Advancing Web Analytics

Kelly Storm

Department of Computer Science, UGA. Black Box Operations, LLC

February 14, 2011

Web analytics consists of collecting and analyzing internet data, such as the information found within web server access logs, in order to generate information for the purposes of understanding web usage and web user behavior. Web analytics is divided into two categories: off-site and on-site. Off-site analytics determines the website's audience, share of voice, and buzz in relation to the internet as a whole, whereas on-site analytics consists of measurements that require the information generated by users once they are on your website. Web server logfile analysis, page tagging, and a hybrid method containing both are the three technological approaches to collecting data for on-site analytics. Log file analysis relies on the interpretation of information found within web server log files in which user requests for pages and information are stored, and page tagging uses JavaScript to notify a server housing the website's usage information when information is requested by a web user. The hybrid approach simply combines these two approaches in order to capture previous, current, and future usage information. Our research can best be described as the analysis of web server access logs in an off-site analytics approach to better understand user behavior.

Web analytics software consists of packages, platforms, or applications that create reports or views of information with the data retrieved through on-site analytics approaches. A few of the key metrics used by analytics packages are hits, a user request for a file, page views, a user request for a page, and visits/sessions, a sequential series of page views requests by a single user. While analytics packages often provide a plethora of information that may help site administrators better understand their usage data, current methodologies and current software frameworks lack the functionality and interactivity necessary to support administrators in answering certain classes of questions. We have addressed several of these limitations in the form of a framework that provides interpretations of information and levels of interaction that are required to answer questions more closely related to the site's usage: Ajaxalytics. For example, current analytics applications are capable of generating reports detailing referring sites that contribute heavily to a site's inbound traffic, but are unable to detail, dissect, or visualize the usage patterns of these users in any meaningful fashion.

To better understand the significance of our contributions we must first holistically comprehend the current state of site analysis as well as the evolution of research conducted and products created within this domain.

Since the inception of ARPANET on October 29th, 1969, the Internet has evolved from a small set of commands that allowed peers to send and receive packets of information between a few computers around the United States to a massive hierarchical routing system allowing for the support of the World Wide Web, an integral platform that almost all businesses, organizations, educational institutions, and people utilize and rely upon for their daily activities.

As the web evolved, researchers developed techniques to extract useful information or knowledge from web page contents, a collection of techniques known as web content mining [15, 1, 5]. Researchers developed web structure mining techniques to help site administrators understand the relationship between resources by analyzing the link structure of the hyperlinks within web documents and web sites, and across multiple web sites [7, 3, 22]. Web usage mining involves discovering the users' patterns of behavior online, and usage mining techniques were developed to extract useful information from resources containing usage data such as *web server access logs*, logs created and maintained by the web server containing information about user requests [19, 13, 16]. Web mining can thus be described as the collection of content, structure, and usage mining techniques.

We believe that site analysis' origins can be traced back to the researchers who first applied content mining techniques traditionally reserved for classifying, categorizing, and annotating physical documents, such as plain text files, books, and articles [15, 5]. Given this point of origin, we followed the evolution of site analysis through several distinct domains and determined key junctures that served to strengthen site analysis, mainly through incorporating other domain's strengths to existing methodologies. The details of this evolution will be available in later publications.

Long story short, site analysis has evolved from naive text parsing algorithms to methodologies that parse usage data and present reports and visualizations that help aid administrators in better understanding their site's users' behavior. Over the past few years we have been defining, understanding, and overcoming several technical challenges present in the struggle to create strong methodologies for site analysis. The truth of the matter is that the information and systems people are most familiar with have been around for a relatively short period of time, roughly a decade. This shift in perspective, where the raw power of site analysis hinged almost entirely on a handful of metrics, is largely due to W3C's definitions, generalizations, and categorizations, and Google's entrance into the market in the late 90s as a technological juggernaut [2]. After that, analytics applications were being created and deployed or made available for download at a faster pace [21, 4, 6, 17]. One of the larger issues that arose as these events unfolded can be seen by observing the evolution of site analysis and identifying the key junctures mentioned above that facilitated significant changes in the methodologies comprising usage data analysis: we have reached the next juncture, and innovation is required in order to help other domains properly utilize and better understand usage data.

Boil that down to a sound bite and it goes something like this: We realized early on that analyzing usage data helps us better understand how users are requesting virtual resources and traversing virtual structures. While advancements have been made in the past that provided contributions that changed the landscape by integrating the strengths and methodologies of similar domains with site analysis, the time has come to start breaking some ground and re-imagining the existing procedures.

Instead of the anticipated evolution, however, things have gone stale.

