MET CS688 C1

Web Analytics and Mining

Zlatko Vasilkoski

Class 1

General Course Information

- Instructor: Zlatko Vasilkoski (email: zlatko@bu.edu or zlatko.vasilkoski@gmail.com)
- Office hours: by appointment (you can email me with questions at any time)
- Class Time: from 6:00pm to 9:00pm, CAS 208 (725 Comm Ave.)
- Course Prerequisites:
  - MET CS 544 - Foundations of Analytics or
  - MET CS 555 - Data Analysis and Visualization
- Course Grading Policy:
  - Quizzes/Lab Projects (10%)
  - Assignments (20%)
  - Midterm exam (35%)
  - Term project (35%)
- Course Topics:
  - Web Analytics and Web Analytics Tools
  - Text Mining
  - Web Mining
  - Mining the Social Web, Twitter, Facebook, LinkedIn, Game Analytics
  - Data Visualization, Google visualization APIs illustrated on above mining examples
  - Basics of machine learning – **Let me know if there is an interest for it!**
- No specific textbook (reference textbooks are listed in the syllabus)

My Background

- PhD. in physics from Tufts University working with David Weaver and Martin Karplus on computational implementation of the diffusion collision multi scale model of protein folding, to which the 2013 Nobel Prize for chemistry was awarded. The algorithms I designed as part of my thesis enabled the diffusion collision model to be applicable to much larger proteins.
- My college teaching experience at Tufts, MIT, Suffolk, BU and Bentley goes back at least for 10 years.
- I have been developing the curriculum and working as a lecturer at the Metropolitan College Computer Science department since 2012.

- My work experience includes
  - Chief Data Scientist at Facility ConneX
  - Senior research scientist at Migma Systems Inc., Neurala Inc.
  - Postdoctoral research work at MIT and Northeastern
  - Worked in the area of Neural Network's learning laws at Department of Cognitive and Neural Systems, at BU.
  - Worked on medical data analytics at Harvard Medical School.

- My current research interests include algorithm development in computational physics, biomedical image processing, computer graphics, computer vision, machine learning and neural network systems for adaptive complex behavior in robots.

Class overview

- This course covers the theoretical and practical aspects of
  - Web Analytics
  - Text Mining
  - Web Mining
  - Internet of Things (IoT)
  - Mining the Social Web
  - Game Analytics

- The web analytics part of the course studies
  - the metrics of web sites
  - their content
  - user behavior during web site visit
  - reporting

- In this class Google analytics tool is used for collection of web site data and implementing the analysis. The use of Google Trends and Google Correlate will be also illustrated.

Class overview

- The text mining part covers the analysis of text and it includes
  - Preprocessing and content extraction from various file types
  - The mathematical (matrix) representation of the extracted text
  - String matching, fuzzy string matching, and their measures of closeness
  - Documents matching in the "concept space" and the simple math behind it
  - Aspects of supervised learning, Tagging, Classification, and Categorization
- The web mining (structure & content) part covers aspects such as
  - Web crawling (gathering pages from the web)
  - Indexing (to support a search engine)
  - Understanding Search Performance and how to measure it
  - The graph representation of the web pages and ranking the web pages
  - Practical applications to the social web and game data

- Illustrations of these concepts are given using R.

- Please indicate how many of you have used R before.

Please Introduce Yourself

- Introduce yourself to me and the other students
- Tell us about your background, your interests, hobbies, etc., so that we can get to know each other better.
- Please describe two or three objectives you hope to accomplish by the end of the course, e.g.
  - How does this course fit into your academic and professional objectives;
  - What do you hope to gain from the course.
- Please describe the type of data you work with and what pattern you typically look for in it.

Introduction

- Most of the information we use today is stored online. There are claims that the data generated over the last 2 years is few fold larger than the data generated previously in the history of mankind.

- Most of this newly generated data is text, images and videos in a form of email, Google, YouTube, Facebook, Twitter, blogs, and most of the other technologies that define our digital age. To this we should also add the new communication tools such as social networks, instant messaging, Yammer, Twitter, Facebook, LinkedIn etc. too.

