Ten Supplementary Analyses to Improve E-commerce Web Sites

Ron Kohavi
Blue Martini Software
2600 Campus Dr.
San Mateo, CA 94025
+1-650-356-4000
ronnyk@cs.stanford.edu

Rajesh Parekh
Blue Martini Software
2600 Campus Dr.
San Mateo, CA 94025
+1-650-356-4000
rparekh@bluemartini.com

ABSTRACT

Typical web analytic packages provide basic key performance indicators and standard reports to help assess traffic patterns on the website, evaluate site performance, and identify potential problems such as bad links resulting in page not found errors. Based on our experience in mining data for multiple retail e-commerce sites, we offer several recommendations for supplementary analyses that we have found to be very useful in practice. These include analysis of human vs. bot traffic, univariate data, session timeouts, form errors, micro-conversions, search, real estate usage, product affinities (associations), migrators, and geographical data. Several of these advanced analyses are based on the construction of a customer signature, which in turn benefits from additional overlays, such as third-party demographic attributes. We describe the construction of such a signature and the challenges faced by businesses attempting to construct it.

Keywords

Web Analytics, Web Mining, E-commerce, Web Metrics

1. INTRODUCTION

There are now many commercial and freeware software packages that provide basic statistics about web sites, including number of page views, hits, traffic patterns by day-of-week or hour-of-day, etc. These tools help ensure the correct operation of web sites (e.g., they may identify page not found errors) and can aid in identifying basic trends, such as traffic growth over time, or patterns such as differences between weekday and weekend traffic. There is no doubt that this basic information is necessary for day-to-day operation of the site and for tactical decisions (e.g., deciding when to take the site down for maintenance based on hour-of-day loads). With growing pressure to make e-commerce sites more profitable, however, additional analyses are usually requested. These analyses are usually deeper, involving discovery of factors, and more strategic to the business.

In this paper we describe ten analyses that we call "supplementary analyses." They are supplementary because they are *not* the first analyses one should run. Every site should start with basic statistics about their operations as provided by various commonly available packages. It is important to make sure that the site is live and stable, that response time is good, and that there are no broken links before performing deeper analyses. Organizations need to know how to crawl efficiently before attempting to walk and run.

Before describing our analyses, we would like to share with the readers our prior experience, both to convey the practical experiences we have had and also to highlight the limited settings and domains, which may bias our experience and may not necessarily generalize. At Blue Martini Software, we had the opportunity to develop a data mining system for business users and data analysts from the ground up, including data collection, creation of a data warehouse, transformations, and associated business intelligence systems that include reporting, visualization, and data mining. The architecture collects clickstreams and events at the application server level thereby avoiding the need for complex efforts of sessionization from web log data [3][13]. The architecture also collects higher-level information, such as search keywords, cart events, registration, and checkout. More details about the architecture are available in Ansari et al. [2]. We have worked with multiple clients to mine their data, answer business questions, and make recommendations. Case studies [4][5] were recently written detailing some of our findings. Most of our clients have retail business to consumer e-commerce sites, limiting our experience to that segment (e.g., our experience with Business-to-Business data is extremely limited).

For the purpose of this paper, we are assuming that clickstream data have been sessionized and that transactions (e.g., purchases) have a key to identify the session in which they occurred. In systems where such information is not collected at the application server layer, sessionization heuristics can be applied to web log data [3][13].

When faced with a choice of deciding which analyses to recommend as the top ten supplementary analyses we chose to prioritize based on the following criteria:

1. General. Is the analysis useful across multiple clients? While each client may have some special analyses that they deem critical, not every analysis generalizes well.
2. Actionable. Does the analysis lead to actionable patterns with high return-on-investment (ROI)? We found many interesting patterns in data, but interesting patterns are not always practically useful or actionable.
3. Non-overlapping. Are the analyses providing wide coverage? We chose to present analyses that cover different areas. While we can share a variety of analyses on a single topic (for example, search), we wanted to cover multiple areas, ranging from usability to real-estate allocation to conversion rates.

