

From Server Logs to Query Logs

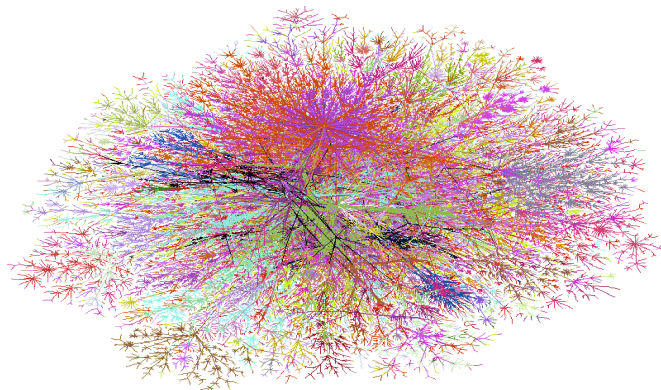
Professor Mark Levene
Information Management and
Web Technologies Research Group

27 May 2009

School of Computer Science
& Information Systems

londonknowledgelab

Observation 1: **Complexity**
The Web is a Complex Network
(Map of the Internet Bell Labs 1998)



27 May 2009

School of Computer Science
& Information Systems

londonknowledgelab

Observation 2: Too much Data

We need to make sense of practically an infinite amount of information

- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:00:48:40 +0100] "GET /~/mark/handheld.html HTTP/1.0" 200 1730 "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:00:49:16 +0100] "GET /~/mark/games.html HTTP/1.0" 200 6582 "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:01:02:21 +0100] "GET /~/mark/bookshops.html HTTP/1.0" 200 3568 "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:01:41:04 +0100] "GET /~/mark/web.html HTTP/1.0" 200 14639 "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:01:42:17 +0100] "GET /~/mark/download/optdb_plan.pdf HTTP/1.0" 304 - "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- ip68-98-199-25.mc.at.cox.net - [21/Sep/2003:02:10:27 +0100] "GET /~/mark/download/optdb_integrity_constraints.pdf HTTP/1.0" 200 32768 "http://search.yahoo.com/search?p=definition+of+superkeys&ei=UTF-8&fr=fp-top&n=20&fl=0&x=wrt" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
- ip68-98-199-25.mc.at.cox.net - [21/Sep/2003:02:10:28 +0100] "GET /~/mark/download/optdb_integrity_constraints.pdf HTTP/1.0" 206 158146 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
- adsl-68-74-73-241.dsl.emhrl.ameritech.net - [21/Sep/2003:02:39:29 +0100] "GET /~/mark/book.html HTTP/1.1" 200 3373 "http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=relational+databases+basic" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
- adsl-68-74-73-241.dsl.emhrl.ameritech.net - [21/Sep/2003:02:39:30 +0100] "GET /~/mark/front_cover.gif HTTP/1.1" 200 64168 "http://www.dcs.bbk.ac.uk/~mark/book.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
- crawler4.googlebot.com - [21/Sep/2003:03:35:52 +0100] "GET /~/mark/games.html HTTP/1.0" 200 6582 "-" "Googlebot/2.1 (+http://www.googlebot.com/bot.html)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:04:15:59 +0100] "GET /~/mark/download/optdb_table.pdf HTTP/1.0" 304 - "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- drone10.sv.av.com - [21/Sep/2003:04:47:09 +0100] "GET /~/mark/ HTTP/1.0" 200 5183 "-" "Scooter/3.3_SF"
- crawler4.googlebot.com - [21/Sep/2003:04:49:22 +0100] "GET /~/mark HTTP/1.0" 301 309 "-" "Googlebot/2.1 (+http://www.googlebot.com/bot.html)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:05:18:46 +0100] "GET /~/mark/optdb_mailing_list.html HTTP/1.0" 200 622 "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"
- pool-68-162-19-184.nwrk.east.verizon.net - [21/Sep/2003:05:35:01 +0100] "GET /~/mark/download/optdb_erd.pdf HTTP/1.1" 200 0 "http://www.google.com/search?ie=entity+relationship+concept&hl=zh-TW&lr=&ie=UTF-8&oe=UTF-8&start=10&sa=N" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; YComp 5.0.2.6)"
- pool-68-162-19-184.nwrk.east.verizon.net - [21/Sep/2003:05:35:02 +0100] "GET /~/mark/download/optdb_erd.pdf HTTP/1.1" 206 0 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; YComp 5.0.2.6)"
- pool-68-162-19-184.nwrk.east.verizon.net - [21/Sep/2003:05:35:07 +0100] "GET /~/mark/download/optdb_erd.pdf HTTP/1.1" 206 275480 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; YComp 5.0.2.6)"
- crawler4.googlebot.com - [21/Sep/2003:05:49:47 +0100] "GET /~/mark/ HTTP/1.0" 200 5183 "-" "Googlebot/2.1 (+http://www.googlebot.com/bot.html)"
- cr008r01-3.sac2.fastsearch.net - [21/Sep/2003:06:14:12 +0100] "GET /~/mark/download/webgraph.pdf HTTP/1.0" 304 - "-" "FAST-WebCrawler/3.8 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)"

