, a particular article page or an advertisement page. With the existing Web crawlers and software robots which automatically generate extra Web traffic, the integrity of such figures and statistics are more questionable.

TABLE I
WEB ANALYTICS ASSOCIATION WEB ANALYTICS DEFINITIONS [1]

Term	**WAA Definition**
Page	A page is an analyst definable unit of content.
Page Views	The number of times a page (an analyst-definable unit of content) was viewed.
Visit / Sessions	A visit is an interaction, by an individual, with a website consisting of one or more requests for an analyst-definable unit of content (i.e. "page view"). If an individual has not taken another action (typically additional page views) on the site within a specified time period, the visit session will terminate.
Unique Visitors	The number of inferred individual people (filtered for spiders and robots), within a designated reporting timeframe, with activity consisting of one or more visits to a site. Each individual is counted only once in the unique visitor measure for the reporting period.
New Visitor	The number of Unique Visitors with activity including a first-ever Visit to a site during a reporting period.
Repeat Visitor	The number of Unique Visitors with activity consisting of two or more Visits to a site during a reporting period.
Entry Page	The first page of a visit.
Landing Page	A page intended to identify the beginning of the user experience resulting from a defined marketing effort.
Exit Page	The last page on a site accessed during a visit, signifying the end of a visit/session.
Visit Duration	The length of time in a session. Calculation is typically the timestamp of the last activity in the session minus the timestamp of the first activity of the session.
Referrer	The referrer is the page URL that originally generated the request for the current page view or object.
Internal Referrer	The internal referrer is a page URL that is internal to the website or a web-property within the website as defined by the user.
External Referrer	The external referrer is a page URL where the traffic is external or outside of the website or a web property defined by the user.
Visit Referrer	The visit referrer is the first referrer in a session, whether internal, external or null.
Original Referrer	The original referrer is the first referrer in a visitor's first session, whether internal, external or null.
Click-through	Number of times a link was clicked by a visitor.
Click-through Rate/Ratio	The number of click-throughs for a specific link divided by the number of times that link was viewed.
Page Views per Visit	The number of page views in a reporting period divided by number of visits in the same reporting period.
Page Exit Ratio	Number of exits from a page divided by total number of page views of that page.
Single-Page Visits	Visits that consist of one page regardless of the number of times the page was viewed.
Single Page	Visits that consist of one page view.
Bounce Rate	Single page view visits divided by entry pages.
Event	Any logged or recorded action that has a specific date and time assigned to it by either the browser or server.
Conversion	A visitor completing a target action.

Traditionally, "cookies" are used as a tool to track visitors.

But according to Jupiter Research, 58 percent of users delete their cookies regularly, with 40 percent deleting them every month. [2] Which means that metrics relying on tracking visitors via cookies are not as reliable as people believe.

Below are some popular terms used in web analysis, but each has its own downside. See below.

1) Bounce Rate / Single Page Visit

One of the arguments about web analysis is "Bounce Rate" or "Single Page Visit". Many web metrics program cannot compute the time duration people have spent on the page or on website. It certainly records the page request and the start of the session, but without the second page visit, it does not know how long someone has been there without the second page visit.

2) Last Page Visit

Similar to Bounce Rate, since time stamp is only recorded when a page within the website is loaded, when the user exits from the website by closing his/her browser, type in a URL of a different site or click on a link on site to go to different site, no time stamp will be recorded. Thus the time spent on the last page cannot be recorded and be taken into account.

3) Idle Browser

"Idle Browser" means the user just leave the browser unattended but not closing it. Many session-based metrics have a default time-out limit of 30 minutes, which means a new page view will be counted if a page is opened for more than 30 minutes.

B. Autonomy

"Autonomy" refers to systems capable of operating in the real-world environment without any form of external control for an extended period of time [10]. Thus, living systems are the prototypes of autonomous system, they can survive in a dynamic environment for some extended period, maintain their internal structures and processes, use the environment to locate and obtain materials for sustenance, and exhibit a variety of behaviours. They are also, within the limit, capable of adapting to environmental change. "Capable of adapting" implies that these systems perform some function or task. This function may be that intended by their human creator, or it may be an unexpected, emergent behaviour. As these systems become more complex, they are likely to exhibit more and more unexpected behaviours.