The ten supplementary analyses are divided into four sections. Section 2 begins with the foundation and data audit analyses, followed by operational analyses in Section 3. Section 4 describes the tactical analyses, and Section 5 deals with strategic analyses. We conclude with a summary in Section 6.

2. FOUNDATION AND DATA AUDIT ANALYSES

To generate strategic insights and effective models in customer-centric analyses, it is imperative to generate rich customer signatures covering all aspects of customers' interaction with the company. Information from sources such as registration data, browsing patterns, purchase history, campaign response, and third party demographic and socio-economic data providers can be used to build the customer signature. In our work with client data we built customer signatures with hundreds of attributes. They include aggregated data such as number of visits, time spent on the site, frequency of visits, days since last visit (recency), total spending, average spending, spending per product category, etc. Mining algorithms can automatically determine the important attributes from the signature that are useful in predicting interesting customer characteristics.

Rich data, however, are not enough—we must ensure that the data are of high quality. To establish a solid foundation for analyses, we recommend performing data audits. Bot analysis and univariate analysis of data are the two data audit analyses that must be performed as a precursor to any deep web site analysis.

2.1 Bot Analysis

Web robots, spiders, crawlers, and aggregators, which we collectively call *bots* [8], are automated programs that create traffic to websites. Bots include search engines, such as Google, web monitoring software, such as Keynote and Gomez, and shopping comparison agents, such as mySimon. Because such bots crawl sites and may bring in additional human traffic through referrals, it is not a good idea for websites to block them from accessing the site. In addition to these "good bots," there are e-mail harvesters, which try to look for e-mails that are sold as e-mail lists, offline browsers (e.g., Internet Explorer has such an option), and many experimental bots by students and companies trying out new ideas.

In our experience, bots account for 5 to 40% of sessions. Due to the volume and type of traffic that they generate, bots can dramatically skew site statistics. Accurately identifying bots and eliminating them before performing any type of analysis on the website is therefore critical.

If we define the term **browser family** as browsers with the same user agent or same IP (these are part of the HTTP header for every request), we typically identify about 150 different browser families that are bots using fairly conservative heuristics described below. Sites like www.robotstxt.org list over 250 different bots.

Many commercial web analytic packages include basic bot detection through a list of known bots, identified by their user agent or IP. However, such lists must be updated regularly to keep track of new evolving and mutating bots. In addition to using a list of known bots, we have added heuristics that dynamically identify new bots. These heuristics include:

1. A browser family that never logs in (an extremely powerful heuristic for sites that have registration).
2. Multiple visits by the same browser family to a zero-width link (trapdoor) that is not seen by humans using browsers.
3. Large number of very long or one-click sessions.
4. Highly regular traffic patterns.
5. No referrers.
6. No gzip support (which most bot authors never consider important to implement).

These heuristics have served us well in identifying most bots, but they are not perfect. We therefore always start with bot detection as our first analysis, looking at browser traffic by browser family (making sure new families are correctly classified), and reviewing very long sessions and spikes in traffic to see if bots cause them. Tan and Kumar [11] discuss additional heuristics and building classifiers for bot detection.

In the remaining analyses we assume that bots have been identified and filtered out.

2.2 Univariate Analysis of Data

The process of performing a simple univariate analysis of data (including the rich customer signatures) cannot be emphasized enough. This simple analysis of looking at the distribution of the different values for an attribute, the number of null values, the minimum, maximum, and average of the values (when appropriate), and outliers provides a lot of insightful information about the data. Often a visual depiction of these basic univariate metrics as shown in Figure 1 is sufficient to obtain a good overview.

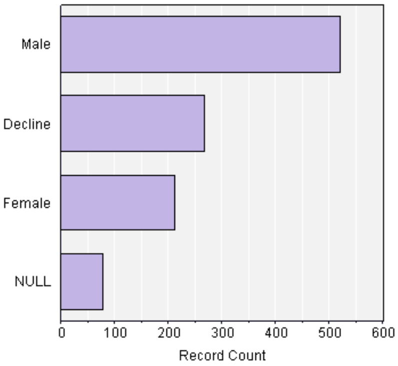

Figure 1: Visual Depiction of Attribute Values

3. OPERATIONAL ANALYSES

In this section we describe three analyses that impact the website operations.

