



E-CUSTOMER ANALYTICS

WHITE PAPER

Authors: Pushpa Ramachandran M, Ragavendra S, Kunal Turakhia

Abstract of the paper

World Wide Web has become an important channel for conducting the businesses. The Internet provides the opportunity to access the global market for the enterprises. Increasingly the Internet is becoming the first point of contact with prospects and the customers for the businesses. This leads to the necessity for understanding the behavior of the web site visitors, by analyzing and profiling their behavior. Data Warehousing and Data Mining technologies can help the e-businesses to better understand their e-customers.

The analysis of the web site visitors poses unique problems like:

- Identification of the origin of the visitor is required
- Calculation of the Dwell time for a content page
- Identification of a User Session
- Managing Web-site Structure Information

This paper addresses the specific problems and possible solutions in providing the analytical solution.

Table of Content

Motivation	01
Challenges.....	02
Identification of the Origin of the Visitor	03
Calculation of Dwell Time	04
Identification of User Session	05
Managing Web-site Structure Information.....	06
Conclusion	08
Author Bio's	09
About Wipro	10
Wipro in Business Intelligence & Data Warehousing.....	10

Motivation

Today, almost every organization has a web interface to the world. The Internet revolution has changed the way companies do their business, communicate to their customers, vendors, and partners and do their day-to-day transactions. The web sites have become the critical components of the overall business strategy of a company. Increasingly, the first point of contact between a company and its customer is at its web site. The organizations that harness their web site data and use this knowledge to create profitable, lasting customer relationships will win. This white paper proposes a solution to leverage on the web based data and addresses the challenges in implementing the solution.

Web Housing & Web Mining

The web site plays an important role in the e-Commerce era, as most of the commercial transactions are through the web. A single online transaction is of more value than the sale itself, in this era, as it provides a way to target the right customer and to understand the needs of the customer.

Gigabytes of data showing the visitor's browsing pattern in the web site, is collected every day from the web servers. This is called the click-stream or web traffic data. Web log analysis reports basic traffic information based on web server log files. The main purpose for web log analysis has traditionally been to gain a general understanding of what is happening on the site like the amount of traffic and the type of errors etc. This information is typically used for web site management purposes.

To identify the right customer and his interests, sophisticated methods are needed to store and analyze the data collected over a period of time. Data Web housing is the new technology of storing the click-stream data with other data from other sources in a data warehouse in the form of star schema with dimensions characterizing the fact measures, to facilitate the analysis of the data. This warehouse is also called the Data Web house. The data can be drilled down to get more detailed information and drilled up to look at the data at more aggregated /consolidated form. The data can be sliced and diced to get an insight into a specific portion of the data, as well.

Web Mining is the integration of web traffic with other traditional business data like sales automation systems, inventory management, accounting, customer profile database, and e-commerce databases to enable the discovery of business co-relations and trends. This leverages on the principles of data mining, which is automated detection of predictive information from large databases.

Some of the applications of web housing and web mining are:

- Targeting the right customer at right time, target marketing
- Personalizing the content of the web site, to own the experience of the user
- Customer profiling to provide a 360 view of the customer behavior
- Site level statistics, to help organize the organization's web space in a better way
- Identifying customer propensity and cross sell & up sell opportunities



By integrating with the external data such as demographics, consumer and household data and business data such as marketing, sales and product management a broader picture and more accurate solutions can be arrived at. The diagram shows a possible architecture for the web house.

The Challenges

Building a successful data warehouse itself is a challenging task and building a data mining model on the data poses lot of challenges, starting from the understanding of the business problem, data preparation to the building and deploying the mining model. Web poses specific challenges in terms of cleaning, transforming and loading the data for the purpose of analysis, as normally 90% of the click-stream data is of not much importance from analytic perspective. Following is the list of possible challenges:

Identification of the origin of the visitor is required. To get the more out of the click stream data it is required to characterize the web site visitors, based on their demographics. A web site visitor can be identified by making use of cookies, online forms etc. If these options are not there, then the customers are to be identified only by the IP address of the connection from which he is accessing the web site. The origin of the visitor is to be identified to have more insight on the visitor behavior using one of these methods.