This is known that many autonomous systems are not fully autonomous, within the scope of the preceding definition. They are not capable of surviving and performing useful tasks in the real world for an extended period, except under highly structured situations. However, if the environment is sufficiently stable and the disturbances to it are not too severe, it can indeed survive and perform useful tasks for an extended period.

There are autonomy systems in physical world and in software. In physical world, they called "Robot", in software, they named "agent".

C. Agent Technology

The term 'agent' is distinguished for two general usages [3], weak and strong notion of agency:

1) Weak Notion of Agency

Weak Notion of Agency is perhaps the most general way in which the term agent is used to denote hardware or software-based computer system that enjoys the following properties:

Autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state [4].

Social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language [5].

Reactivity: agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it.

Pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behaviour by taking the initiative.

2) Stronger Notion of Agency

Stronger Notion of Agency means a computer system that, in addition to having the properties identified above, is either conceptualized or implemented using concepts that are more usually applied to humans. For example, it is quite common in AI to characterize an agent using mentalistic notions, such as knowledge, belief, intention, and obligation [6]. Some AI researchers have gone further, and considered emotional agents [7]. Another way of giving agents human-like attributes is to represent them visually, perhaps by using a cartoon-like graphical icon or an animated face [8] — for obvious reasons, such agents are of particular importance to those interested in human-computer interfaces. This kind of agent will be the focus in this paper.

3) Network Performance

Frequent transfer of data from client to server will sometimes overload the network and the performance will also be limited by the network traffic. Synchronous mode of data transfer can be a waste of resource as user may just want to know about the tread of the performance but not the most up-to-date information.

As agent will share most of the calculation, server is mainly used as a centralized storage, server loading, thus, can be reduced. Since the increase in the number of clients will not affect the loading of server, the scalability of the whole system can be greatly enhanced.

Since server will not need the most up-to-date information from client for calculation, if network congestion occurs, clients can temporarily store the calculated result and send it to server later. On the other hand, a loss of one particular calculated result will not affect the whole statistics as the whole client record will not be counted.

4) Agent Environment

There are mainly 4 definitions added on agent [9]:

An agent owner, (AO) is the human or artificial agent that has the power to launch the agent, as well as make the decision whether the agent should be shut down or be assigned new preferences. The owner expresses its preferences to the agent, and gets it to work toward the given preferences.

An agent designer, (AD) is the human or artificial agent that has designed and possibly implemented the control mechanism of an agent. With control, we mean the internal evaluation of the environment and the owner preferences.

A designer of an environment, (ED) is the human or artificial agent that has designed and possibly implemented the rules and conditions under which agents are able to act in the environment.

An environment owner, (EO) is the human or artificial agent whose run-time preferences are reflected in the dynamics of the rules and the conditions under which agents are able to act in the environment.

III. Methodology

For e-Publication such as e-magazine, WCET on one singe target page is defined as the duration of display of target page and its adjacent page by flipping action, if any (for instance, cover page and back-cover page have no adjacent page). For example, if one is reading an article on left page where right page is an advertisement, the engagement time of the advertisement page is the same as the article page. We first defined some terms used in e-Magazine or e-Publication system.

A. Terms used in e-Magazine or e-Publication System

eMag Readership – Number of Readers per Issue

Unique Reader – Number of IPs visited per Issue

eMag Readership – Overall Number of Readers

eMag Engagement Time – Sum of Engagement Time per Issue

Average Engagement Time – Average Reading Time per Issue per Day

Maximum Engagement Time – Maximum Reading Time per Issue per Day

Minimum Engagement Time – Minimum Reading Time per Issue per Day

Page View – Number of Display of target page

Web Content Engagement Time – Duration of Display of target page and its adjacent page, including Bounce

Bounce – Number of Display of target page and its adjacent page without Flipping, with Zero Duration

In particular, Web Content Engagement Time (WCET) is the focus among all. This metric reflects the combination effect of reader loyalty and page views. WCET gives us how long the reader is reading a page, reflecting the true loyalty of readers on a particular page or section, thus implying which publication or which section of the publications is most interesting to the readers. This index is beneficial for both publishers and advertisers, or even readers.