3.1 Session Timeout Analysis

Enhancing the user browsing experience is an important goal for website developers. One hindrance to a smooth browsing experience is the occurrence of a session timeout. A user session is determined by the application logic to have timed out (ended) after a certain predefined period of inactivity. Depending on the timeout implementation, further activity after session timeout can either silently start a new session or first display a timeout page to inform the user of the session timeout and then start a new session. In either case, users who have active shopping carts at the time of a session timeout might lose their state and possibly the shopping cart when a new session starts. Setting the session timeout threshold too high would mean that fewer users would experience timeout thereby improving the user experience. At the same time, a larger number of sessions would have to be kept active (in memory) at the website thereby resulting in a higher load on the website system resources. Setting an appropriate session timeout threshold involves a trade-off between website memory utilization (which may impact performance) and user experience.

In prior work [6], it was determined that the mean time between two browser events across all users was 9.3 minutes and a session timeout threshold of 25.5 minutes (1½ standard deviations from the mean) was recommended. In one of our studies, we found that users were experiencing timeouts as a result of a low timeout threshold. We conducted an experiment to study the impact of session timeout on lost shopping carts for two large clients. For the purpose of this experiment, we arranged successive web page requests from each distinct browser cookie in chronological order. These ordered requests were sessionized manually as follows.

- If the duration between any two successive web page requests from the same cookie was over 3 hours then the session was considered to have terminated with the first of the two successive requests and a new session was considered to start with the second request.
- If any web page request corresponded to ‘logout’ or ‘checkout confirmation’ then the following requests for the same cookie were considered to be part of a new session.

Following this manual sessionization, we determined the maximum duration between any two web page requests of a session. If the session timeout threshold for the website is smaller than the maximum time duration between any two successive web page requests of a session then that session would timeout out prematurely.

Figure 2 and Figure 3 show the impact of different session timeout thresholds set at 10-minute intervals for the two large clients. If the session timeout threshold were set to 25 minutes then for client A, 7% of all sessions would experience timeout and 8.25% of sessions with active shopping carts would lose their carts as a result. However, for client B, the numbers are 3.5% and 5% respectively.

Thus, the recommended session timeout threshold would be different for the two clients. Further, the smooth curves in both the charts do not suggest any specific threshold beyond which the impact of session timeout would be minimal for either of the two clients. It is therefore imperative for website administrators to run a similar analysis on their data to determine a suitable session timeout threshold. As a general rule, we recommend that the session timeout for e-commerce sites be set to no less than 60 minutes. Note that is more than double the sessionization time recommended in [6].

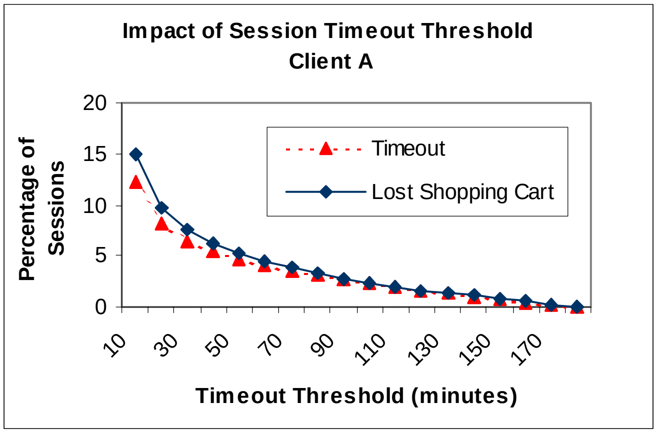

Figure 2: Session Timeout Threshold Analysis (Client A)

