Modelling human behaviour on the web

Nikolaj Tatti
Department of Computer Science
Helsinki University of Technology
ntatti@cc.hut.fi

November 2003

What about people?

- So far we have studied how the web may look like or how search engines work.
- The web is used (mostly) by people. It would be nice to study their navigational behaviour and try to build some models explaining their actions.
- Also it would be nice to model search engine queries.

Menu:
Collecting data
Modelling the navigational behaviour
Understanding the search queries

Collecting data

- There are two different ways of collecting navigational data of user.
- Server-side data: we examine the web server page request logs.
- Client-side data: we monitor the user's computer.
- Server-side is much easier to get, but it has to be analysed much carefully.

Server-side data

- Web server log contains human and robot page requests. These must be separated from each other.
- We can examine the behaviour of the user. For example, if the user asks pages too fast, we can assume that the user is a robot and delete all its page requests from our data.
- Page caching: The browser and the proxy server cache pages, so not all page requests are seen in logs.
- There can be multiple users browsing from the same IP address. This can be solved by using some other identification methods than IP address.

Client-side data

- Much more reliable than server-side data.
- Other events than page requests are retrieved also. For example, 'back button' usage or scrolling.
- Much harder to get. Users must be asked for a permission.

Empirical tests

- There has been a variety of empirical studies how users use the web.
- However, there can be biases in results since the test subjects are usually computer science faculty staff or graduate students (definitely abnormal people).
- Among early studies most cited are Catledge and Pitkow (1995) and Taushcer and Greenberg (1997).
- The more recent study is Cockburn and McKenzie (2002)

Early studies

- Catledge and Pitkow collected on 107 users over three weeks in 1994. Totally, there were 31134 navigation commands (back-button usage, bookmarking..) and 14 page requests per user per day.
- Tauscher and Greenberg had 23 subjects over a six-week period in 1995. Data consisted of 19000 navigation commands and about 21 page requests per user per day.
- Both studies showed that clicking anchor-links was the most common web browsing action (50 %).
- The second most common action was the usage of 'back button' (41 % in CP and 30 % in TG).

Early studies

- Tauscher and Greenberg analysed how often a page is revisited. The probability that page is revisited was 0.58 for 1995 data.
- Tauscher and Greenberg also re-analysed a subset of 1994 data and estimate the revisitation probability to be 0.61.
- *Recency* effect: Page is revisited more probably if it has been visited very recently. This copes with the usage of 'back button'.

The Cockburn and McKenzie study

- Cockburn and McKenzie analysed *history.dat* files produced by Netscape browser for 17 users between October 1999 and January 2000. The subject were (again) faculty, staff and graduate students.
- There were 42 page requests per user per day. This is much higher than in early studies.
- The revisitation probability was estimated to be 0.81. This is also higher than in early studies.
- The usage of the web has evolved from an exploratory mode to a utilitarian mode. For example, there are pages (www.helsinginsanomat.fi or www.dilbert.com) which are visited daily by some particular user.

Video-based analysis of Web usage

- There has been studies where users browsing were video-taped.
- For, example Byrne *et al.* (1999) analysed video-taped recordings of eight different users.
- A lot of time is spent scrolling pages (40 min out of 5 h).
- Also, a lot of time is spent waiting for pages to load (50 min out of 5 h).

Modelling browsing behaviour

- Assume that you have several session data sets. One session data consists of a sequence of pages requested by user. For example, $\{ABABCD, BCAACBA\}$, where different letters represent different pages.
- Combine all sessions in one big sequence by adding a special symbol 'E' in the of the sequences. For example, $\{ABABCD, BCAACBA\} \rightarrow ABABCDEBCAACBAE$
- The symbol 'E' represents the end of a session.
- Model this sequence using k-order Markov chains.

Markov chains

- When using Markov chains it is assumed that the probability of the following page in the sequence s_t depends only on k previous pages.

$$p(s_t \mid \cdot) = p(s_t \mid s_{t-1}, \cdots, s_{t-k}).$$

- If there are M possible symbols in the sequence ($M-1$ different pages and a symbol 'E'), then there are M^{k+1} parameters in a k-order Markov chain.
- Let θ_{ij} be the probability of a symbol i occurring immediately after a subsequence j of length k.
- Let n_{ij} be the number of times a symbol i occurring immediately after a subsequence j of length k in the data set.

Markov chains - continues

- The likelihood $\mathcal{L}(\theta)$ is equal to

$$\mathcal{L}(\theta) = \prod_{i,j} \theta_{ij}^{n_{ij}}.$$

- The log-likelihood is equal to $l(\theta) = \sum_{i,j} n_{ij} \log \theta_{ij}$.
- The ML estimation is equal to

$$\theta_{ij}^{ML} = \frac{n_{ij}}{n_j},$$

where n_j ensures that $\sum_i \theta_{ij}^{ML} = 1$. Thus $n_j = \sum_i n_{ij}$.

Markov chains - extensions

- Cadez *et al.* (2003) and Hansen (2003) proposed mixtures of Markov chains.
- The probability of a symbol is a mixture of N first-order Markov chains

$$p(s_t \mid \cdot) = \sum_{k=1}^{N} p(s_t \mid s_{t-1}, c = k) P(c = k),$$

  where c denotes the mixture component and k runs over all mixture components.
- The parameters of this model can be estimated using EM algorithm.

Modelling runlengths within states

- If the symbols in the data sequence represent pages, then the transition probability to the same symbol should be zero (there is little sense linking to the same page).
- On the other hand, if the symbols represent web servers, then there are positive transition probability T_i to the same symbol.
- The probability of the runlength r for state i is $P_i(r) = T_i^r(1 - T_i)$, that is, the probability that we stay in the state i for r steps.
- This is a geometric distribution having mode at 1 and mean at $(1-T_i)^{-1}$.
- For example, news web sites may have mode at 2.
- This can be solved using semi-Markov models. At each state i a runlength is drawn from some probability $P_i(r)$ and after r time-steps some other state is picked.

Session lengths

- Session lengths seem to follow more or less power law.
- However, the Markov chain predicts that the distribution is geometric.
- Also it was shown by Huberman *et al.* (1998) that under some assumptions the length distribution is an inverse Gaussian.
- None of these works properly.

Search Engine Queries

- There have been several studies of search engine queries: Lau and Horvitz (1999), Silverstein *et al.* (1998), Spink *et al.* (2002) and Xie and O'Hallaron (2002).
- The engines examined in these studies were AltaVista, Excite and Vivisimo.

Search Engine Queries - some results

- The average number of terms in a query range from 2.2 to 2.6 across the studies.
- The mode for terms in a query were 2 in all studies.
- Most of the users didn't refine their search.
- There seems to be shift in the distribution of query topics.

Category	LH (1999)	XO (2002)
Adult content	16.7%	8.3%
Entertainment	20%	7%
Commerce, Travel,People,Places,Things	20%	45%