- By some general estimates a **third** of our time is spent on searching for information and another **quarter** analyzing it. It is widely believed that more and more data will be generated in the near future (IoT) and the time managing this data must be as productive as possible.

- This is just one aspect of what this course is about! We have an exciting journey ahead as we acquire the skills regarding this subject step-by-step.

Examples of Web analytics use

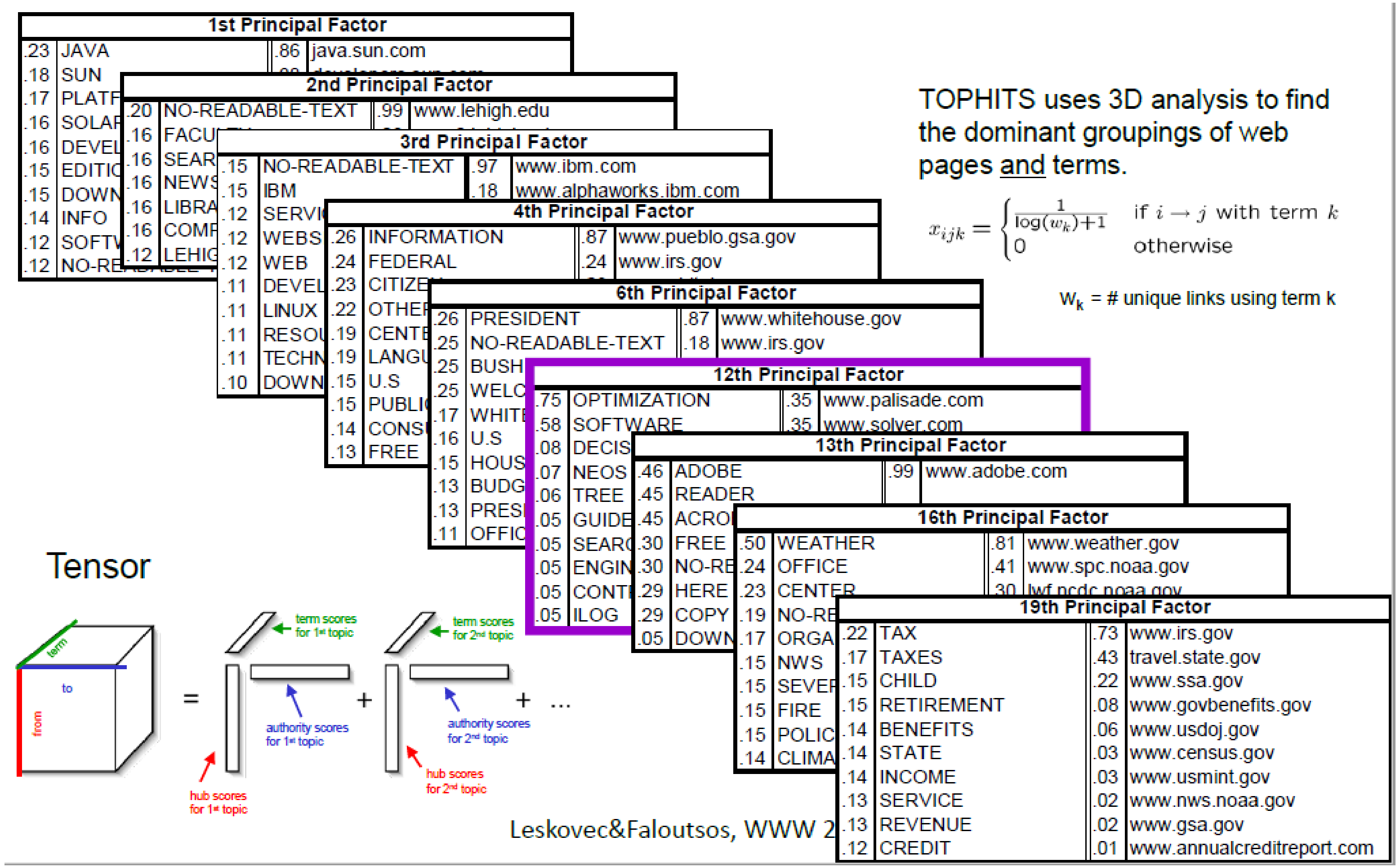
Web analytics is commonly used to give:

- Real-time visibility into web site performance,
- Order status,
- Inventory levels,
- Warehouse management systems.