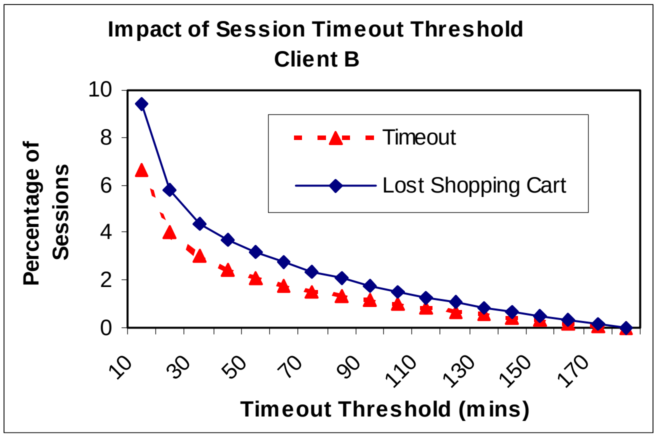

Figure 3: Session Timeout Threshold Analysis (Client B)

3.2 Form Error Analysis

This analysis can only be performed if there is proper architectural support for collecting form errors. We built the architectural support into our system based on client demand and it has been very useful. The idea is to log any type of error where the user is not allowed to continue. Examples include:

- Login failure.
- Mandatory form questions that are not filled in.
- Incorrect format (e.g., email address that does not conform to the specification or a US phone number without 10 digits).

An architecture to perform form validation using some basic regular expression matching or specific logic to validate individual fields could be implemented on any website.

One example of the usefulness of this event happened with one of our customers. Two weeks after go-live with our system, we looked at form failures and found thousands of failures on the home page every day! There was only one form field on the home page: a place to enter your email address to join their mailing list. Thousands of visitors were typing search keywords into the box and because they failed to validate as emails, the architecture logged them. The fix was easy and the customer added a clear search box on the home page and defaulted this email box to the word “email.”

Another example comes from our own company’s site (www.bluemartini.com). When we analyzed our form errors, we noted a large number of failures in the login. Looking at the values typed, it was clear that many users attempted to login with the username and password from our developer site (developer.bluemartini.com). In the short term, we made it clear that the developer site is independent. Longer term, synchronizing the accounts is desired. The large number of login errors provided us with the data to justify developing a synchronized account system.

As with many proposals, it is sometimes clear that proposal A (e.g., synchronized accounts) is superior to proposal B (e.g., having independent logins on the two sites), but the cost (e.g., in terms of a poor user experience) of implementing an inferior proposal is not always known. Companies are under constant pressure to do more with fewer resources and will usually pick to implement proposals that are “good enough.” Doing the analysis allowed us to better quantify the user experience and how it would improve with a better design. That can then be traded-off against the cost of a more expensive proposal.

3.3 Micro-conversions Analysis

A **conversion** is the completion of a task or process. Classical examples of conversions at e-commerce sites are checkout (i.e., purchase) and registration. Other tasks include filling a survey or finding a nearby store. Conversion tasks are typically multi-step, i.e., they require a user to fill a few forms entries, such as their name, address, and credit card number.

Most organizations know their conversion rates for critical processes, such as purchases or registration. However, a deeper analysis may reveal where in the conversion process visitors abandon. The intermediate steps of a conversion are called *micro-conversions* [9]. For example, while purchase conversion may be 2%, it may be that half the users abandon when seeing the page with the shipping charge, and an experiment might be justified to see if the conversion rate improves dramatically if shipping costs are reduced for some segments.

Figure 4 shows micro-conversions in the checkout process for Debenhams, UK [4]:

- Visitor adds to cart: 6.0%
- Visitor who added to cart initiates checkout: 45%
- Visitor who initiates checkout purchases: 83%

Looking at these numbers, 55% abandon their shopping carts before initiating checkout and of the remaining 45%, 17% abandon during checkout. The total cart abandonment of 62% is not abnormal, but leaves significant room for improvements. For example, further investigation revealed that 9.5% of users lost their carts due to session timeouts that were set very low.