1) Bounce / Single Page Visit

The case of bounce or single page visit on e-Magazine or e-Publication, that means the user open the cover page, or any two of the pages, and then close the browser. In this case, a close event signal will be fired before the browser close, the event is recorded by the agent and thus engagement time can be calculated.

2) Last Page Visit

Situation is similar for Single Page Visit. Before the user close the browser, a close event signal will be fired for agent to record. Thus engagement time can be calculated.

3) Idle Browser

In case of e-Magazine or e-Publication, especially professional documents, this is quite normal for reader will read the document for more than 30 minutes, so this is not suitable for a session time of maximum 30 minutes. Unlike the traditional cookie-based methodology, which is not reliable as mentioned, an agent-based methodology is proposed.

B. Agent-based methodology versus Server-Client methodology

The proposed methodology mainly uses agents resided in client to collect the data, manipulate it and send the final result to the centralized server. Fig. 1 and Fig. 2 illustrate the difference between the traditional and proposed methodology. The server, on the other hand, is used as repository. Different agents are collaborated as follow.

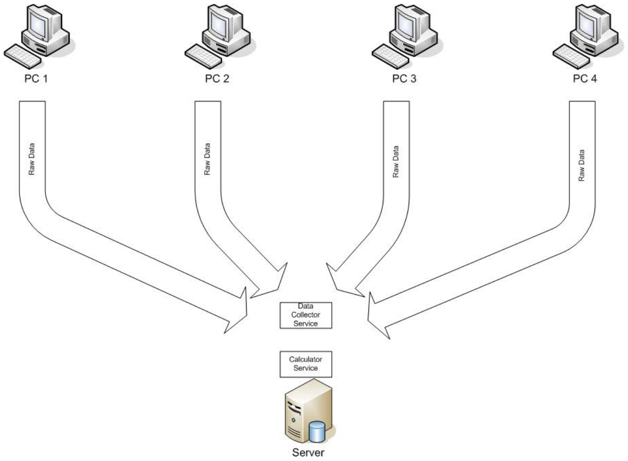


Fig. 1. Server-Client Methodology

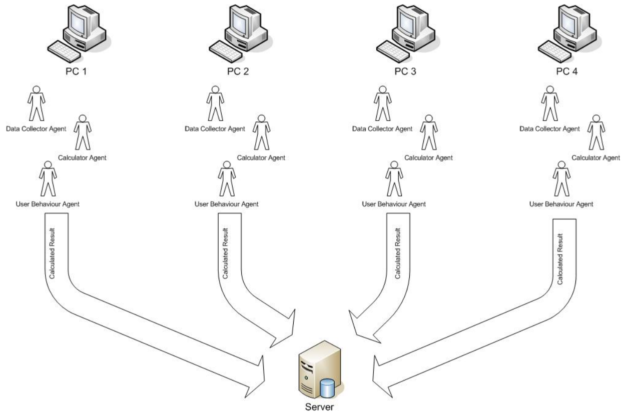


Fig. 2. Agent-based Methodology

1) Data Collector Agent

Data Collector Agent is used to collect user events such as flipping event in e-Publication case. The agent listens to page display and leave signal, record the timestamp, and send the timestamp, as well as publisher information, issue information, page information, user's IP, etc, to calculator agent for processing.

2) Calculator Agent

After receiveing the required information, **Calculator Agent** will evaluate the total time spent on a page, that is, Engagement Time, using the timestamps recorded. Calculator Agent will also compute the eMag Readership, Unique Reader, eMag Engagement Time and Page View. This information will be sent to centralized server for recording as well as to User Behaviour Agent for further processing.

3) User Behaviour Agent

User Behaviour Agent will use information sent from Calculator Agent for manipulation. Which section of pages the user reads most often? Which page the user stays for the longest time? What publication or which issue the user likes the most? Is there any reading habit about the user? All these information can be collected and evaluated by this agent.