27 May 2009

School of Computer Science & Information Systems

londonknowledgelab

Observation 3: Wireless Mobility

Wherever you are, if you have a mobile wireless device your activities are logged

(Map of mobile phone base stations in WC1E 7HX from www.sitfinder.ofcom.org.uk)



27 May 2009

School of Computer Science & Information Systems

londonknowledgelab


Observation 4: Search

Whenever you use a search engine your searches are being logged

Web Images Video Maps News Shopping Mail more ▾ [Sign in](#)

Google 5 Southampton Street, London, WC2E 7HA [Advanced Search](#) [Features](#)

Web [Show options...](#) Results 1 - 10 of about 14,100 for 5 Southampton Street, London, WC2E 7HA. (0.32 seconds)

 [5 Southampton St London WC2E 7HA, UK](#)
maps.google.com
Start address
 Remember this location

[Sponsored Links](#)
London Street View
Explore London neighbourhoods with street level imagery. Try it! maps.google.co.uk

[Southampton Street, London](#) borough of Camden
Address: Davidson Building, 5 Southampton Street London WC2E 7HA View LECG's profile
- Leonardo Media Marketing Consultants & Services (Head Office) based ...
[www.londonline.co.uk/area/Southampton_Street_WC2E/](#) - 23k - [Cached](#) - [Similar pages](#)

[The Jubilee Hall](#) [Southampton Street](#) [London](#) [WC2E 7HA](#)
Tel: +44 (0) 20 7240 7405; Fax: Email: Information; Location: Southampton Street, Covent Garden, London, WC2E 7HA. Nearest station: Covent Garden, London ...
[www.londonnet.co.uk/listings/shopsamenities/exhibitioncentres/thejubileehallincoventgarden/](#) - 37k - [Cached](#) - [Similar pages](#)

[Cotswold Outdoor](#) - 8 [Southampton Street](#) [London](#) [WC2E 7HA](#)
Tel: +44 (0) 20 7379 3060; Fax: Email: Information; Click here for more information; Location: 8 Southampton Street, Covent Garden, London, WC2E 7HA ...
[www.londonnet.co.uk/listings/shopsamenities/campingequipmentsupplies/cotswoldoutdoorincoventgarden/](#) - 37k - [Cached](#) - [Similar pages](#)

[Maps to our offices: Contact & Help - BCS](#)
London Office BCS First Floor The Davidson Building, 5 Southampton Street London, WC2E 7HA. Contact BCS - Map and directions (PDF) ...
[www.bcs.org/maps](#) - 10k - [Cached](#) - [Similar pages](#)

27 May 2009

School of Computer Science
& Information Systems



Log Analysis/Web Analytics - Applications

- Prediction (Recommender systems)
- Personalisation (Collaborative Filtering)
- Social search
- Topical search
- Search engine advertising
- Clickstream analysis, e.g. e-Commrece
- Prefetching and caching of SERPs and web pages

27 May 2009

School of Computer Science
& Information Systems



Availability of Log Data

- Publically available log data
- Limited distribution log data
- Log files may not be enough – demographic and other external user profile data is useful
- There is a lack of recent log data for the research community (the AOL search log fiasco raised some important issues)
- There is an issue with verifiability and repeatability of experiments, where the access to the data is limited

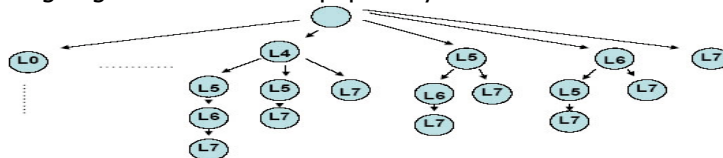
27 May 2009

School of Computer Science
& Information Systems

londonknowledgelab

Server Log Data Analysis (with Jose Borges)

- Variable Length Markov Chain model.
- Use suffix to represent trails.
- Try and predict next link the user will follow from the longest suffix of a trail that can be matched in the suffix tree.
- Evaluate with hit and miss or average rank of link followed.
- Can determine order of model (maximum length of trails).
- Can determine window size to avoid concept drift.
- Ongoing work to combine popularity with time and content.