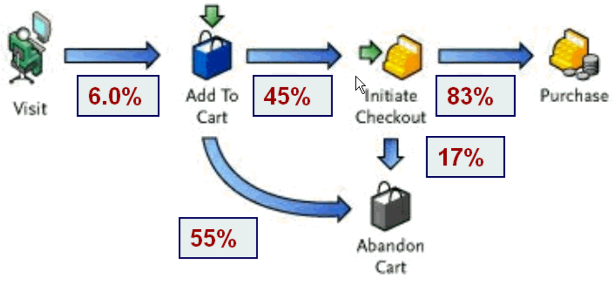

Figure 4: Micro Conversions at Debenhams, UK

In another example, a client was evaluating micro-conversion rates for a process different than checkout. Micro-conversion analysis revealed a subprocess that had significant abandonment. By making slight changes to the subprocess, the overall conversion rate increased by over 50%, which had a significant impact on revenues.

Micro-conversion rate analyses helps to quickly identify areas of the site with potential problems so that web site administrators can take appropriate steps to quickly fix the problems.

4. TACTICAL ANALYSES

In this section we describe three tactical analyses that can have a positive impact on the e-commerce website.

4.1 Search Analysis

Search is a critical activity at e-commerce web sites. Across multiple sites we have seen that a search in a session increases the conversion rate for the visitor by 50% to 150%. Knowing that visitors who search are very important, we have analyzed search from many angles. Visit to purchase conversion and average revenues per visit are important key-performance indicators for comparing visitors who search with those who do not. Beyond these simple metrics, it is important to look at:

- Failed searches: what searches return zero results (requires architectural support for collecting information about results returned by search). Failed searches help in two important ways:
  - They help to identify the need for synonyms in the search thesaurus. For example, at Publix Direct, a supermarket chain, one of the top selling assortments was “bathroom tissue.” The top failed search was “toilet paper.” Adding synonyms solved this problem. Another good tactic is to add common misspellings of popular search key words such as ‘jewelry’ as synonyms.
  - They help merchandisers identify trends and products that visitors expect them to carry. For example, at one site visitors were regularly searching for ‘gift certificates’ that the site did not sell.
- Too many results: it is easy to return too many results and overwhelm visitors. For example, in an attempt to avoid failed searches, many sites implement a default disjunctive query (“or”) in a multi-word query, i.e., find either of the words. For example, when visitors search for “digital cameras” the search returns all digital products and all cameras. Sites should default to disjunctive queries only when the conjunction is not found. At a minimum, results containing all search keywords should rank higher.
- Revenues associated with specific keywords: certain searches may identify high-spending segments. These can help determine words for ad placement at search engines, such as Google AdWords.
- Correlated attributes: customer and behavioral attributes that correlate with visitors searching (or not searching) may help identify trends or problems. For example, we found that low screen resolution in a group of customers correlated with lack of search activity. It turns out that the search box on the upper right was not visible at the low resolution and visitors had to scroll horizontally to see it in their browser.

4.2 Real-estate Usage Analysis

Web analytics packages allow users to analyze the effectiveness of individual web pages or of specific hyper-links on each web page. This analysis typically includes a study of user browsing patterns to determine what percentage of user sessions view web pages, follow certain paths, or abandon. This analysis should be taken to the next level by tying in sales information. Table 1 shows the percentage of all sessions that clicked on the specific links immediately after getting to the homepage and the revenue per session associated with those links. The revenue is normalized to that associated with the ‘Kids’ link which is designated as X in order to avoid disclosing actual sales amounts.

Table 1: Click through Rates and Sales for Specific Links

Link Name	Click through, Normalized Sales
Womens	14%, 1.4X
Mens	3%, 2.3X
Kids	2%, X
Wedding	13%, 4.2X
Flowers	0.6%,5.0X
Catalogue	2%, 10.2X

Consider the ‘Catalogue’ link. Clicking on this link takes users to the catalogue page where they can type in the product code supplied in the paper catalogue to purchase the item(s) they are interested in. Although a relatively smaller percentage of the sessions clicked through on the catalogue link the revenue associated with those sessions is the highest among all the other links on the page. The original design of the homepage had all of these links clustered together (in no specific order) with the same amount of real estate assigned to each of the links. Web site designers can take the information provided by the click through rate and attributed sales metrics above into account to suitably redesign the website, giving prominence to the links that have high click through rates and/or higher attributed sales while de-emphasizing links that have much lower numbers for these metrics.