The last notable integration was business marketing with basic site metrics, which bred applications and methodologies that gave us indicators such as return, bounce, and click-through rates. There are a plethora of other significant contributions such as integrating and evolving clustering algorithms to better group users based on their behavior, but there are very few instances of research that created applications that are either deployed and stick, or are adopted, by the mainstream business arena.

Innovation has slowed to a crawl and been replaced by production through replication. The machine's rusted, and its cogs may turn and shift but only out of mechanical response.

We believe this is due to the fact that current applications and methodologies provide enough correct information, but only to some of the people some of the time. For example, Google Analytics will provide information that can be used to directly increase sales, but only for companies and organizations whose business models and business-to-customer processes, such as simple product sales, adhere to necessary virtual structure standards. Google's JavaScript engine used to capture and return user requests and information, Urchin, is simply a plugin, and unless the plugin is used correctly in all instances of deployment and the conversion campaign is managed correctly through Google's web interface, the results may be significantly skewed. Not every business has control over their web store. Not every business has the hired geek power necessary to ensure proper incorporation, implementation, and management of such a tool. My experiences with sites both in academia and in business, while subjective and anecdotal, have shown me that very few people get it right and thus are provided with incomplete information. They learn to not trust it, they evolve to not need it. Furthermore, analytics platforms create additional views and reports in response to any information the existing views or reports lack. We're not strengthening the current methodologies, we're adding fluff to the figures. You can continue pumping air into your flat, but you will eventually need to remove the nail.

So what happened?

The necessity to incorporate site analysis into the development cycle either never occurred or fell by the wayside. Site analysis utilities are seen as superfluous, nothing more than additions or plugins that may contribute to, but are not solely responsible for, administrators and developers better understanding or validating existing assumptions about their user base. Site analysis isn't treated as an integral part of site development and management and thus its already minuscule importance, due largely to the lack of utility, waned. Decreasing significance begets fewer tools. Fewer tools, lower adoption rates. Lower adoption rates, decreased general reliance. Scientific investigation and proper interpretation of usage data lost the battle to educated guessing. While websites became web applications with global user bases, the evolution of site analysis methodologies and implementations slowed and the reliance on tools offering analyses is slowly disappearing. Businesses and organizations consider such methods an afterthought instead of a frontline offense, and this translates directly to higher development costs and longer project times because the only other viable option is educated guessing. Think about that for a second. Apply it to a different field. This would be akin to pharmaceutical companies relying on lab work at a staggeringly decreasing rate due to it's lack of significant information, and instead taking the medication themselves and holding meetings to discuss why they believe the walls are melting and Tom looks oddly like your third grade teacher. It has become common practice for developers and administrators to kick around ideas, ask a few users a few questions, draw some pretty graphs, and hinge entire revisions on the subsequent conclusions and results.

User studies exist, sure, but they're site specific, locally scoped. Research has been conducted to create better methodologies for behavioral clustering, behavior interpretation, view creation, etc., but research in this realm is sparse and often abandoned after the initial investigation and evaluation. As with any research, results are often skewed and evidence contradicting a researcher's theory is often omitted. When committed researchers conduct thorough investigations and generate signification results, their impact is felt. Menasce's [8, 9] and Nasraoui's [11, 14, 10, 12] contributions directly influenced our research, but we don't have a multi-billion dollar corporation ready to take our methodologies and turn them into stable hosted services, generalized for the purposes of allowing any site manager the ability to apply state of the art technology born from fresh, viable research directly to their site's usage data. Incorporating genetic algorithms into clustering techniques to better group users based on their traversal patterns is neat. And it stops at "neat". Innovation slowed because analytics was either removed or never properly inserted into the web development cycle.

Focus up front class. Flip to the present page. We have global economy chock full of businesses that rely heavily on websites and web applications. The NIH and CDC are directly funding projects such as the EuPathDB project so that scientists can turn to a web resource and treat it as a service to more effectively and efficiently study and take down a variety of nasty beasties. Telecommunications is booming and the demand for VoIP services is becoming common practice for even small businesses. Entire industries exist virtually and we can even sell virtual products for real money while governments scramble to figure out the tax repercussions of selling something that doesn't technically exist. The internet changed everything, and the ensuing information goldrush has forced the world to run at a different pace. Legislation isn't written fast enough to keep up and any attempt to enforce such legislation, think Torrenting music and movies, is met with backlash. Opening a brick and mortar business and ignoring the significance of managing an online presence is similar to filing Chapter 7 early and liquidating the merchandise before it's even shelved. Blizzard, the creators of the popular MMORPG World of Warcraft, make $4,000 a minute, and the users have been able to sell their virtual characters and their character's virtual gold and items, for hundreds of thousands of dollars.