Data Analytics

- Real data are often in high dimensions with multiple aspects (modes)

- Matrices and tensors provide elegant theory and algorithms



1st Principal Factor			
.23	JAVA	.86	java.sun.com
.18	SUN	.38	developer.sun.com
.17	PLATF		
.16	SOLAR		
.16	DEVEL		
.15	EDITIO		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-RE		

2nd Principal Factor			
.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACU		
.16	SEAR		
.16	NEWS		
.16	LIBRA		
.16	COMP		
.12	LEHIG		

3rd Principal Factor			
.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM	.18	www.alphaworks.ibm.com
.12	SERVI		
.12	WEBS		
.12	WEB		
.11	DEVEL		
.11	LINUX		
.11	RESOU		
.11	TECHN		
.10	DOWN		

4th Principal Factor			
.26	INFORMATION	.87	www.pueblo.gsa.gov
.24	FEDERAL	.24	www.irs.gov
.23	CITIZE		
.22	OTHER		
.19	CENTE		
.19	LANGU		
.15	U.S		
.15	PUBLI		
.14	CONSU		
.13	FREE		

6th Principal Factor			
.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHITE		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRESI		
.11	OFFIC		

12th Principal Factor			
.75	OPTIMIZATION	.35	www.palisade.com
.58	SOFTWARE	.35	www.solver.com
.08	DECIS		
.07	NEOS		
.06	TREE		
.05	GUIDE		
.05	SEARC		
.05	ENGIN		
.05	CONTI		
.05	ILOG		

13th Principal Factor			
.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACRO		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

16th Principal Factor			
.50	WEATHER	.81	www.weather.gov
.24	OFFICE	.41	www.spc.noaa.gov
.23	CENTER	.30	lwf.ncdc.noaa.gov
.19	NO-RE		
.17	ORGA		
.15	NWS		
.15	SEVER		
.15	FIRE		
.15	POLIC		
.14	CLIMA		

19th Principal Factor			
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

TOPHITS uses 3D analysis to find the dominant groupings of web pages and terms.

$$x_{ijk} = \begin{cases} \frac{1}{\log(w_k)+1} & \text{if } i \rightarrow j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

w_k = # unique links using term k

Tensor

term
to
from
=
term scores for 1st topic
authority scores for 1st topic
hub scores for 1st topic
+
term scores for 2nd topic
authority scores for 2nd topic
hub scores for 2nd topic
+ …

Leskovec&Faloutsos, WWW 2

Introduction: Working with data

- Lots of digital data generated over the last few years with expectations that more and more data will be generated in the near future.
- Questions arise how to optimally manage it in less time.
- Typically at some point all this unstructured data is reduced to digital text as most practical form.

- **Analytics** - discovery of relevant patterns in data.
- **Mining** - extracting useful information from data.
- Either way the goal is to learn from data.
- In the language of **machine learning** (to learn without explicitly being programmed) these learning categorizations can be subdivided as
  - **Supervised learning** (we know the categories into which we need to separate the data). Task: for new feature x, predict y)
    - Regression (continuous y data predictions from features x)
    - Classification (discrete y data predictions from features x)
  - **Unsupervised learning** (we don't have any knowledge into what categories the data can be subdivided) - find structure in large data set.
    - Clustering

What is an analytic?