4.3 Market Basket Analysis and an Automatic Product Recommender

Cross-sell and up-sell recommendations are an effective way to increase market basket size and thereby boost revenue. Retailers employ expert merchandisers who make cross-sell and up-sell recommendations based on their experience and domain expertise. Retailers often deal with extensive merchandise hierarchies comprising of hundreds of thousands of individual SKUs. It is practically impossible for marketers to make effective recommendations while analyzing such extensive volumes of data. Automated product recommendation systems based on associations rules mining [1][15] provide an efficient method for identifying statistically significant rules for making cross-sell recommendations. Automated product recommenders provide significant advantages over the manual approach.

1. They automatically generate statistically significant rules with little or no manual intervention.
2. They make multi-way recommendations based on two or more products purchased by the customer
3. They can be customized to make recommendations on a per basket or per customer level considering all products ever purchased by the customer. Further, they can make recommendations not just on individual products or SKUs but also on higher abstractions in the merchandise hierarchy such as product families or product categories.
4. They are efficient and can be re-run on a regular basis to capture new associations among product purchases.

In our experience with mining retail e-commerce data we have found that the associations rules algorithm can discover several interesting product associations that were missed by the retailers. For example, in the case of Debenhams, it was determined that customers who purchase ‘Fully Reversible Mats’ also purchase ‘Egyptian Cotton Towels’. This rule had a 41% confidence and a lift of 456! The recommendation provided on the website when users purchased ‘Fully Reversible Mats’ was for ‘Jasper Towels’. The confidence of this human-defined cross-sell was just 1.4% indicating that using the automated product recommender can provide significant cross-sell opportunities.

On the other hand, expert marketers are able to bring their experience and domain expertise to bear on the task of making appropriate recommendations. For example, marketers might decide to promote new products or brands by recommending them with some of the top selling products instead of relying solely on products that associate with the top selling products. In practice, retailers might prefer to combine the efficiency of the automatic product recommender with the experience and expertise of the marketers to formulate an appropriate cross-sell and up-sell recommendation strategy.

5. STRATEGIC ANALYSES

In this section we discuss two strategic analyses that help to identify interesting segments of the customer base and enable the marketers to design targeted campaigns to woo these customer segments.

5.1 Analysis of Migratory Customers

A common question asked by business users is “who are the heavy spenders and what characterizes them?” We have used this analysis for a long time, but always found that significant manual effort was required for interesting insights. The reason was that too many attributes in the customer signature were **leaks**, i.e., attributes that obviously correlated to heavy spending because either they were the immediate trigger (e.g., a purchase of an expensive product) or because they happened *because* of the expensive purchase (e.g., paying a lot of tax). We often ended up spending large amounts of time removing these leaks and wondering whether some attributes are really leaks. For example, a purchase of a single diamond is typically enough to move the customer into the heavy spender class, but is any purchase in the jewelry department a leak? After removing the "total purchase in jewelry" attribute, our rules engine discovered rules of the form "if they purchased zero in clothing, zero in shoes, zero in bags then they are likely a heavy spender." Well, the reason is that since they appear in the training set, they must have purchased and the only remaining department was jewelry!

Instead of a "heavy spender" analysis, a much better analysis is the "migrator" analysis, which does not suffer from purchase leaks. A **migrator** is defined as a customer who makes small purchases over one time period (say the first year) but migrates to a heavy spender over the next time period (say year two). To characterize migrators we identify their characteristics during the earlier year and use these attributes to predict their behavior the following year. Because the customer signature only uses purchases from the first period, it is impossible to create leaks that seem obvious. If the fact that a customer purchases a diamond in the first year is predictive that they will be a heavy spender in year two, then this is interesting (and not obvious).

The migrator study allowed us to identify interesting characteristics at Mountain Equipment Co-Op (MEC), Canada's leading supplier of outdoor gear and clothing [5]. For example, we were able to identify specific product line purchases that triggered a migration (t-shirts and accessories) and specific product lines (e.g., child carriers), which indicated reverse migrations, or attrition.