Fig. 3 depicts how agents are communicated.

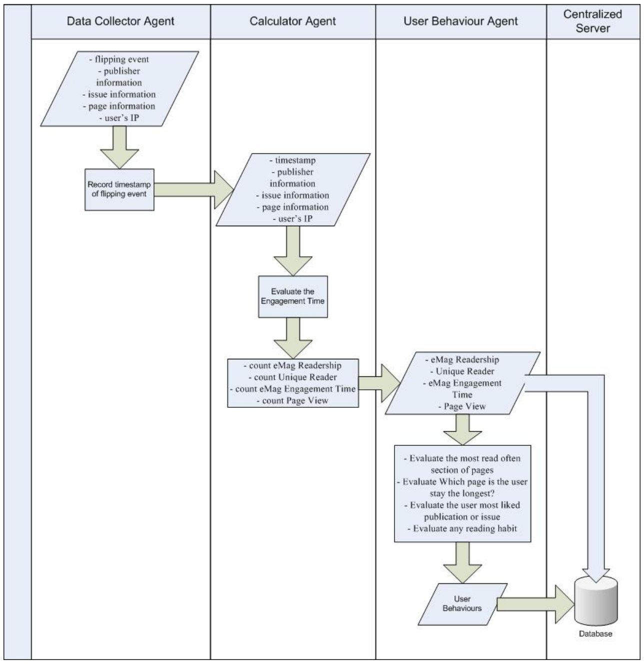


Fig. 3. Communications between agents

C. How Advertisers evaluate their ROI

After the information is analyzed by the Calculator Agent and User Behaviour Agent, all these information will be sent to the centralized server. For the advertisers, who always rely on readership on traditional publication for evaluation, the result is not accurate and lack of detailed information. For example, the questions asked in the pervious section cannot be answered using the traditional statistical method.

Advertisers can check statistics about their own advertisements on the centralized server. They can evaluate whether their eAds within the e-magazine are viewed by the readers or not, and how long the readers are paying attention to their eAds. Afterwards, advertisers can reconsider their marketing strategy such as relocating the advertisements, targeting on the most viewed or most engaged section of publication, or putting more affects on the profitable part.

D. How Publishers evaluate their Readership

For the publishers, this information can be used to evaluate whether particular section or topic in the publication is most viewed or most engaged. Authors who are most popular can be found also. Publishers can then recompose the structure for better user experience, enhance the most attractive section or remove the least one.

E. How Engagement Time can be used as a platform to evaluate e-publication content performance

Engagement Time can be used as a tool like page views or loyalty to evaluate the quality of web content or e-publication content performance. As in normal case, readers will spend more time on quality article, if different publishers are using the same platform, Engagement Time can become an index on whether the readers are interested in particular publications or authors, so different publishers can compare their own performance over the others.

In a long run, when enough Engagement Time entries are recorded, publishers can review their sections or contents, if particular sections or contents have an attractive long period of Engagement Time, this can be used to attract advertisers for putting advertisement on their publications, set the pricing, review the authors or target what the customers actually wants.

IV. Implementation and Experimental Results

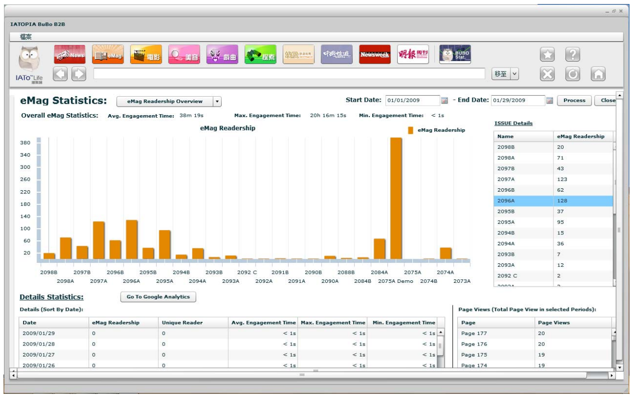
From the implementation perspective, MingPaoWeekly, one of most popular Chinese entertainment magazines with over 0.5 million readership in Hong Kong and Chinese communities in overseas, launched the e-magazine in Nov 2008 at IAToMPW.com Web Channel is an example to show how the WCET system works. The following information of a reader will be gathered in the client side:

1. A unique session id that generated whenever an issue is opened, for calculating the readership.
2. IP of the reader, for distinguishing the unique reader.
3. Duration of a page, which computed by subtracting the page load timestamp from the page unload timestamp.
4. Duration of an issue, which computed by summing up all the page durations.

