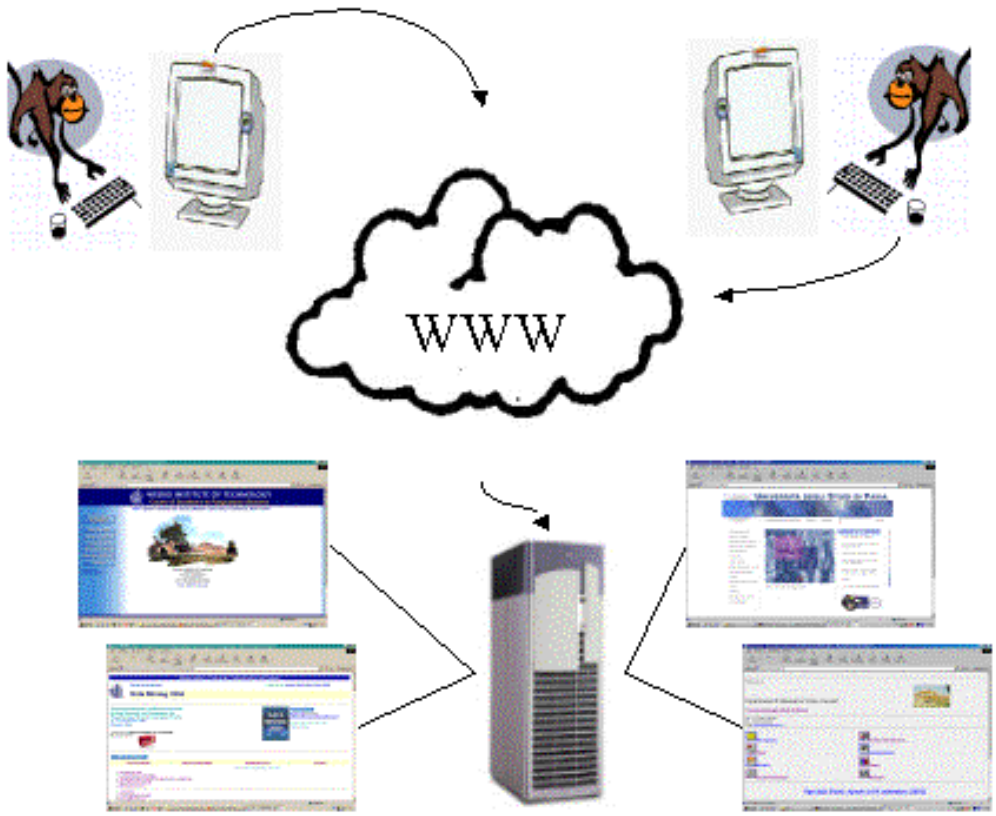
DATA MINING 2002
Bologna—Italy, September 25–27, 2002

Probabilistic Modelling for Clickstream Analysis

Lilla Di Scala & Luca La Rocca
University of Pavia—Italy

Site-centric clickstream data



Common Log Format includes:

- surfer's IP address
- URL of accessed file
- date of request

Data pre-processing

- pruning of irrelevant requests
  (e.g. files containing images)

- surfers' identification
  (e.g. by means of cookies)

- sessions' identification
  (e.g. based on inter-click times)

Surfer Identifier	Date of Request	Page Accessed
e74c4561...668c7fc5	NA	external
e74c4561...668c7fc5	08JUN98:11:30:07	home
e74c4561...668c7fc5	08JUN98:11:30:28	program
e74c4561...668c7fc5	08JUN98:11:33:14	login
e74c4561...668c7fc5	08JUN98:11:33:47	logpost
e74c4561...668c7fc5	NA	external
e708dc4e...b6d6f919	NA	external
e708dc4e...b6d6f919	17JUN98:16:58:07	program
e708dc4e...b6d6f919	17JUN98:16:59:53	addcart
e708dc4e...b6d6f919	17JUN98:17:00:48	product
e708dc4e...b6d6f919	17JUN98:17:02:01	freeze
e708dc4e...b6d6f919	NA	external
e708dc4e...b6d6f919	17JUN98:17:09:53	download
e708dc4e...b6d6f919	17JUN98:17:10:25	shelf
e708dc4e...b6d6f919	NA	external

A couple of interesting issues

- Which page will be requested next?
  (The answer can help develop an efficient server-side caching mechanism)

- How many different surfing styles?
  (The answer can help profile surfers, perhaps for marketing purposes)

$$* * *$$

We shall denote by

$$X_t^k$$

the t^{th} page visited by surfer k.

Modelling surfing behaviour

Surfing behaviour of surfer k described by

$$\mathcal{P}(X^k_{t+1} = i_{t+1} \,|\, X^k_t = i_t, \ldots, X_1 = i_1)$$

which we assume given by

$$P^c(i_{t-m+1}, \ldots, i_t; i_{t+1})$$

transition probabilities for surfers belonging to class c, whose members have memory m.

The resulting **likelihood** is

$$L(P, \pi) = \sum_c \pi_c \cdot \prod_t P^c(\ldots; i_t)$$

where π_c is the population-weight of class c.

