Click stream reporting & analysis for website optimization

Richard Doherty
e-intelligence Program Manager
SAS Institute EMEA
What is Click Stream Reporting?

- Potential customers, or *visitors*, navigate the Web by selecting a sequence of Web pages.
- Each time a page is requested, an entry is made in the Web log.
- Analysis of data stored in Web logs is often referred to as *click stream reporting*.
What is Click Stream Analysis?

- Using Web logs, it is possible to identify a visitor’s path through the site.
- Visitors often request several pages before leaving the site, ending the visit.
- Analysis of the pages accessed during a visit is often referred to as *click stream or path analysis*. 
Why bother with Click Stream Reporting & Analysis?

- Companies want to enhance their understanding of how users are navigating their web site in order to:
  - increase conversion rate - browsers to buyers
  - reduce abandonment - half full shopping carts!
  - enable personalisation
  - improve web site design
  - increase ease of use and navigation
Why bother with Click Stream Reporting & Analysis?

- On-line businesses want to be like the corner shop used to be
- Understand their customers
- Target them with useful/relevant content
- Build up a mutually beneficial relationship
- Keep the customer coming back
Data formats

- Traditional
  - Common log format (CLF)
  - Extended log format (ELF)
  - Internet Information Server (IIS) format

- Custom
  - Dynamic HTML
  - Java server pages (JSP) & servlets
  - Content servers
  - Packet sniffers
Overview of Data Collection on Web Visits
The SAS Solution for e-Intelligence includes generic readers for:
- Microsoft IIS.
- CLF (Common Log Format).
- ELF (Extended Log Format).
- Custom logs
  - Becoming increasingly popular as data volumes increase.
- A quickstart visual interface eases loading of these e-data sources.
What does the data look like?

- Client IP.
- Identification field.
- Authenticated User Id.
- Date/Time and GMT.
- HTTP Request.
- Status Code Field.
What does the data look like?

- Transfer volume.
- Error log.
- Referrer log e.g. http://search.yahoo.com/search?p=data+mining+Websites
- Agent log e.g. "Mozilla/4.61 [en} (WinNT;I)"
Some Terminology

- **Hits**
  Each request for content is a hit. The HTML page, any GIF references….

- **Visits/Sessions:**
  A distinct user who browses the web site within some defined time period.
  
  A session ends after 30 minutes of inactivity.
  
  Can have multiple sessions per day.
Typical reporting requirements

- Produce frequencies on accessed Pages: Month, week, day, hour.
- Produces Frequencies on Entry points and Exit Points.
- Produces Frequencies on Browsers used.
- Top 10 pages, Bottom 10 pages.
- Etc..
Click Stream Reporting using the SAS Solution for e-Intelligence

SAS Institute's web tracking and analysis tool
Click Stream Analysis

- Consists of hyperlinked set of pages. Where the nodes are web pages and edges are hyperlinks.

This structure could also be viewed using the tree lens viewer - coming soon from SAS Institute
Visitor steps in a website

- A forward Step: A to B
- A backward Step: D to C
- A jump: H to E
The log files

- Log servers record data on accessed pages and in the accessed order.
- Several to many web servers for a single web site.
- Large number of users.
- Files can become very, very large.
Steps involved for Path analysis.

- Merging all Web logs together.
  - Possible synchronization problems
  - A visit might span multiple servers
- Cleaning web logs.
- Identify accessed page for each visitor.
- Determining user sessions for each visitor.
- Identify frequently visited paths per sessions or across sessions.
Association Analysis
(a.k.a. Market Basket Analysis)

- Association analysis is a descriptive technique that determines how often events occur with other events.
- Stores use this type of analysis to determine which products were purchased together, hence the name *market basket analysis*. 
Questions addressed by Association Analysis

- A store may want to know things such as
  - What proportion of the people who purchase eggs also purchase milk?
  - What proportion of the people who purchase eggs and milk also purchase bread?

- An online vendor may want to know things like
  - What proportion of people who select a particular Web page also select a different Web page?
Sequence Analysis

- Sequence analysis is often considered a special case of association analysis.
- Sequence analysis answers the same types of questions answered by association analysis, but it takes into account the order of the purchases.
Questions addressed by Sequence Analysis

- A store may want to know things such as
  - What proportion of the people who purchase eggs *later* purchase milk?
  - What proportion of the people who purchase eggs and milk *later* purchase bread?
- An online vendor may want to know things like
  - What proportion of people who select a particular Web page *later* select a different Web page?
Path analysis issues

- The Web is stateless
  There is no connection or relationship from one request to another request.
- User identification:
  Many IP or domain name per user and many users per IP.
- Local and proxy Caching
User identification

- Require a login to access site
- Serving content via:
  - CGI
  - servlets
- Cookies.
Cookies

- **What is a cookie?**
  A text file that is server specific and stored on the local users machine.

- **How do they work?**
  Initially sent to browsers by the server. The browser stores the text file. Unique cookie id added to each web log record.

- **Can be created using client-side (JavaScript) or server-side (CGI/servlet) tools.**
Cookies - the bad

- Perception...Cookies are not inherently bad.
  - Cookies are being used by some companies to track web surfer patterns.
- Some users elect to disable cookies.
- Cookies can be deleted.
- Older browsers may not support cookie usage.
- Cookies are not secure. Sensitive information should never be stored in a cookie.
- Mobile devices
A session is a mechanism used to maintain state about a series of requests from the same user (request originating from the same browser) across some period of time.

- Can provide an alternative means to cookies to capture user information.
- Doesn’t suffer from size and type (simple text), restrictions of cookies.
The SAS Solution for e-intelligence

- Utilises the award winning Enterprise Miner software
  - Sequences
  - Associations
  - Visual exploration
- See case study presentations for more details
Next steps

To add real value

- warehouse the e-data
- combine with other sources of information in the organisation
- control/document with SAS/Warehouse Administrator
- create knowledge using OLAP, Q&R, data mining, etc.

What would this look like?
Next steps

- e-intelligence knowledge can feed other projects/initiatives
  - CRM
  - SRM
  - BSC
  - ……
In Summary

- Click Stream reporting & analysis:
  - Challenging.
  - Complex.
  - Analytical.
  - High volumes.
The SAS Solution for e-Intelligence

- highly scalable
- industry leading analytics
- industry leading data mining
- automated
- data warehouse centric
- web enabled
- customisable
- 24 years of experience and R&D
THE END