Calculation of the Dwell time for a content page. The time spent by the visitor on a particular page provides a good measure showing the interests of the visitor. Direct ways are not available to calculate the dwell time of a visitor on a page.

Identification of a User Session. A visitor can be characterized by studying his browsing behavior in a session, which is a collection of web based transactions related by time. Computing the start and end of a session is a complex process.

Managing Web-site Structure Information: The structure of the web site is an important information. With the continuous changes in creating and maintaining electronic docu-

ments, there are multiple challenges in the ETL process for loading and maintaining the web site structure. The challenges include handling dynamic pages, handling ancillary pages, extracting page title and category and handling frequently changes in the pages served in the web site.

Let us look into each of the problems at a detailed level and more importantly how to address them.

Challenge 1: Identification of the Origin of the Visitor

Web is the most anonymous thing on the earth and the web site visitors want to be anonymous. It is a great challenge to discover the personalities of these anonymous visitors based on their behavior during the time they interact with your web site, and capturing enough information to do so without infringing into their privacy.

There are four levels in which a user can be identified, viz.

- Based on Visitor's IP Address
- A persistent identifier for that session only
- A persistent identifier that lets know the same web browser on a particular computer has returned for a repeat session
- A persistent identifier that lets know the particular human being has returned to our web site

Based on the Visitor's IP address get the country rather than the person name. It is better to know atleast the country of the visitor instead of anonymity. Knowing the country of the visitor provides with opportunities to a personalize the web site for his needs as well in gaining the browsing behavior of the person with respect to the local time of the user.

The IP addresses are allocated dynamically by the Internet Service Providers (ISPs) to their customers. The IP address is not the unique way to identify a web site visitor.

There are databases maintained for each part of the globe which gives the country, contact person of the ISP, his mail-id, phone number, fax number, IP address allocating authority and the route to the IP address etc. This helps to identify the part of the globe from which the visitor is originating.

A persistent identifier for that session only can be passed through URLs, hidden fields or session identifiers. This will help avoid the problem of proxy servers. But only current session can be recorded No way of tracking repeat visits and the browser Caching. Clicking of the back button is not recorded in the web server log. This makes it impossible to have a complete map of user's actions. A possible solution for this could be the use of No-Cache tags in the HTML content

A persistent identifier that lets know the same web browser on a particular computer has returned for a repeat session can be implemented through persistent cookies stored on the client machine. The cookie is a record placed on a user's PC by a web browser in response to a request from a web server. The cookie contents are specified by the web server and can only be read from the domain that is specified the cookie. This provides a way to identify the machine from which the user is accessing

the net and not the user. The problems with cookies are that the user might have disabled the cookies. Even if the cookies are enabled the user may delete it at any point of time.

A persistent identifier that lets know the particular human being has returned to the web site is normally implemented via access through user/password. Online forms like registration or preferences for customization are an excellent source to link customers to clicks generated by them. By far, it is the most effective method of gathering visitor information. Online forms also have problems. It is believed that when asked for their name on an Internet form, men will enter a pseudonym 50 percent of the time, and women will use a pseudonym 80 percent of the time. It is not preferable to ask the user to fill in the form while he is visiting the site for the first time, as it can be repulsive.

