

Data Quality Challenges and Solution Approaches in YAHOO!'s Massive Data Environment

ABSTRACT-----

After significant initial Data Quality wins in Yahoo!'s Audience pipeline, there are many technical challenges in identifying next level DQ issues and organizational challenges of institutionalizing the data quality program in Yahoo!'s massive data environment. Below are some of the areas of focus:

- Application of time series analysis for detection of data anomalies
- Lean Manufacturing and Six Sigma application to data pipeline analysis
- Standardized DQ Assessments across all products
- Application of value of information techniques in prioritization of quality improvements

BIOGRAPHY-----

Dan Defend

Yahoo!

Dan Defend earned his Masters in Computer Science from University of Illinois. He has experience at Motorola as an Engineering Manager in Unix OS and embedded cell phone software where he also led analysis and data-driven improvements using Digital Six Sigma methodology. He is currently leading the Data Quality program at Yahoo! which relies heavily on organizational leverage and distributed ownership and accountability.




Aparna Vani


Yahoo!

Aparna Vani earned her Masters in ECE from University of Houston. She worked as a Hardware Design engineer at Compaq and development lead at TVGuide. She has extensive analysis, quality and architecture experience at Motorola. Currently, she is the Chief DQ Architect at Yahoo, responsible for DQ strategy and design, and responsible for organization-wide DQ improvement projects involving centralized monitoring and data cleansing across Yahoo!'s data pipelines.






MIT Information Quality Industry Symposium July 09




Data Quality Challenges and Solution approaches in YAHOO!'s massive data environment

Data Quality Team, Yahoo!
qa-dq-champaign@yahoo-inc.com

Executive Summary/Abstract: After significant initial Data Quality wins in Yahoo!'s Audience pipeline, there are many technical challenges in identifying next level DQ issues and organizational challenges of institutionalizing the data quality program in Yahoo!'s massive data environment. Areas of focus includes: Data Monitoring and alerting challenges, Browser Cookie Churn, Traffic Protection (robot filtering).



MIT Information Quality Industry Symposium July 09



Data Quality Challenges

- Detection of data anomalies in traffic monitoring
- Browser Cookie Churn
- Traffic Protection

MIT Information Quality Industry Symposium July 09



Data Quality Challenges

- Detection of data anomalies in traffic monitoring
- Browser Cookie Churn
- Traffic Protection


MIT Information Quality Industry Symposium July 09

Detection of data anomalies in traffic monitoring


- Yahoo! usage typically consistent day over day
- Fluctuations in traffic usage can exist due to
 - geography concerns
 - Seasonality
 - Special events
- Goal: Ensure our data is properly vetted for our customers





Source: Nielsen Ratings
Yahoo! Sports doubles (year over year) number of users to 18.7 million due to Olympics.
Oct 3, 2008 - Reuters



Source: BlogPulse
Yahoo! News was the second most cited news source on information regarding hurricane Katrina
Sept 6, 2005 - Greg Jarboe

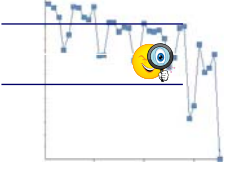


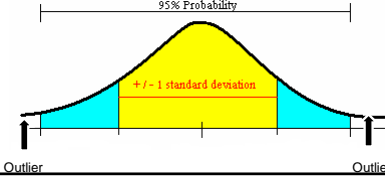
Source: Nielsen Ratings
Yahoo! News unique users spike over 13% for the 2009 Inauguration.
Jan 23, 2009 - Nielsen Ratings


MIT Information Quality Industry Symposium July 09


Trending Analysis

- To validate trends, we incorporate statistical analysis to ensure accurate numbers
- Monitoring system uses **standard deviation** of specific metric from given input source.
- Calculating the **mean of past x days** and the number of standard deviations away from mean
- Using the **test statistic** to determine “out of bounds”
- Ensure our **new value is within normal range of the mean**







95% Probability

+/- 1 standard deviation

Outlier

1. Notify Stakeholders
2. Identify Root Cause
3. Act as appropriate


MIT Information Quality Industry Symposium July 09






Trending Analysis – Challenges with Usage Patterns

Alpha risk (missing the actual warnings):

- Irregular Patterns in traffic volume
- Huge difference between weekday and weekend traffic pattern
- Outliers like Holidays, Seasonal events changes the std deviation

Beta risk (False Alarm):

- Holidays, Seasonal events
- Pages with short shelf life
- Highly fluctuating, content dependent properties

			
All-year, Weekday versus Weekend Or Seasonal or cyclical traffic patterns	Pages with shelf life of 1 day Or Malicious traffic spike	Pages which see increased, constant activity Or New Page Launch	Pages which fluctuate depending on content Or Consistent Robotic fluctuations

MIT Information Quality Industry Symposium July 09

Data Quality Challenges

- Detection of data anomalies in traffic monitoring
- Browser Cookie Churn**
- Traffic Protection


MIT Information Quality Industry Symposium July 09

Browser Cookie Churn:


What is Cookie Churn?