- Methodology of revealing a meaningful pattern in recorded information (data) to quantify a performance.
- Relies on the simultaneous application of several steps and techniques including research and data storage managed by a computer system.
  - Database
    - Typically used for large, long-lived data.
  - The knowledge base data storage (called an ontology) is
    - A dynamic resource
    - An object model with classes, subclasses, and instances.
    - Benefits - being able to store, analyze, and reuse knowledge
    - Typically used to arrive at a specific answer to a problem.
- Use of computer systems to implement
  - Statistics techniques
  - Including machine learning
- Analytics is used to drive decision making
  - By identifying which data is useful and meaningful

Analytics Components

- **Descriptive** analytics – describe what is happening in your system.
  - Simply describes past events and can allow for interpretation in preventing future negative impacts.
- **Predictive** analytics – predict what could happen.
  - Utilizes a variety of statistical, modeling, data mining and machine learning techniques to study historical and recent data.
  - Allows analysts to make predictions about future (positive or negative) events.
  - Currently being able to foresee positive and negative events is extremely powerful feature used to derive marginal advantages over competitors, if not to gain competitive advantage.
  - Predictive analytics takes into account all historical data, allowing for linking of key data points over time to provide predictive features related to operational cost effectiveness and possible downtrends.
- All this is used to prescribe solutions to help mitigate issues along the way.
- This new and growing area of predictive analytics gives the probability of an event and gives the data points needed to mitigate it.
- The combination of analytics and predictive analytics (coding) offers more advanced and cost-effective operation.
- In addition, machine based predictive analytics capabilities can help find the meaning and subset useful data based on input from the user.
- An intuitive predictive coding workflow can validate performance and allow weighing the costs and benefits based on specific measure thresholds (such as precision and recall).
- It can recommend one or more courses of action – and show the likely outcome of each scenario.

Web Analytics Overview

- Web analytics goal:
  - Improving the online experience of your customers and potential customers by data analysis from your business (and the competition).
  - Many examples you might be familiar with (real estate, pharmaceuticals, travel etc.). Can you suggest few more?
- Analytics program includes aspects of
  - Collecting relevant raw data
  - Understand significance contained in the data
- The focus is to understand the interaction with the customer
  - How the search keywords (and advertisements) influence the search process so that the business can have more visitors (potential customers).
  - Understand (measure) the user experience, behavior and satisfaction with the web site.
- Needed to achieve this
  - Infrastructure to collect and process data (technology)
  - Skills to analyze and interpret that data (qualified people)
- Common misunderstanding of comprehensive web analytics program
  - Focusing on collecting and reporting raw data without understanding their significance
- Skills needed: understanding statistics, mathematics.
  - Recently attempts to create tools to compensate for these skills.
- Fluidic and evolving subject in nature, roles are still being defined
  - Who is responsible for web analytics? Marketing or IT?

Web Analytics Overview

- Goal - Illustrate how to establish a comprehensive analytics program.
- Good tools are needed for any implementation of analytics
- Different tools is needed for different types of businesses.

- Brief History of Web Analytics
- Every time someone locates a website the web server logs data
- Before 1995, simple reports based on information that is automatically collected
  - filename
  - time, referrer (i.e., the website forwarding the request)
  - browser
  - operating system
  - computer data
- Later hit counter introduced on websites, but not accurate so they vanished
- WebTrends usually associated with first commercial web analytics programs

Example 1

- To better target, acquire, and retain customers, marketers need to use data analytics, content marketing, and customer engagement.

- For example an office furniture retailer could increase its return on what it spends for advertising by use of data analytics in the following way.

- To acquire new customers, the retailer has to figure out better ways to find potential clients to target for its ads. To do that, however, the company needs huge amounts of **intent**—or **in-market**—data.

- **Intent data** is data collected about online users’ activities—indicating some future action, or intent, such as ordering a product. This can be achieved by obtaining web site's key performance indicators using web (or free Google) analytics as described in the first part of this semester class notes.

- When potential customers is looking to furnish a new office may interact with the retailer's web site in a variety of ways, such as browse through architectural sites for design ideas, visit various office retail sites to evaluate items or perhaps even do some comparative analysis on a product review site. All these actions signal that the consumer is actively browsing, researching or comparing the types of products online furniture retailers sell. Customer's intents, preferences, and loyalties create impressions and Web/Google analytics tools can then capture that data enabling marketers to act on it.