Challenge 2: Calculation of Dwell Time

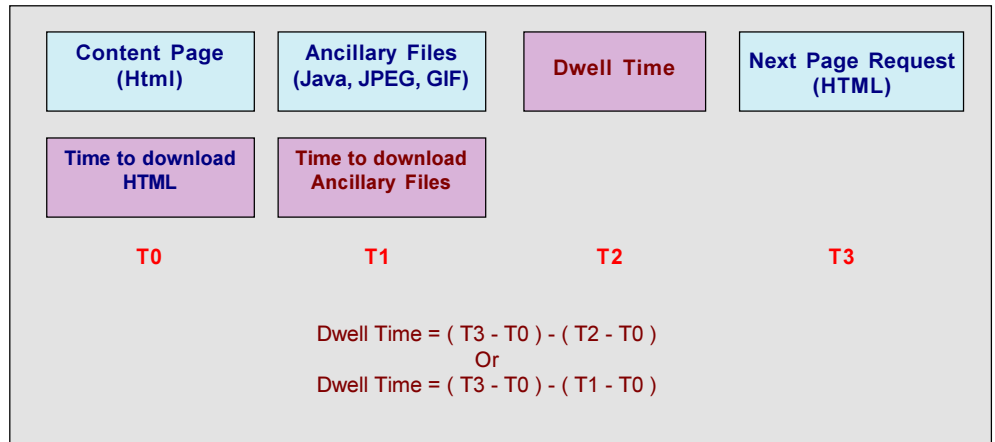
Dwell time is the time spent by the visitor on a content page. It is an important measure of the relevance of the content for the user and effectiveness of the page in attracting the visitor.

The dwell time can be calculated by finding the difference between the 2 content page requests and subtracting the time required to load the content page and the ancillary files from the value. But the time required to load streaming media files like real audio and mpeg may not be considered for the dwell time computation. In this case, the dwell time is to be computed using the beginning of the streaming media download regardless of whether the rest of the content is fully downloaded.

$$\text{Dwell time of content page 1} = \text{Time of request of content page 2} - \text{Time of request of content page 1} - (\text{Time required to load the content page} + \text{Time required to load the ancillary files})$$

The pros and cons for this approach are:

- Using this method the time spent on viewing the fully loaded page can be calculated.
- When it takes time to load the ancillary files, the user may go through the content of the page, as the text portion of the content page is visible. The user may decide to go to next page even before all the ancillary files are loaded. In this case the formula used for calculating the dwell time will underestimate the actual dwell time.



As an alternate the dwell time can be calculated by finding the difference between the 2 content page requests and subtracting the time required to load the content page alone from the value. But the time required to load all kinds of ancillary files are not considered for the dwell time computation.

$$\text{Dwell time of content page 1} = \text{Time of request of content page 2} - \text{Time of request of content page 1} - \text{Time required to load the content page}$$

The pros and cons for this approach are:

- The time spent by the user on text portion of the content page is taken into account.
- The dwell time computation may not be too accurate as the load time for the ancillary files are taken into account.

Based on the importance of the ancillary files in the web page, one of these two approaches can be chosen. If the pages are containing pictorial representations of the information, then it may be appropriate to adopt the first approach otherwise the second approach will be more appropriate.

Challenge 3: Identification of User Session

The start and end of a user session is to be identified in order to analyze the user behavior in a session as well as for measuring the effectiveness of the design of the web site in keeping the visitor for more time in the site. This also helps in identifying the various 'entry pages', the page through which a visitor enters the web site and effectively design these pages by providing links to other pages and putting appropriate ad-banners in the pages depending upon the context. Any page in a web site can be the entry page for a visitor as key word search in search engines can lead the visitor to any page in the web site.

The identification of a session also helps in identifying the most popular exit pages, which could be the session killers. Identifying the session killers and effectively redesigning them may keep the users in the web site for more time. But there is no direct way to identify the start and the end of a user session.

Entry pages. The very first request from an IP address at any point of time is considered as the start of a user session. A page request with out any referrer or an external referrer can be considered as the start of a session. Referrer is a URL from which the current page is reached. If there is no referrer, then it means that the user has directly keyed in the URL of the current page and reached the web page. In this case the page is definitely the entry page of the session for that particular user. If the referrer is an external URL, then the user might have followed a link from that page to this page or searched for a keyword and the search engine led him to the page or might have clicked on an ad-banner and reached the page. In this case also the page is definitely the entry page of the session for the visitor.

Exit pages are pages that mark the end of a session. But there is nothing like as decisive as hanging up a phone. So, it becomes difficult to identify the end of a session and the exit page. One possible solution for this problem is to use the inactivity timeout, as a criterion to identify the end of a session. For example, if there is no request for more than 5 minutes then that can be considered as the end of a session. The subsequent events are assumed to be from a follow-on session. The time limit is to be decided appropriately, as this puts an upper bound on the dwell time that a visitor spends on a page.