27 May 2009

School of Computer Science
& Information Systems

londonknowledgelab

Context-Topic Association Rules Discovered from Search Engine Logs (with Carlos Hurtado)

- We assume that the *topic* of a query can be determined from its terms: e.g. chat, cars, "going out", lottery, jobs.
- The *context* is a set of features that can be extracted from the log: e.g. date, time, IP.
- *Given a query topic, what are the "interesting" contexts for that topic?*
- Concentrate on temporal contexts: *dayOfWeek* and *hourOfDay*
- C-T rule format:
 $\{\text{Fri,Sat}\},\{19,20,21\} \rightarrow \text{"going out"}$

Semantics of C-T Rules

- Estimate the probability of the topic given the context $P(t/c)$.
- Estimate the probability of the topic $P(t)$.
- Estimate a lower percentage limit confidence interval for the ratio $P(t/c) / P(t)$ at a given level, say β .
- The rule is (ρ, β) *interesting* if the lower limit at level β is greater than ρ .
- Some (1,70%) interesting rules from 2000 manually classified queries from TodoCl (2004).

<i>dayOfWeek</i>	<i>hourOfDay</i>	<i>Topic</i>
<i>Saturday</i>	0	<i>chat</i>
<i>Sunday</i>	1	<i>chat</i>
<i>Friday, Saturday</i>	20	<i>goingOut</i>
<i>Monday</i>	8, 9	<i>gamesOfChance</i>
<i>Sunday</i>	22	<i>gamesOfChance</i>
<i>Monday</i>	11	<i>jobs</i>
<i>Tuesday</i>	9	<i>health</i>
<i>Monday, Tuesday</i>	20	<i>culture</i>
<i>Sunday</i>	12	<i>culture</i>
<i>Tuesday</i>	20, 21	<i>culture</i>
<i>Wednesday</i>	19	<i>culture</i>

Query Classification (with J. Bar-Ilan, I. Cox and Z. Zhu)

- Use a searcher's ontology of 30 top-level categories.
- It's a multi-class, multi-label problem.
- Difficult problem as queries are short, noisy and ambiguous.
- Queries need to be enriched with terms from:
 - *Search engine result web pages or snippets.*
 - Related queries (e.g. Yahoo Explore Concepts).
 - Training set of categorised web pages.
- We use Bigram SVM, with balancing of small classes.

27 May 2009

Experiment with AOL Query Log

- Test set of about 33k manually classified queries, 10k from a Masters class assignment and 23K from AOL research (2006).
- Evaluate using Micro-F1.
- Best results training and testing (10-fold cross validation), with top-10 Google snippets (10g).

TE	TR	recall	precision	F1
2g	2g	47.22%	47.46%	47.32%
5g	5g	51.42%	57.06%	54.09%
10g	10g	52.57%	59.70%	55.91%
20g	20g	51.65%	60.62%	55.78%

TE	TR	recall	precision	F1
q	q	53.78%	17.01%	25.83%
e	e	50.81%	38.14%	43.56%
2g	2g	47.22%	47.46%	47.32%

27 May 2009

Topical Analysis of An MSN Query Log

- We use the classifier to categorise 15 million queries (2006).
- To test the quality of the classification we asked 30 users, each to assess 20 sampled queries from each large class and 10 from small classes.; "Y" meant the class was reasonable and "N" not.
- Precision, recall and F1 were consistent at about 75%, with an error of about 5% due to the sample size.

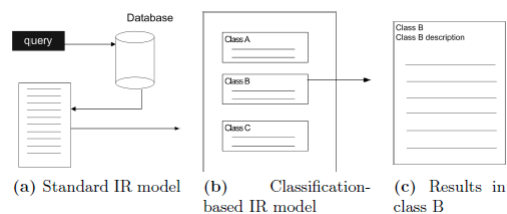
27 May 2009

School of Computer Science
& Information Systems

londonknowledgelab

Classification-Based Search

- With the aid of snippets from SERPs a query classifier can help categorise the results.
- Such a strategy can be useful when
 - The query is ambiguous..
 - The user knows which class he/she is interested in.



27 May 2009

School of Computer Science
& Information Systems

londonknowledgelab

Wireless Usage Log Data

- Ongoing work:
 - Prediction – where will the user go next?
 - Navigation patterns – can we classify users according to their behaviour?
 - Visual analytics of movement and social interaction.
- Can mobile search logs be made available?

27 May 2009

School of Computer Science
& Information Systems

 londonknowledgelab

The Many Facets of Log Analysis

- There are many other sources of web log data that can help us understand how the web is used (e.g. social networks).
- Given enough data we can leverage the analysis to improve the user experience (e.g. recommendation) and the delivery of content (e.g. advertising).
- But without log data we cannot do data mining!

27 May 2009

School of Computer Science
& Information Systems

 londonknowledgelab