The information will be sent to the centralized servers, hosted in IATOPIA Web Farm Cluster, with one datacenters (of 100Mbps broadband) in Hong Kong and Los Angeles, US. The servers will then store the information and calculate the overall statistics (engagement time, average, maximum, minimum engagement time, pageviews, readership, etc) from the information collected from clients.

A. Engagement Time and Page View

Fig. 4. shows the user interface of the WCET system within the BuBo browser. The period in January 2009 is selected for the study. The following table give some figures extracted by the WCET system:-



eMag Statistics: eMag Readership Overview

Start Date: 01/01/2009 - End Date: 01/29/2009 Process Close

Overall eMag Statistics: Avg. Engagement Time: 38m 19s Max. Engagement Time: 20h 16m 15s Min. Engagement Time: < 1s

ISSUE Details

Name	eMag Readership
2098B	20
2098A	71
2097B	43
2097A	123
2096B	62
2096A	128
2095B	37
2095A	95
2094B	15
2094A	36
2093B	7
2093A	12
2092 C	2

Details Statistics: Go To Google Analytics

Details (Sort By Date):

Date	eMag Readership	Unique Reader	Avg. Engagement Time	Max. Engagement Time	Min. Engagement Time
2009/01/29	0	0	< 1s	< 1s	< 1s
2009/01/28	0	0	< 1s	< 1s	< 1s
2009/01/27	0	0	< 1s	< 1s	< 1s
2009/01/26	0	0	< 1s	< 1s	< 1s

Page Views (Total Page View in selected Periods):

Page	Page Views
Page 177	20
Page 176	20
Page 175	19
Page 174	19

Fig. 4. WCET results

TABLE II
Extract of Average, Maximum and Minimum Engagement Time

Date	Avg. Engagement Time	Max. Engagement Time	Min. Engagement Time
2009/01/18	1h 10m 19s	2h 6m 31s	14m 8s
2009/01/21	48m 22s	1h 52m 30s	24s
2009/01/14	38m 47s	3h 16m 46s	1s
2009/01/17	36m 44s	3h 37m 2s	1m 17s
2009/01/12	32m 57s	2h 37m 3s	18s
2009/01/11	30m 28s	6h 38m 40s	< 1s
2009/01/23	16m 20s	31m 25s	1m 15s
2009/01/19	13m 6s	19m 3s	7m 9s
2009/01/10	11m 35s	58m 58s	< 1s
2009/01/15	8m 53s	36m 17s	< 1s

Sorted by Average WCET, it shows that the peak Average Engagement Time appearing on 18 January. That means each reader spent over 1 hour and 10 minutes on a particular e-magazine on average, which is a promising and surprising result as compared with the average Web page duration of the popular Web sites over the world which is talking about 10 to 15 minutes (maximum). The longest period of reading is 2 hours 6 minutes and 31 seconds. We next look for which pages are the most viewed by readers. Below is an extract of figures:

TABLE III
Extract of Page Views

Page	Page Views
Page 0	100
Page 1	100
Page 56	67
Page 57	67
Page 2	55
Page 3	55
Page 4	49
Page 5	49
Page 58	48
Page 59	48

Using a focus group of 1050 e-readers involved in reading a particular MingPaoWeekly emagazine as example. The PageView summary of a particular date is shown in Table III. The Page Views are all in pairs because the target page and its adjacent page are viewed for the same period of time. Page 1 represents the cover page and page 0 represents the blank area adjacent to cover page. The top viewed page is cover page, which is viewed 100 times, the second rank are Pages 56 and 57 involving an article of the cover story with the “neighborhood” page of an eAds. Such figures are not only important to the publisher, but also to the advertisers to evaluate the impact of their eAds and promotions.