The local time for the user of the session is also important attribute for characterizing the user sessions. The web servers log only the local time or the server time. The local time of the user can be calculated by identifying the country to which the visitor belongs. The Greenwich Mean Time can be calculated by knowing the difference in time between the server time and the GMT. As most of the countries have a standard time difference with GMT, the local time of the user can be calculated once the country, from which a user is accessing the web site, is known. While calculating the local time, the day light saving hours should also be taken care of. There is one more challenge in determining the local time, as some countries can have more than one time zone. Some strategy is to be arrived at for handling this situation.

Challenge 4: Managing Web-site Structure Information

Web sites may serve static or dynamic pages or a combination of both and each page served may contain or have links different type of files like documents, images, multimedia, embedded scripts, etc. Pages can be static html documents or can just consist of a template and an Application Server can serve the content for the different components of the template. The type of files served may change frequently. Many new pages can be added on a daily or weekly basis and the old pages may be superceded. According to its purpose, the files may have a classification like Company information, Product catalogue, Technical support, Ordering page, etc. Content pages may have page titles, which will be required for analysis and it should be extracted and loaded to page dimension.

Dynamic pages. Pages can be generated and served dynamically based on the parameters given by the visitor in a previous page. A dynamic page can consist of a template with different components and the content for each component can be generated dynamically based on a given set of parameters. The page used will be the same but the content served will be different at different instances of time. Storing all the instances of the dynamic page will drastically increase the size of the page dimension.

The solution for this problem will be to identify the function that characterizes the page and use that function to describe that page. The dynamic pages generated from the same template can then be grouped by similar function and type.

Frequent changing of the pages served. The pages served at a web site can change frequently. Many new pages can be added on a daily or weekly basis and the old pages may be superceded. Also pages can just consist of a template and an Application Server can serve the content for the different components of the template. This is a normal situation in web sites, which give daily news, stock updates, etc. This may lead to frequent update of the page dimension.

The data on static pages can be stored page dimension with one record for each static page. For dynamic pages, data on the templates can be stored, as each template will be used for a specific type of page. The page dimension has to be updated for any new static pages or templates added to the site.

Non-Content (Ancillary) pages. The Web site may serve a huge number of graphics, images, sound files and other ancillary content. The number of such files may be high and storing the server hits for these files in the click stream may be expensive. The access log entries for these files cannot be discarded when as they will be required in calculating dwell time, which is the likely time that a specific page was actively displayed on the visitor's browser. Ancillary pages will have more importance in cases like a product specification page that shows the product features in one or more picture files. Hence, the page dimension should provide a way to associate each such ancillary file with one or more content pages.

For small sites with a few thousand files, including all the content and ancillary files, the data about both the content files and the ancillary files can be stored in the page dimension. For large sites using several thousands of files, the information on ancillary files can be stored separately from page dimension and each ancillary file can have a page key to link to the content page that contain it.

Extracting Page Title. The new pages from the access log entries will be loaded to the Page dimension by looking up at the records already available in the page dimension. For these records, the page title cannot be retrieved at run time while loading from the access log.

A process has to be defined which will lookup the page titles for the new pages that are not already available in the page dimension. This process can save the page titles in the staging area. This process should to be automatically called by the ETL process for loading page dimension and should be synchronized with it. Alternately, while loading the page dimension, the ETL process can call a function for each new content page to return its title.

Extracting Page Category. The pages served by the web site can be categorized according to the business to reflect different usage of the page. Examples of such classification are Home page, Company information, Product catalogue, Technical support, Ordering page, etc. Such classification will be helpful for analysis. The issue with this is that, the classification is highly subjective and can be done only by the designers of the Web site. Assigning the category information for thousands of pages in the web site will be a tedious task.

The page category is highly subjective to the business and can vary widely for different sites. Page category will be a decision taken by the site designers and the mechanism for assigning categories to pages should be designed by them. If page category is stored for all content pages in some form, the same can be used. A process for extracting page category for each content page has to be defined, which will be similar to the process for extracting page titles dynamically.