5.2 Geographical Analysis

Geographical analysis based on collected customer data can be extremely helpful in understanding customer preferences and in making strategic decisions such as advertising and shipping promotions. Analysis of data from Debenhams, UK, showed that although a majority of the orders were placed from within the UK, customers from the US, South Africa, and Australia accounted for a non-significant number of orders[4]. The retailer ships only in the UK. The orders placed from outside the UK were gifts for friends and relatives residing in the UK. This simple analysis alerted the client to the potential market existing outside the UK so they could formulate suitable advertising strategies specifically targeted to customers in the US, South Africa, and Australia.

Overlaying customer address information such as city or zip code with latitude and longitude data that is commercially available opens up a whole new realm of distance based analysis. Retailers who have bricks-and-mortar presence in addition to an e-commerce web site can determine the customers' distance to the nearest physical store and then analyze their purchasing patterns both online and in the retail stores. Figure 5 shows the relationship between the customers' distance to the nearest physical store and the online revenue for MEC [5]. People who reside farther away from the physical store tend to spend more online both in terms of average and total revenues. Further, a map of the location of online customers might help find regions with a large number of online customers with no bricks-and-mortar store close-by. Retailers can use this information to select possible locations for their future bricks-and-mortar stores.

Figure 6 shows the relationship between the average revenue per visit to a bricks-and-mortar store and the distance from home of the bricks-and-mortar store shopped at, for a major retailer in the US. It is interesting to note that people who travel far to shop at the store spend more on average per visit than those who shop locally. This information can be very useful for the retailer in charting out a suitable advertising strategy. Specific ad campaigns can be designed to attract customers to stores located in cities that are popular tourist destinations or important business centers.