Example 2

- ENERGY STAR, a score from 0 to 100, is a measure of energy consumption performance. A score of 40 means performance worse than 40% of similar buildings nationwide. A score of 75 or higher makes you eligible for ENERGY STAR certification.

- Energy Management Associates, Inc. Help customers achieve ENERGY STAR certification rating by reviewing customer's portfolio, and providing calls with questions or suggestions, to raise customer's score and provide up to 20% savings opportunities.

- According to EAM's web site, customers are encouraged to provide a copy of their most recent utility invoice (12 to 14 month usage history) for each account (gas and electric). This data is contained in EMA's database that with help of **predictive data analytics**, can be used to enhance the detailed customer profile and predict the most likely factors affecting the ENERGY STAR score, by integrating a automatic analysis of the key performance indicators with historical data, by region, type of energy etc. and tied together with external data available on the web, such as location specific temperature and weather conditions.

Example 3

- The use of **predictive data analytics** enables the retailer, insurance company, the travel agency or any similar eCommerce (click to order) business to predict the most likely purchase type from a given customer.

- This is typically achieved by leveraging historical data (sometimes tied together with a third-party data) to paint a detailed picture of the buyer personas. From that information, it can be determined:
  - Which customer will purchase
  - What product they will purchase
  - What message they will respond to
  - Which customers will focus on

- Predictive data analytics can recognize patterns and behaviors for more effective
  - messaging
  - plan persona-specific campaigns based on customer's habits and past preferences.

Example 4

- Data analytics is also commonly used to:
  - Give real-time visibility of inventory and warehouse management systems.
  - Sifting through an abundance of social media information. Monitoring, analyzing and reporting on the voice of its customers on social media including Facebook, Twitter, and various blogs and forums.
  - To more accurately analyze the voice of the customer, predictive analytics (and NLP - Natural Language Processing) can be used to score data attributes and determine which social media posts are actionable and relevant thus filtering out the noise of irrelevant posts.

Brief History of Web Analytics

- Several years later, introduced website analytics software that was able to measure click density or site overlay and heat maps.
  - Enabled to understand exactly, which links and where on a page, visitors were clicking.
  - Reports could described the consumer behavior by reporting number of clicks or overall percentage for the identified web pages.
- Big web analytics vendors:
  - Google, WebTrends, Coremetrics, Omniture, WebSideStory, IBM and Adobe.
  - Please have a look at SiteCatalyst a tool from Omniture (Adobe marketing cloud) to see the features and functionalities that can be found in modern tools.
- Ever changing and growing industry, gained much more popularity recently with Big Data.

Addressing the “What” question

- Large amount of data collected by vendors from every web site visit and click – **clickstream**.
  - It describes what data is collected.
  - It does not contain any insight into the significance of the data.
  - Too large to be useful.