- Cookie Churn is well known complex issue that impacts most web hosting and web analytics companies.
- Cookie Churn is phenomenon where new cookies are issued to the same browser/same user.
- New cookies are issued due to lack of specific browser cookie key in top level domain in client cookie jar.

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	- Site Reads 3 distinct cookies - Same visitor is counted 3 times				



MIT Information Quality Industry Symposium July 09




Cookie Churn Impacts*


The implications of cookie deletion are far-reaching, affecting both site-centric analytics and ad server analytics, and ultimately leading to inaccuracies for those choosing to rely solely on server-based data.

- **Cookie deletion leads to the following inaccuracies in site-centric measurement:**
 - Overstatement of unique visitor counts
 - Understatement of repeat visitor counts
 - Understatement of conversion rates
- **Cookie deletion leads to the following inaccuracies in ad server measurement:**
 - Overstatement of reach
 - Understatement of frequency

**Source: White Paper - ComScore Cookie Deletion Study (June 2007)*




MIT Information Quality Industry Symposium July 09




Sources of Cookie Churn

- Cookie Cache overflow
- Users flushing cookies intentionally
- Robots, crawlers, spiders accepting cookies
- Users flushing cookies thru tools like Norton anti-virus, IT departments auto-flushing
- Flash apps running within FireFox on Windows and invoking the 'SWFupload' method will actually submit a request with cookies from the IE cookiejar
- Cookie Blocking




MIT Information Quality Industry Symposium July 09




Measurement and Challenges

- **Browser Cookies per User ID – How many browser cookies do we see per user ID?**
 - People logging from different machines (home/work)
 - Multiple browsers by same user
 - Complex and data intensive
- **Based on Global Unique Identifier:**
 - Multiple Global Identifier per browser cookie (Due to different network interface)
 - Multiple browser cookies per global identifier due to multiple users on one system





MIT Information Quality Industry Symposium July 09



Data Quality Challenges

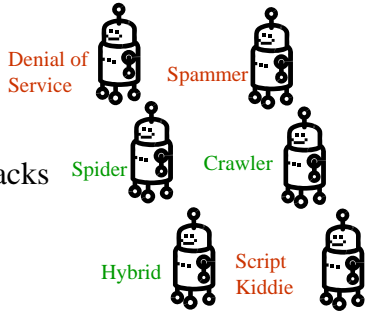
- Detection of data anomalies in traffic monitoring
- Browser Cookie Churn
- **Traffic Protection**




MIT Information Quality Industry Symposium July 09


Traffic Protection (Robot Filtering)

What are robots?

- Crawlers and spiders - the programs that run on systems whose job it is to index the contents of the web. They identify them in 'User Agent' and can be easily filtered out.
- Abuse/Malicious Users
 - Not 'real human beings'
 - Programming scripts
 - Distributed denial-of-service attacks
 - Ad click frauds
 - Spam mail users





MIT Information Quality Industry Symposium July 09



Traffic Protection (Robot Filtering)

Why High-Quality Robot Filtering is Important

- Various critical business functions rely on accurate representation of real user traffic:
 - Assessment of property and network health and growth trending (Robots add noise to data by affecting average page views, time spent, sessions and network retention)
 - Strategic planning decision support
 - User engagement analysis
 - Content optimization
 - Corporate reporting
 - Behavioral Targeting




MIT Information Quality Industry Symposium July 09




Traffic Protection (Robot Filtering)

Robot Filtering Challenges

- Filtering algorithm complexity and impact on data availability (Service level agreement)
- Identifying 'User DNA' to account for filtering based yahoo user ID, browser cookie, uncooked, IP address.
- Different User patterns and behaviors on different yahoo properties and also various regions worldwide.
- Overfiltering vs underfiltering.
- Cost of implementation at initial stage of pipeline with huge data volume.
- Real-time filtering for low-latency data customers.



MIT Information Quality Industry Symposium July 09



Questions ?