Conclusion

Web houses - data warehouses that integrate and summarize the web-based data for analysis - have tremendous business value. This article lists the challenges that are very different from the other types of Data Warehousing projects:

- Identifying the anonymous users uniquely
- Computing the time spent by the visitors in the web site at the page level
- Identifying the user sessions by calculating the pages visited by the user in a sequence and the total time spent by the user in the web site
- Managing Web-site structure information

Web housing should aim at providing a flexible architecture that adapt gracefully to new end user queries, new dimensions, new attributes within those dimensions and new facts. This will help addressing the customer needs and continuously improving the web-based services.

References

- www.webminer.com
- www.kdnuggets.com
- www.dmreview.com
- www.w3c.org
- www.wdvl.com
- www.arin.net
- www.ripe.net
- www.apnic.net
- Data Web House Toolkit by Ralph Kimball et al.

About the Author's

Pushpa Ramachandran M, is a Consultant with the Business Intelligence and Data Warehousing group in Wipro Technologies. He has worked on Data Warehousing and Data Mining projects for companies in Finance, Telecom, Customer service and Airline industry on issues such as managing customers, targeting promotions, detecting fraud, and pricing. He holds a Masters degree in Computer Applications.

Kunal Turakhia, is a Consultant with the Business Intelligence and Data Warehousing group in Wipro Technologies. He has worked on Data Warehousing and Data Mining projects for companies in Insurance and Utilities industry on issues such as detecting fraud, optimizing commercial decision making, tracking of business performance and meeting regulatory requirements. He holds a MBA in Information Systems.

Ragavendran Sripad, is a Consultant with the Business Intelligence and Data Warehousing group in Wipro Technologies. He has worked on Data Warehousing, Process Modeling for Data Warehousing and Web Mining. Has also worked the implementation of Oracle Financials. Currently involved in Data Warehousing projects in the area of Asset Management.



About Wipro

Wipro (NYSE: WIT) is the first SEI CMM Level 5 certified IT Services Company operating in the global market. Wipro provides software solutions and services to global corporate enterprises and Research and Development services to Telecom and Electronic product companies. In the Indian market, Wipro is a leader in providing IT solutions and services for the corporate segment offering system integration, network integration and IT services.

Wipro in Business Intelligence & Data Warehousing

Wipro provides end to end Data Warehousing and Business Intelligence services to Global Corporate Enterprises. Wipro has implemented Business Intelligence and Data Warehousing solutions for over 30 Fortune 1000 customers across various vertical industries like Finance, Insurance, Utilities, Telecom, Retail, Logistics, Manufacturing and Healthcare.

Wipro has evolved its "Insta Intelligence" project management and delivery methodology built around leading edge technologies in the areas of Data Acquisition, Data Modeling, Data Management, OLAP, Data Mining, and Meta-Data Management to deliver innovative, surefire solutions to its customers. It has entered into Business and Technology alliances with some of the leading vendors like IBM, Informatica, Cognos, Microstrategy, Brio, SAS to offer customized solutions to its customers.

© Copyright 2001. Wipro Technologies. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without express written permission from Wipro Technologies. Specifications subject to change without notice. All other trademarks mentioned herein are the property of their respective owners. Specifications subject to change without notice.

America

1995 El Camino Real, Suite 200
Santa Clara, CA 95050, USA
Phone: +1 (408) 2496345
Fax: +1 (408) 6157174/6157178

Europe

137, Euston Road
London NW12AA, UK
Phone: + (44) 020 73870606
Fax: + (44) 020 73870605

Japan

Saint Paul Bldg, 5-14-11
Higashi-Oi, Shinagawa-Ku,
Tokyo 140-0011, Japan
Phone: + (81) 354627921
Fax: + (81) 354627922

India-Worldwide HD

Doddakannelli, Sarjapur Road
Bangalore-560 035, India
Phone: + (91) 808440011 -15
Fax: + (91) 808440254

www.wipro.com

eMail: info@wipro.com

To access more white-papers, please visit <http://www.wipro.com/shortcuts/downloads.htm>