All of that is happening right now. Citing sources is unnecessary. “Google it”.

Given this information, we’re left to wonder where usage analysis fell through the cracks. Wouldn’t it make more sense to build your million dollar idea around user expectations? Wouldn’t it make more sense to let user behavior determine, or at least largely direct, site revisions and additions? If the developers of multifaceted web application funded by grants from government organizations don’t have tools to help them better understand their user base, then they’re only option is to hope that the changes made given the small amount of available information stick. Even then, how do we count the noodles we threw against the wall?

We can’t. It doesn’t exist. Current methodologies consistently fall short in very distinct ways. It’s not as if those developers intentionally ignored user expectations, it has more to do with the difficulty of obtaining information that directly reflects and intelligently and clearly conveys user behavior.

This translates into frustrated developers sitting through an endless barrage of meetings and sifting through user studies in hopes of detecting or intelligently isolating information that can be used to implement changes. Changes that the majority of their user base will approve of, adopt, and use to better reach their own goals. These same developers have no way of determining whether those changes are better on any statistically significant level. Look at it this way; Baskin-Robbins creates a new flavor, asphalt and anchovies, and asks a handful of ice cream experts a series of questions regarding the new flavor’s texture, consistency, etc. If these experts agree, or if the researchers can infer that they agreed by positioning the results in an attractive fashion, then we have ourselves a new flavor and Baskin-Robbins will go on to ruin birthday parties around the world.

This translates into business owners coordinating the company’s next online move with incomplete or incorrect information, then move forward in a direction they believe may yield profit in the form of increased sales, higher conversion rates, etc. When the profit doesn't come people are fired. Hire a person and give them a sponge, then fire them when they can't build a house.

This translates into graphic designers developing clunky websites based on a requirements elicitation concocted by marketing strategist. These strategist are fueled by information that hasn't been cross referenced or validated against the actual behavior exhibited by real users.

Then we have user studies and the observer-expectancy effect; the researchers' implicit, unintentional cognitive bias subconsciously influencing the participants of any given experiment. We also have conscious, intentional tweaks to make it appear as if the information retrieved by any given user study is "good". There's only so much information people, directed and aware of their observers, will provide. User studies have provided and will continue to provide a wealth of significant information to site administrators and developers.

But we're sitting on a gold mine here. Why host workshops for a few professionals? In order to ascertain significant usage information, sure. Researchers then weaponize this information in order to pierce through the very heart of shoddy development and design. However, we can pull an arsenal out of log file analysis, page tagging, and hybrid approaches, but current implementations give us sporks instead. Why?

It's because Google already did it or they will in the future. It's because the open source market is flooded with knock offs. It's because researchers tried using the applications but didn't know what to do, or what the reports were attempting to convey. It's because a business's marketing campaign said they had a 56% increase in conversions but they're down $3500 this month. It's because people genuinely don't care that 17 people came from Google, and that telling them they typed in a few keywords provides nothing significant enough to make effective alterations or additions to their site or application. It's because knowing the location of one's users does little more than allow for broad generalizations and assumptions based on existing cultural or sociological biases. It's because the only thing a person can do with generic reports is print them and show them off. Large green arrows and numbers may impress a boss, but the numbers are contextually meaningless and the information is hollow.

It's because people genuinely have questions and no one's giving them answers.

The answers don't have to be entirely correct, they don't have to be earth shattering conclusions that serve as the singular reason behind a business's success. They just have to be better than what we're currently providing to developers and administrators.

Creating methodologies and techniques that enable administrators to reasonably comprehend the answers to their questions is vital to properly integrating or reintroducing site analysis into the development cycle. Providing reliable platforms, abstracted and generalized, to allow for any site developer to gracefully incorporate site analysis into their system should be a top priority. Best bet? Develop it, strengthen it, prove it, provide it, increase the customer base, call Google, sell it. Got time? Hang up the phone, and draw a line back to the strengthening bit, keep at it.

So where do we begin? We can't simply create another view with contextually meaningless numbers or export another PDF with graphs and visualizations representing said contextually meaningless numbers, it isn't working. So how do we go about strengthening site analytics to the point that it is properly reintroduced, adopted, and treated as part of the general development cycle? We start by bending ideas. History has taught us that true innovation is mapped neatly to points in time where researchers combined the strengths of one domain with the strengths of another then evaluated the subsequent hybrid approach and generated meaningful results.