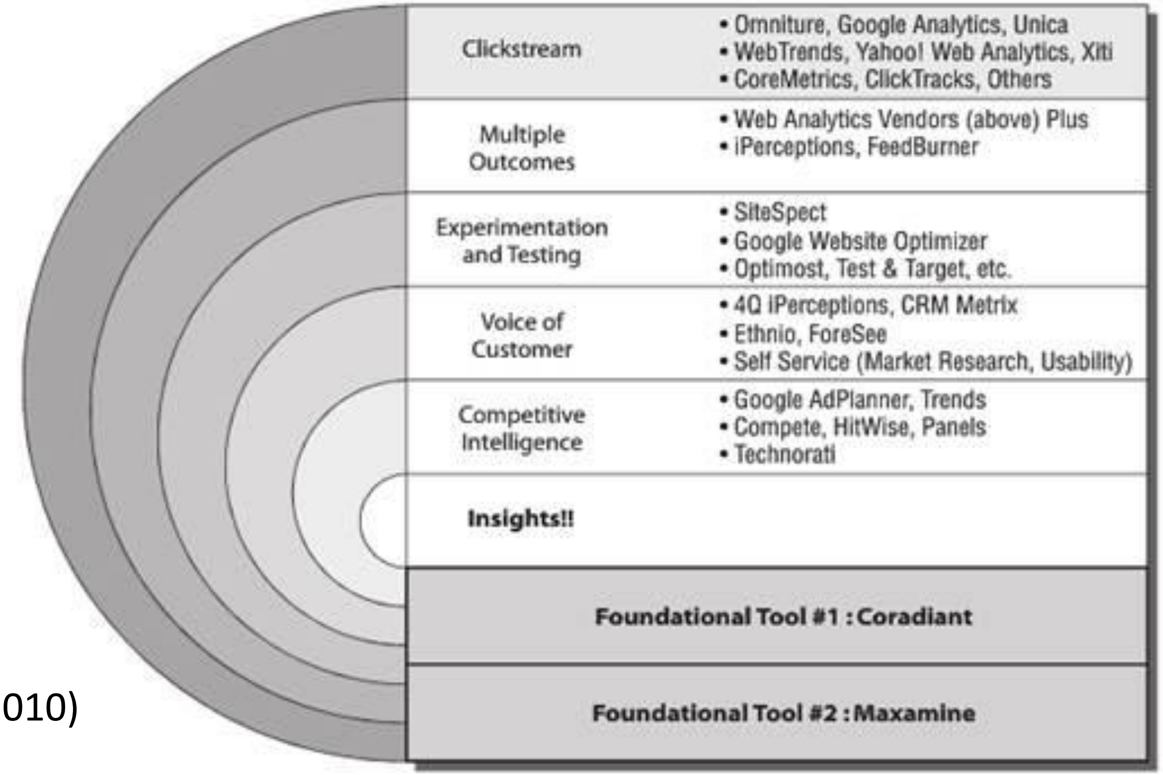
- Important “What happened” questions to address
  - What pages did people view on your website?
  - What products did people purchase?
  - What was the average time spent?
  - What sources did they come from?
  - What keywords are campaigns produced clicks?

Addressing the “Why” question

- Even more critical is to know “why” people do the things they do on your web site
  - This can be used to make intelligent decisions about your web presence.
- Important “why” questions to address
  - Analyze qualitative and quantitative data from our web site and a competitor’s web site
  - Focus on continual improvement of the customer’s online experience
  - Better translate data into desired outcomes, both on- and off- line
  - Decide whether we should go with channels like TV and radio advertising, rather than online advertising
  - Obtain a strategic advantage over competitors who focus only on clickstream
- Examples:
  - dropped shopping carts or
  - dropped registrations forms

Web Analytics Tools

- Which web analytics tools to use?
- Depends on the size of your business and the available resources
  - Small Business – Click stream, outcomes, voice of the customer
  - Medium Business – All of the above plus Testing
  - Large Business – All of the above plus Competitive intelligence, Deep back-end analysis, Site structure gaps



Clickstream	• Omniture, Google Analytics, Unica • WebTrends, Yahoo! Web Analytics, Xiti • CoreMetrics, ClickTracks, Others
Multiple Outcomes	• Web Analytics Vendors (above) Plus • iPerceptions, FeedBurner
Experimentation and Testing	• SiteSpect • Google Website Optimizer • Optimost, Test & Target, etc.
Voice of Customer	• 4Q iPerceptions, CRM Metrix • Ethnio, ForeSee • Self Service (Market Research, Usability)
Competitive Intelligence	• Google AdPlanner, Trends • Compete, HitWise, Panels • Technorati
Insights!!	

Foundational Tool #1 : Coradiant

Foundational Tool #2 : Maxamine

From “Strategy and Tools”
Webanalytics 2.0 by Kaushik (Wiley 2010)

Goals, Context,

- It is Important to know
  - Goals of Analysis
    - Target that lets us measure success
    - Assumes understanding both the “what and the why”
    - Examples:
      - Knowing where visitors come from.
      - Understanding what they did in the website.
      - Did they meet the goal?
      - Did the consumers respond to the online marketing plan in the manner you expected?
      - Does the data analytics tool validate that?
  - Context Study
    - Website visitors are made up of groups of people behave differently and have different objectives.
    - Their action or behavior can be analyzed with the web analytics tools.

and Segments

- Segment (determine as much as possible features for classification)
  - Analyze particular groups (segments) of people that come to the website.
  - What this particular group has in common with other groups.
  - Relate visitors by geographical regions (east/west coast, Europe, Asia etc.)
  - Relate visitors by gender and their preferences.
  - Segment data by visitor’s motivation.
  - Which group of people “browsed, shopped and purchased”.
  - Was this the group that you sent out an email marketing outreach newsletter?
  - If so do we need to do more outreach via e-newsletters.