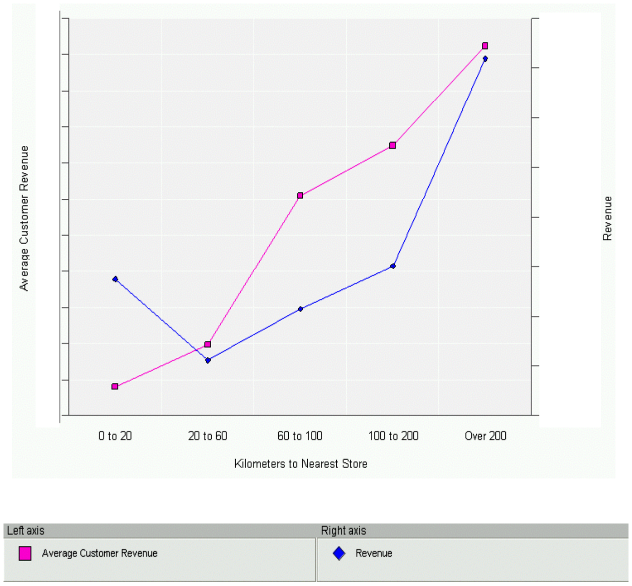

Figure 5: Total and Average Revenue by Distance to Nearest Physical Store

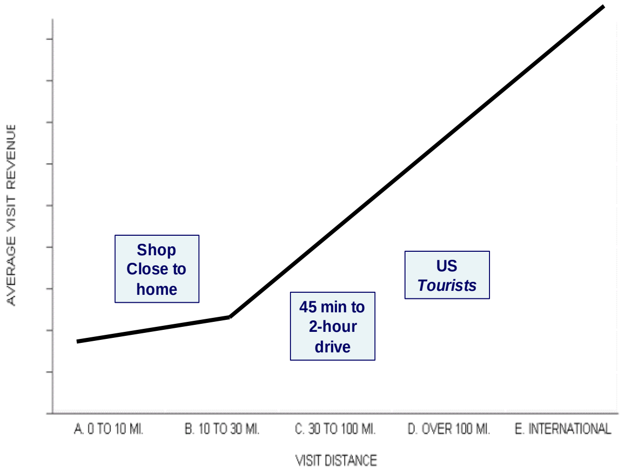

Figure 6: Average Visit Revenue by Distance of Store from Home

6. SUMMARY

We presented ten supplementary analyses that e-commerce web sites should review after running the common reports provided by most web analytics tools. These recommended analyses are based on our experience mining data from multiple retail e-commerce sites over the last three years. The analyses cover a wide range of topics from bot analysis to web site real estate usage analysis, and from analysis of migratory customers to geographical analysis. We have found the insights from these analyses to be very interesting to the business users and actionable in practice. While some analyses require architectural support to collect the right data, the effort required is moderate in most cases and should result in significant Return-On-Investment (ROI). Some analysis, such as customer migration, can benefit directly from a wide customer signature, and many analyses may lead to insight that can be investigated through the use of customer signature. It is therefore critical to develop a rich customer signature and reuse it in analyses. We are still developing domain knowledge and techniques to enable us to build rich customer signatures, and we believe that this is an area that deserves the attention of researchers and practitioners.

ACKNOWLEDGMENTS

We wish to thank Blue Martini's data mining team for many interesting discussions on these analyses. We wish to thank MEC and Debenhams for allowing us to share some insights in case studies.

7. REFERENCES

[1] Agrawal Rakesh and Srikant Ramakrishnan. Fast Algorithms for Mining Association Rules, Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), 1994.

[2] Ansari Suhail, Kohavi Ron, Mason Llew, and Zheng Zijian, Integrating E-Commerce and Data Mining: Architecture and Challenges, ICDM 2001 http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html

[3] Berendt Bettina, Mobasher Bamshad, Spiliopoulou Myra, and Wiltshire Jim. Measuring the accuracy of sessionizers for web usage analysis. In Workshop on Web Mining at the First SIAM International Conference on Data Mining, pages 7-14, April 2001.

[4] Business Intelligence at Debenhams: Case Study, http://www.bluemartini.com/bi

[5] Business Intelligence at Work: MEC Case Study http://www.bluemartini.com/bi

[6] Catledge Lara and Pitkow James, Characterizing Browsing Strategies in the World Wide Web, Computer Networks and ISDN Systems, 27(6), 1065-1073, 1995.

[7] Cohen William, Learning Trees and Rules with Set-Valued Features, Proceedings of the AAAI/IAAI Conference, vol. 1, pages 709-716, 1996.

[8] Heaton Jeff, Programming Spiders, Bots, and Aggregators in Java, Sybex Book, 2002.

[9] Lee, Juhnyoung, Podlaseck, Mark, Schonberg, Edith, and Hoch, Robert, Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising, Data Mining and Knowledge Discovery, 5(1/2), 2001.

[10] Kohavi Ron, Brodley Carla, Frasca Brian, Mason Llew, and Zheng Zijian, KDD-Cup 2000 Organizers' Report: Peeling the Onion. SIGKDD Explorations Volume 2, issue 2, 2000. http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html

[11] Pang-Ning Tan, Vipin Kumar, *Discovery of Web Robot Sessions based on their Navigational Patterns*, Data Mining and Knowledge Discovery, 6(1): 9-35 (2002) http://www-users.cs.umn.edu/~ptan/Papers/DMKD.ps.gz

[12] Pyle Dorian, Data Preparation for Data Mining, Morgan Kauffman Publishers, 1999

[13] Spiliopoulou Myra, Mobasher Bamshad, Berendt Bettina, and Nakagawa Miki. A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage In INFORMS Journal of Computing, Special Issue on Mining Web-Based Data for E-Business Applications, Vol. 15 No. 2, 2003. http://maya.cs.depaul.edu/~mobasher/papers/SMBN03.pdf

[14] Zhang Jun, Silvescu Adrian, and Honavar Vasant, Ontology-Driven Induction of Decision Trees at Multiple Levels of Abstraction, Proceedings of Symposium on Abstraction, Reformulation, and Approximation, volume 2371 of Lecture Notes in Artificial Intelligence, Springer-Verlag, 2002.

[15] Zheng Zijian, Kohavi Ron, Mason Llew, Real World Performance of Association Rule Algorithms, Proceedings of the Knowledge Discovery and Data Mining conference (KDD 2001), pages 401-406, 2001. http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html