We started by applying Menasce's customer behavior model graphs to sessionized access logs. The generated graphs were too large, the state-space contained too much information,

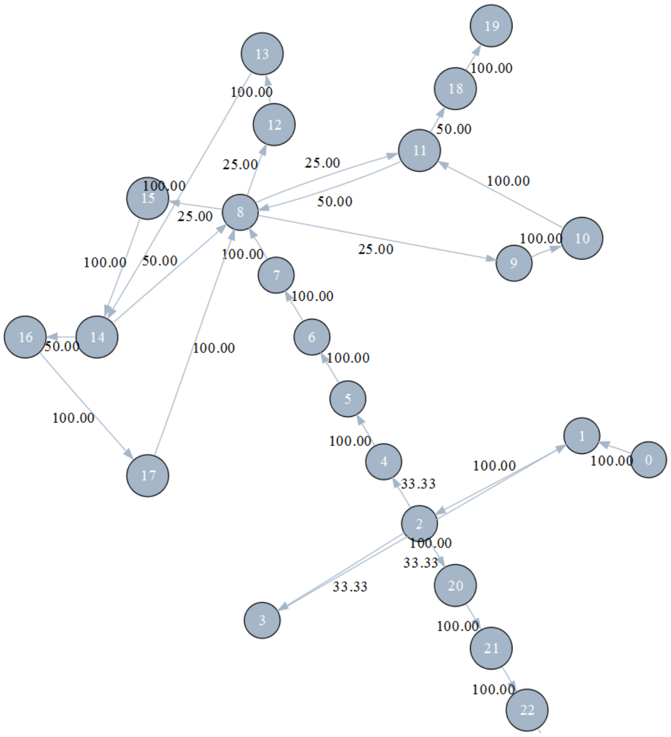


Figure 1: CBMG created from ApiDB usage data

and the evaluators were ultimately unable to determine anything meaningful or significant from the visualizations.

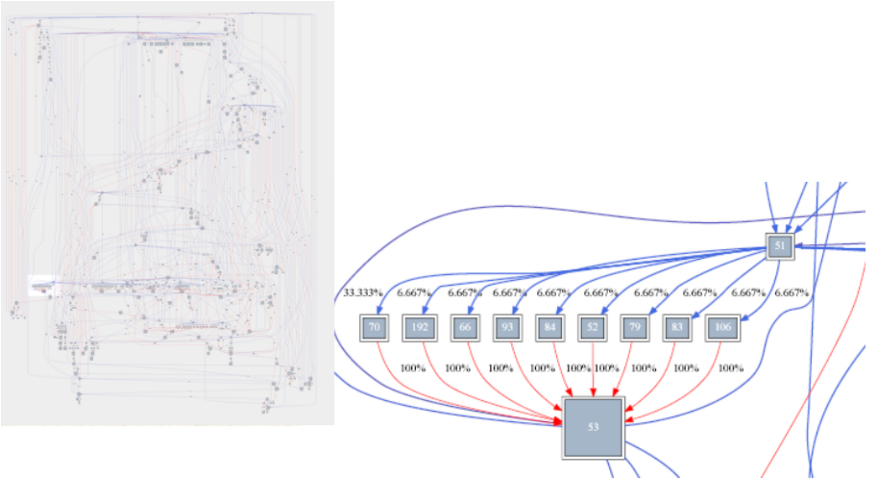


Figure 2: ECBMG created from ApiDB usage data

In an attempt to increase the visibility of potentially significant information, to offset the signal-to-noise ratio, we enhanced the customer behavior model graphs. Higher transition probability? Visually depict the cold to hot color range then. Higher request frequency? Have the size of the state reflect its frequency of occurrence. These ideas came from analyzing basic graphical representations. When a state is larger, it implies some numerical representation is greater than that of a smaller state. If something is blue, or cold, it represents a lower transition probability. Vice versa for red, or hot.

The textual representation of figure 2, where the image on the right hand side is a zoomed in portion of the image on the left hand side, is roughly translated as “strike two”.

There's still too much information. Sengupta introduced the concept of transaction-oriented views (TOVs) [18] and we thought applying TOVs might help minimize the state-space. We manually generated customer behavior model graphs after grouping users based on their behavior, selecting series of requests which could be meaningfully categorized as transactions, and determining two distinct periods of time for a comparative evaluation. We called the result a Time Series Analysis Selected Episode Graph (TSA-SEG), figure 3 as an example. The evaluators were able to infer meaningful information from the observations made against the generated TSA-SEG, and we went on to publish our results in WSE09 [20] which went over well with the community. Not much fuss was made over the graphical representation of data that allowed for the efficient extraction of information and insights, the research world was already flooded with such approaches. Still, forward progress was made. We took our ideas and the results of our investigation and pressed on.