Defining Basic Analytics Metrics

- In analytics it is essential to know which numbers are important and why.
- Note that the metric focuses on “Why”.

- Basic Analytics Metrics
  - Visits and Visitor Sessions
  - Referrals
  - Bounce & Exit Rate
  - Conversion Rate
  - Engagement
  - SEO, Social Media, Emails and Metrics

Visits and Visitor Sessions

- **Visitor** - an individual (not necessarily a human) or device such as browser which accesses a Web site within a specific time period.
  - Unique visitor within a specific reporting period (no double counting).
- **Visit** (**Sessions**) – an interaction with a data source (examp: text and/or graphics downloads) from a single browser (device) during a single session.
  - A visit can consist of a series of page views that a single visitor makes during a period of browsing activity. A visit ends after the visitor closes the browser, clears cookies, or is inactive for 30 minutes (customizable time period).
  - During each visit, users will engage in one or more interactions with the web site pages.
  - Analytics software will automatically track these interactions as “**pageviews**.” The pageview metric increases every time a page is viewed on your site. Other activity, like watching a video, mouse position, etc. can also be tracked. Such activities are better classified as “**events**” rather than pageviews.
  - **Cookie** (persistent or session) is a file on the user’s device that identifies the user’s unique browser.
  - Tracking code looks for cookies. If a cookie is deleted or blocked incorrectly counts unique visitors.

Referrals - Where do visitors come from?

- **Referrals** indicate the place from which the user clicked to get to the current page.
- It is a valuable to know how someone found our web site. Was it
  - through a search engine
  - positive review
  - social-media talk
  - email or e-newsletters
- Referrals are the lifeline for marketing advertisements.
- It's important to know which campaigns helped draw in new visitors or succeeded in getting loyal customers.

Bounce and Exit Rate

- **Bounce rate** and the exit rate measure whether users find a web site or a web page useful.
- Bounces are counted for users who land on a page and leave immediately. They do not see the page content.
- Reasons can be site-design or usability issues or many other reasons.
- Typically expressed as percentage of single-page sessions.

- If the exit rate is high, the exit-rate metric can be meaningless, and it should not matter.
- Useful to find out if visitors are abandoning the site at a certain point in the middle of an e-commerce transaction.

Conversion Rate

- In the context of studying goals and outcomes, this metrics is a significant one.
- The **conversion rate** (as a percentage), is defined as

  conversion rate = **outcome** /unique visits * 100
- An example of an **outcome** could be something very simple like clicking on ads or coupons or subscribing to a newsletter.
- Should we use Unique Visitors (browsers) or Visits?
- Common for a unique visitor to visit the same page many times (purchase).
  - Can you think of an example?
- In this context conversion rate measures the process of converting a visitor into a buyer.
- Conversion rate can be calculated automatically by integrating analytics software with shopping carts.

Engagement

- This is a qualitative metric hard to measure. The definition can be fuzzy.
- Most analytics software will track event and visit duration.
- This does not provide any information about the quality of engagement during that visit.
- What it matters is the time a visitor spends on a web site **with** engagement.
- The challenge is to distinguish between
  - Reading the information on a web page or
  - Looking for the information and not being able to find it.
- Example: Google Analytics tools allow us to research engagement
  - In-Page Analytics (visual assessment of how users interact with your web pages)
  - Behavior Flow analytics (visualizes the path users traveled from one page or event to the next).
- Both of these contribute to Engagement statistics.
- More on Google Analytics in the next session.