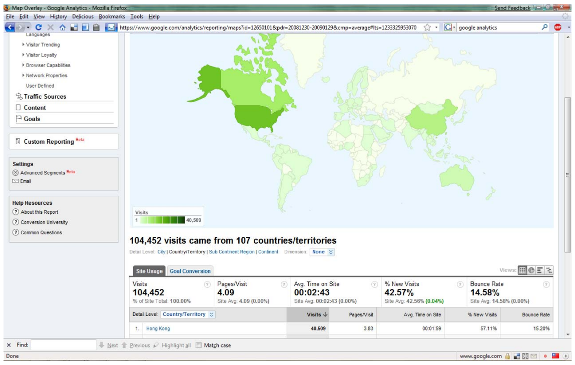
B. *Demographic information*

Fig. 5 and Table IV show the demographic information of all the visitors visiting the IAToMPW.com Web Channel, for viewing the MingPaoWeekly.com e-magazines and all the archives in past years.

TABLE IV

ABSTRACT OF DEMOGRAPHIC INFORMATION

Country/Territory	Visits
Hong Kong	38.78%
United States	21.10%
Canada	12.44%
China	9.05%
Australia	3.74%
United Kingdom	2.41%



104,452 visits came from 107 countries/territories

Detail Level: City | Country/Territory | Sub Continent Region | Continent Dimension: None

Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
104,452 % of Site Total: 100.00%	4.09 Site Avg: 4.09 (0.00%)	00:02:43 Site Avg: 00:02:43 (0.00%)	42.57% Site Avg: 42.56% (0.04%)	14.58% Site Avg: 14.58% (0.00%)

Detail Level: Country/Territory	Visits ↓	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
1. Hong Kong	40,509	3.83	00:01:59	57.11%	15.20%

Fig. 5. Demographic information of IAToMPW.com Web Channel

For MingPaoWeekly case, one-third of visitors come from Hong Kong, following by 20% of visitor comes from United States.

V. CONCLUSION AND FUTURE WORK

This paper focuses on the adoption of agent technology to calculate and evaluate the Web Content Engagement Time (WCET). Unlike traditional client-server based methodology which heavily depends on server performance and inaccurate measurement due to their technology limitation, agent-based methodology shares the workload of server, provide a more accurate and meaningful measurement of web analysis and gives more user-oriented information. Future works will include how to utilize the user-oriented information as feedback to improve the whole system, use the engagement time collected to further examine and analyse the user behaviour such as reading habit.

REFERENCES

[1] http://waablog.webanalyticsassociation.com/2008/09/new-standards-d.html

[2] http://www.jupitermedia.com/corporate/releases/05.03.14-newjupresearch.html

[3] Michael Wooldridge, and Nicholas R. Jennings, Intelligent Agents: Theory and Practice, Knowledge Engineering Review, Oct. 1994.

[4] Castlefranchi, C., Guarantees for Autonomy in Cognitive Agent Architectures. *In*: Intelligent Agents: Theories, Architectures and Languages, Wooldridge, M. and Jennings, N. R., *Eds*., Lecture Notes in Artificial Intelligence, Volume 890, Springer-Verlag: 56-70, 1995

[5] Genesereth, M. R. and Ketchpel, S. P., Software Agents. *In*: Communications of the ACM, 37(7): 48-53, 1994

[6] Shoham, Y. Agent-oriented programming. Artificial Intelligence, 60(1):51-92, 1993.

[7] Bates, J. The role of emotion in believable agents. Communications of the ACM, 37(7):122-125, 1994.

[8] Maes, P. Agents that reduce work and information overload. Communications of the ACM, 37(7):31-40, 1994a.

[9] Stefan J. Johansson, A Preference-Driven Approach to Designing Agent Systems, Proceedings of the 2nd Asia-Pacific Conference on IAT, 2001.

[10] George A. Bekey, Autonomous Robots From Biological Inspiration to Implementation and Control, 2005.