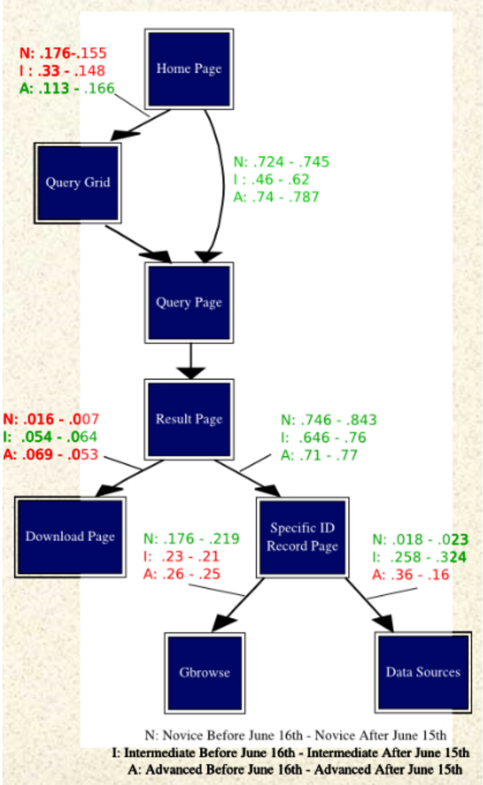
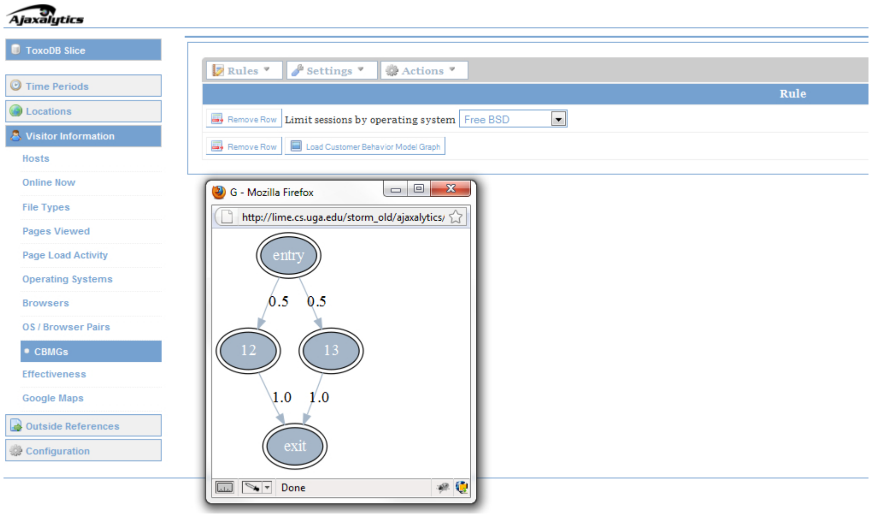


Figure 3: TSA-SEG created from ApiDB usage data



Limit sessions by operating system: Free BSD

Figure 4: Ajaxalytics, Version .1

Upon further investigation we determined that evaluators required the ability to create TSA-SEGs dynamically, allowing for rule sets and graphs to be created and refined in real time at the evaluators' own pace. We then theorized about the real advantage of our approach. Was it the graphs or the automated aspect of creating the graphs? I thought it was the graphs; not so much. It was mainly the automation. It was the rule sets. It was the freedom to freely interact with the usage data. This was staring us right in the face and we didn't see it, so now that we know let's roll up our sleeves get to work.

Enter Ajaxalytics, stage left.

Did we just create an application that allows evaluators to dynamically select information to refine and create CBMGs and TSA-SEGs? Yes we did! High five.

Now finish that whiskey, sleep off your hangover, and get back to work because the system's not that great.

The system was clunky and slow, and the interactivity was limited to rule selection. The graphs were neat, but they weren't interactive and they raised more questions than provided answers. We were contributing to the problem as well as the solution. Two steps forward, one step back. But it's still a step forward right? Right.

Let's keep going then. Back to the Batcave. It's time to learn us some web technologies, hone our skills, figure out how to build a better system, and prepare for the fight so we can come out of our corner swinging. It's time to dig into research and refine existing theories. It's time to create new ones.

But theory only blazes the trail, it doesn't pave the road.

Well then, let's line up the enemies so we can knock them down.

The access logs are enormous, the parser needs to be as fast as it is accurate. The leading open source analytics application, AWStats, parses access logs at a rate of 32MB/minute when given the maximum resource allowance of 512MB of memory according to the benchmarks posted on their website. This requires an hour and thirty-six minutes to parse a 3GB access log. Our own tests generated times closer to three hours.