Surfers' memory (I)

Within a single class of surfers, we focus our attention on determining their memory:

- $m = 0$ is **Independence (MC0)**
- $m = 1$ is **Markov Chain (MC1)**
- $m > 1$ is **High-order Chain (MC#)**

In the last case, parsimonious modelling is welcome and we suggest using the **MTDg** model (Raftery, 1985):

$$P(i_{t-m+1}, \ldots, i_t; i_{t+1}) = \\ = \sum_{l=1}^{m} \lambda_l Q_l(i_{t-l+1}; i_{t+1})$$

where $\lambda_l \geq 0$, $\sum_l \lambda_l = 1$ and each Q_l is a stochastic matrix.

This reduces to **MTD** when there is a single

$$Q = Q_l, \forall l$$

Surfers' memory (II)

- **MC#** models can be estimated by ML, that is by counting transitions.

- **MTDg** models can be estimated by ML, via numerical optimization: Matlab (R) code for Berchtold's (2001) algorithm is freely available on-line (Statlib SW Library, Carnagie Mellon University).

- **Best** model can be chosen by information criteria (models are not nested):

  $$BIC = -2 \log L(\hat{P}) + K \cdot \log N$$

  should be preferred to

  $$AIC = -2 \log L(\hat{P}) + K$$

  following Katz (1981); N is the number of likelihood components, K the actual number of transition probabilities.

Surfers' heterogeneity (I)

Setting $m = 1$, we concentrate on surfers' heterogeneity. Clickstream is thus modelled by a **finite mixture of Markov chains**.

Once the number of classes has been fixed, transition probabilities $P^c(i, j)$ and weights π_c can be estimated via the **EM algorithm**, which also inferentially classifies surfers.

Is an extra class needed?

- $\widehat{P}, \widehat{\pi}$ estimates
- $\widehat{\widehat{P}}, \widehat{\widehat{\pi}}$ estimates with one more class
- increased goodness-of-fit measured by

$$W = -2\left\{\log L\left(\widehat{P}, \widehat{\pi}\right) - \log L\left(\widehat{\widehat{P}}, \widehat{\widehat{\pi}}\right)\right\}$$

which is the **likelihood ratio statistic**

Surfers' heterogeneity (II)

Distribution of W under the null hypothesis?

- no asymptotic theory: null hypothesis lies on the boundary of parameter space
- following Aitkin et al. (1981), distribution is simulated under the fitted model
- we increase the number of classes until the null hypothesis cannot be rejected

Simulated surfers pay a **geometric number of sessions** to the site: come-back-to-site probability is estimated by ML from real data.

Results of analysis (I)

We analysed data from the log-files of an anonymous European e-commerce site.

Attention focused on a limited set of pages: sampling the chain avoids lumpability issues.

Memory

A reduced model is also considered, in which "structural zeroes" are not parameters.

Model	Weeks	(full)	Weeks	(reduced)
MC0	1	2%	0	0%
MC1	47	98%	27	56%
MC2	0	0%	3	6%
MTD2	0	0%	2	4%
MTD3	0	0%	2	4%
MTD4	0	0%	2	4%
MTDg2	0	0%	7	15%
MTDg3	0	0%	4	9%
MTDg4	0	0%	1	2%
Total	48	100%	48	100%

The first lag of memory is by far the most important, in accordance with the entropy-based results of Pirolli & Pitkow (1999).

Results of analysis (II)

The MTDg's superior predictive power can be worth its increased complexity, especially if prior topological information is available.

Heterogeneity

Month	Year	Week 1	Week 2	Week 3	Week 4
January	1998	7 / 1	18 / 1	42 / 2	47 / 1
February	1998	50 / 2	40 / 2	47 / 2	35 / 2
March	1998	32 / 2	54 / 2	57 / 1	39 / 2
April	1998	43 / 3	40 / 2	44 / 2	32 / 2
May	1998	30 / 2	30 / 2	25 / 3	31 / 3
June	1998	42 / 2	31 / 2	49 / 2	59 / 3
July	1998	39 / 2	59 / 1	59 / 2	52 / 2
August	1998	72 / 3	43 / 3	47 / 2	80 / 2
September	1998	69 / 3	82 / 3	80 / 2	70 / 2
October	1998	65 / 2	73 / 2	98 / 3	93 / 2
November	1998	79 / 3	74 / 2	81 / 3	63 / 3
December	1998	70 / 2	81 / 3	90 / 2	48 / 2
January	1999	53 / 2	73 / 2	74 / 3	85 / 4
February	1999	56 / 3	101 / 2	98 / 3	73 / 3
March	1999	86 / 4	98 / 3	112 / 3	95 / 2
April	1999	62 / 2	89 / 3	106 / 3	70 / 2
May	1999	75 / 3	98 / 2	94 / 2	81 / 3
June	1999	57 / 2	86 / 4	86 / 2	111 / 2

As time passes, the mean number of surfers per class roughly increases from 20 to 35.

Possible developments

- Letting $m > 1$, while considering more than one class, with m different for each class, might be a way to **classify surfers** (also) **according to their memory**.

- Since the **MTDg** model proved adequate for clickstream data, even if costly in terms of parameters, one could look for a more parsimonious version.