The sessions are delimited by time stamps, and the requests are ordered by their occurrence within the log file, so we need integrate the time spent by users' on any given page into the sessionization process as well as the models and visualizations so that evaluators can visually determine things like groups of **push pages**, where a page may be a simple redirect to another page or a page is silently, implicitly requesting other pages that the web user never technically sees or utilizes.

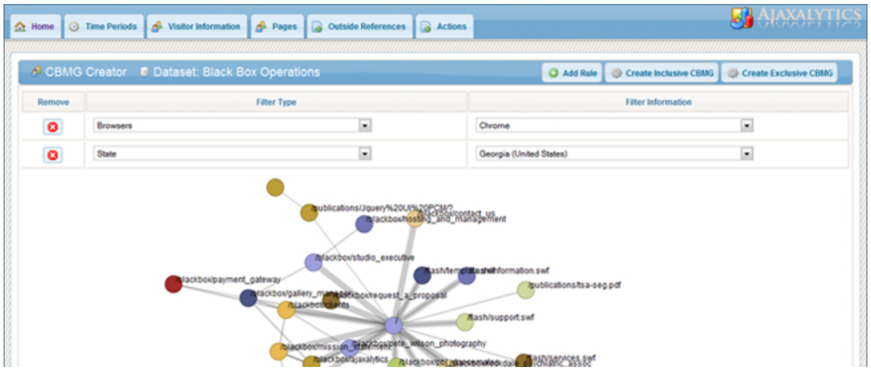
We need to be able to cluster users based on their behavior in contextually meaningful ways, either through a machine learning approach where archetypes are isolated and presented to the user or a rule based approach where the evaluator can create their own groups of users based on a small set of selected sessions exhibiting behavior representative of the group itself.

The graphical representations of user behavior, regardless of what acronym they're given, need to be as interactive as the rules used to create them. The rules should be expanded to provide for meaningful boundaries, such as important episodes or back linked keywords used to locate a site within a given time span.

We could visualize the usage behavior of Firefox users for the month of June, so what? I want to see how my marketing campaign is working, I want to see users who came to my site last week by clicking a link within an email advertisement and landing on this specific target page, I want to see what they did, I want to see where they went, I want to see where they left, I want to group some of those pages as transactions, I want to save it for later use or create another graph and have them side by side so I can do my own comparative analysis, I want to see if this ad campaign was any better than the last ad campaign in terms of users traversing to a small group of pages targeted towards prospect to customer conversion and sales. I want all of that with the few clicks of a mouse, I don't want to spend half my week in a mire of confusion. You know what, scratch that, my marketing campaigns always fall flat, they never produce results, but there are a few keywords people keep typing into search engines that have my site pop up as the first result. I want to see information about those people for these keywords right here, I want everything I asked before to go along with it.

The information that's currently provided by existing analytics applications should also be provided through our implementation. We've observed that creating supplemental approaches will hinder the chances of integrating our ideas into mainstream site analytics. Adoption rates would suffer if we're nothing more than a single-serving tool. Let's take the existing framework and wrap it around our research in a way that allows users to interact with the usage data in a holistic, global manner.

Let me hang my social life from the rafters and get back to work.


Home
Time Periods
Visitor Information
Pages
Outside References
Actions
AJAXALYTICS
CBMG Creator
Dataset: Black Box Operations
Add Rule
Create Inclusive CBMG
Create Exclusive CBMG
/publications/Jquery%20UI%20PCM/?
/blackbox/contact_us
/blackbox/hosting_and_management
/blackbox/studio_executive
/flash/information.swf
/blackbox/payment_gateway
/publications/tsa-seg.pdf
/blackbox/gallery_management
/blackbox/request_a_proposal
/blackbox/clients
/
/flash/support.swf
/blackbox/mission_statement
/blackbox/pete_wilson_photography
/blackbox/ajaxalytics
/flash/services.swf

Remove	Filter Type	Filter Information
☒	Browsers	Chrome
☒	State	Georgia (United States)

Figure 5: Ajaxalytics, Version 1.0

Ajaxalytics version 1.0, shown in figure 5 is released with major improvements across the board. We didn't continue improving Ajaxalytics version .1, we started from scratch.

While similar to version .1, 1.0 is more user friendly and the evaluators are able to use the application to help answer a handful of simple answers.

It's abstracted, it's modularized, it's faster, and with a bit more elbow grease and patience it'll be useful. Useful in the way we've suggested throughout the entirety of this rant.

We released version 1.1 a few weeks after 1.0's release as a response to issues that needed to be addressed immediately.

We're almost finished with Ajaxalytics version 1.2.

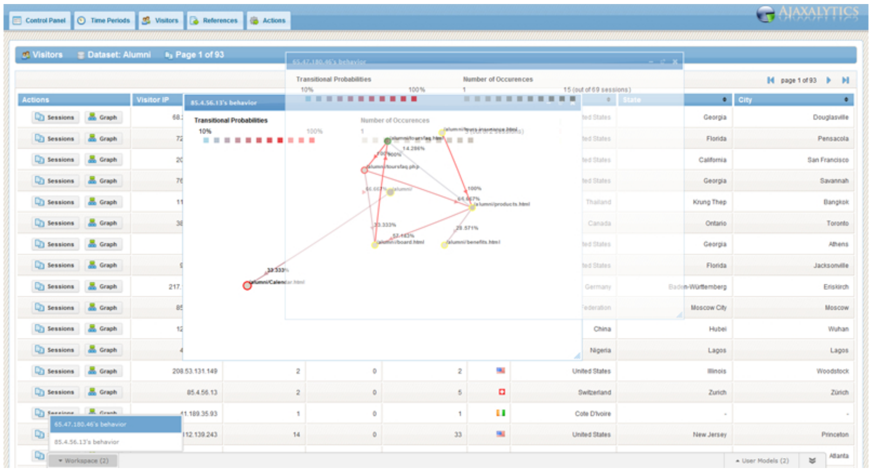
Install processes and additions to the interface have been added to decrease the level of technical knowledge necessary to install the application.

We added an interface that allows for the dynamic creation and control of datasets, or application level views of several different access logs. Instead of analyzing a single access log, evaluators can now move between access logs by selecting a different dataset. Our parsers are crunching the logs at 150MB/minute given 256MB of memory. Half the resource usage and 4.5x faster than AWstats.

We incorporated a windowing system that allows users to minimize, maximize, and restore any view created within the website. The system also allows users to carry the views to other modules and save the views for later user. Instead of clicking back and forth between modules, an evaluator can create views that are presented as draggable, interactive windows then carry these views over to other modules in order to facilitate comparative analyses. If the views are important then a snapshot of the view can be saved and restored when the evaluator returns to the website to continue their investigation. This is a big deal: we've added persistence to dynamically generated information then decoupled/delocalized the information from its parent page in order to exhibit the behavior of a desktop application.

Sessionizing takes timestamp information into consideration, requests are ordered by time instead of occurrence in the actual access log, graphs are adjusted to shorten the distance between states, shown as nodes, when the staying time is unusually low, and grouped together when the staying time is 0 so as to indicate push pages.

Stepwise Comparative Model Graphs were created and integrated into the application. Evaluators are now able to isolate sessions by time, select multiple session representatives for multiple groups in order to cluster based on usage behavior, select potentially meaningful



Actions	Visitor IP						State	City
Sessions Graph	68.					ted States	Georgia	Douglasville
Sessions Graph	72					ted States	Florida	Pensacola
Sessions Graph	20					ted States	California	San Francisco
Sessions Graph	76					ted States	Georgia	Savannah
Sessions Graph	11					Thailand	Krung Thep	Bangkok
Sessions Graph	38					Canada	Ontario	Toronto
Sessions Graph						ted States	Georgia	Athens
Sessions Graph	9					ted States	Florida	Jacksonville
Sessions Graph	217.					Germany	Baden-Württemberg	Eriskirch
Sessions Graph	85					ederation	Moscow City	Moscow
Sessions Graph	12					China	Hubei	Wuhan
Sessions Graph	4					Nigeria	Lagos	Lagos
Sessions Graph	208.53.131.149	2	0	2	🇺🇸	United States	Illinois	Woodstock
Sessions Graph	85.4.56.13	2	0	5	🇨🇭	Switzerland	Zurich	Zürich
Sessions Graph	41.189.35.93	1	0	1	🇨🇮	Cote D'Ivoire	-	-
	112.139.243	14	0	33	🇺🇸	United States	New Jersey	Princeton
								Atlanta

Figure 6: Ajaxalytics, Version 1.2

transactions from a list retrieved by analyzing the subset of selected sessions and isolating longest common subsequences, and interact with visualizations depicting the usage behavior of the groups of sessions, or clusters, by observing the sequential nature of each groups' collective traversal in a stepwise fashion.

Customer Behavior Model Graphs now allow for contextually meaningful rules such as isolating users who traversed from major search engines by typing in a keyword found within the user defined rule set.

The existing modules and utilities created to provide evaluators with the tools they've become accustomed to have been vastly improved. A data hierarchy was created and employed within each view providing the ability to drill down multiple levels.

The next step is to perform case studies for version 1.2 and determine the effectiveness of our ideas when applied to real businesses and organizations.

To be continued.

Bibliography

[1] Daniel Billsus and Michael J. Pazzani. A hybrid user model for news story classification. In *Proceedings of the Seventh International Conference on User Modeling*, 1999.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[3] Soumen Chakrabarti, Byron E. Dom, David Gibson, Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Mining the link structure of the world wide web. *IEEE COMPUTER*, 32:60–67, 1999.

[4] Frei Software Development. Webalizer xtended. Accessed Online, 2010. `http://www.patrickfrei.ch/webalizer/`.

[5] Ronen Feldman, Moshe Fresko, Yakkov Kinar, Yehuda Lindell, Orly Liphstat, Martin Rajman, Yonatan Schler, and Oren Zamir. Text mining at the term level. In *PKDD '98: Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 65–73, London, UK, 1998. Springer-Verlag.

[6] Alentum Software Ltd. Weblog expert - powerful log analyzer. Accessed Online, 2010. `http://www.weblogexpert.com/`.

[7] Sanjay Madria, Sourav Bhowmick, W. Ng, and E. Lim. Research issues in web data mining. In Mukesh Mohania and A Tjoa, editors, *DataWarehousing and Knowledge Discovery*, volume 1676 of *Lecture Notes in Computer Science*, pages 805–805. Springer Berlin / Heidelberg, 1999.

[8] Daniel A. Menascé, Virgilio A. F. Almeida, Rodrigo Fonseca, and Marco A. Mendes. A methodology for workload characterization of e-commerce sites. In *EC '99: Proceedings of the 1st ACM conference on Electronic commerce*, pages 119–128, New York, NY, USA, 1999. ACM.

[9] Daniel A. Menasce and A. F. Almeida Virgilio. *Scaling for E Business: Technologies, Models, Performance, and Capacity Planning.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000.

[10] Olfa Nasraoui. Mining web access logs using relational competitive fuzzy clustering. In *Proceedings of the Eight International Fuzzy Systems Association World Congress*, 1999.

[11] Olfa Nasraoui. R.:an evolutionary approach to mining robust multiresolution web profiles and context sensitive url associations. *International Journal of Computational Intelligence and Applications*, 2:339–348, 2002.

[12] Olfa Nasraoui and Hichem Frigui. Extracting web user profiles using relational competitive fuzzy clustering. In *International Journal on Artificial Intelligence Tools*, volume 9, pages 509–526, 2000.

[13] David Nicholas, Paul Huntington, and Anthony Watkinson. Scholarly journal usage: the results of deep log analysis. *Journal of Documentation*, 61(2):248–280, 2005.

[14] Elizabeth Leon Olfa Nasraoui and Raghu Krishnapuram. Unsupervised niche clustering: Discovering an unknown number of clusters in noisy data sets. *Evolutionary Computation in Data Mining*, pages 157–188, 2005.

[15] Helena Ahonen Oskari, Helena Ahonen, Oskari Heinonen, Mika Klemettinen, and A. Inkeri Verkamo. Applying data mining techniques for descriptive phrase extraction in digital document collections. In *In Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*, pages 2–11, 1998.

[16] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Hua Zhu. Mining access patterns efficiently from web logs. In *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 396–407, London, UK, 2000. Springer-Verlag.

[17] Salvatore Sanfilippo. Visitors, a fast web log analyzer. Accessed Online, 2010. `http://www.hping.org/visitors/`.

[18] Shubhashis Sengupta. Characterizing web workloads - a transaction-oriented view. In Samir Das and Sajal Das, editors, *Distributed Computing - IWDC 2003*, volume 2918 of *Lecture Notes in Computer Science*, pages 833–833. Springer Berlin / Heidelberg, 2003.

[19] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, 2000.

[20] Kelly Storm and Eileen Kraemer. Web site evolution: Usability evaluation using time-series analysis selected episode graphs. In *WSE '09: Proceedings of the Ninth IEEE International Symposium on Web Site Evolution*, pages 27–38, 2009.

[21] K.-L. Wu, P. S. Yu, and A. Ballman. Speedtracer: a web usage mining and analysis tool. *IBM Syst. J.*, 37(1):89–105, 1997.

[22] Shi Zhou, Ingemar Cox, and Vaclav Petricek. Characterising web site link structure. *Web Site Evolution, IEEE International Workshop on*, 0:73–80, 2007.