



# Adaptive Web Presence and Evolution through Web Log Analysis

Xueping Li

University of Tennessee

Co-author: Laigang Song





# Outline

---

- ✎ Background & motivation
- ✎ Data acquisition
- ✎ Data analysis
- ✎ Discovery of some invariants
- ✎ Concluding remarks
- ✎ Q/A

# Background

---

- ✎ WWW becomes a key revenue-generating channel
- ✎ Web presence is an effective way to
  - Enhance the image of a company
  - Increase the brand and product awareness
  - Provide customer services
  - Gather information
  - ...
- ✎ Many web sites remain unchanged for months or even years after they were initiated and developed
  - They lose their customers eventually
- ✎ How to evolve the web presence?



# A significant challenge

---

- ✎ WWW is inherently dynamic and web data is more sophisticated, diverse and dynamic
- ✎ Web mining is one method to gain insights
  - WSM
  - WCM
  - WUM
- ✎ This study presents a framework to establish adaptive web presence and evolution through web log analysis



# Web log data

---

- ✎ Web logs contain potentially useful information
  - Access logs contain the bulk of data including the date and time, users' IP address, requested URL, etc.
  - Agent logs provide the information of the users' browser type, browser version, and operating system.
  - Error logs provide the information of the problematic and erroneous links on the server.
  - Referrer logs provide information about what web pages from where contain the links to documents on the server
- ✎ This study adopts the classical data collected from the web server at the NASA's Kennedy Space Center

# Data format

✎ *A typical form of an access log of a request*

- *hostname - -  
[dd/Mon/yyyy:hh24:mm:ss tz] request status  
bytes*
- *E.g. uplherc.upl.com - -  
[01/Aug/1995:00:00:07 -  
0400] "GET / HTTP/1.0"  
200 12309*

✎ Export to an Oracle Database

Column Name	Remarks
Host	The hostname
REQ	The requested documents
TS	The time stamp
ReplyCode	The status code replied by web server
ReplyBytes	The Bytes that replied to the client
HTTPHead	The HTTP service name such as "GET", "HTTP" etc.
HTTPVer	The version number of http protocol such as "HTTP 1.0"
Duration	The time interval between two consecutive requests from the same host
IntArTime	The time difference between two consecutive requests

# Some basic statistics of the dataset



---

✎ 3,461,612 requests

- 2189225 requests for picture files

✎ 137,978 hosts

- 97680 hosts were recorded by their DNS name
- 40298 hosts were recorded by their IP addresses



# Questions of interest

---

## Questions

- Who visited (will visit) the web site?
- Which topics are hot?
- Which pages should be updated?
- Which pages contain errors?
- ...

## Implication?

- Where to put an ad campaign for a product
- How to target the market



# Correspondence analysis

---

- ✎ A descriptive/exploratory technique designed to analyze simple two-way and multi-way tables
  - Not based on a canonical distribution method or other theoretical distribution
- ✎ It was originally developed primarily in France by Jean-Paul Benzécri in the early 1960's and 1970's
  - Also known as homogeneity analysis, optimal scaling, optimal scoring etc.

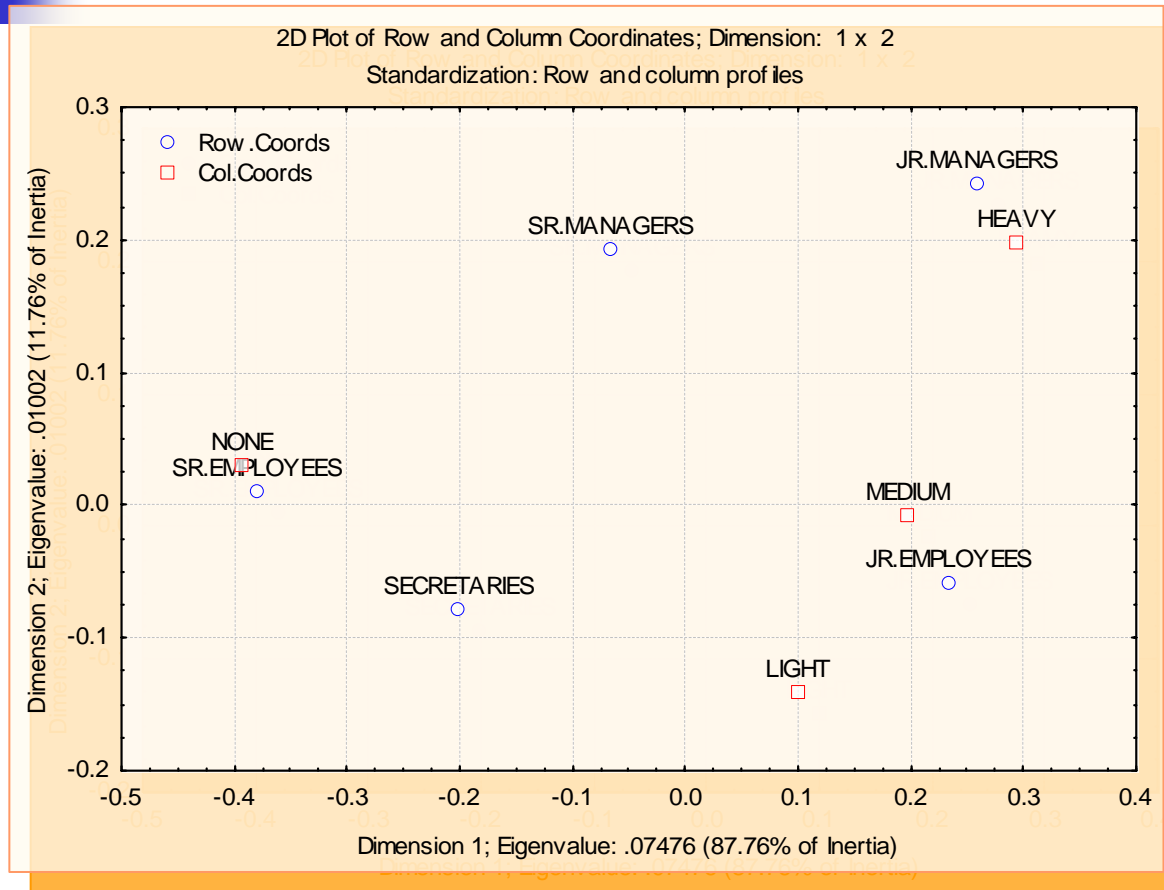
# Correspondence analysis—an example

Staff Group	Smoking Category				Row Totals
	(1) None	(2) Light	(3) Medium	(4) Heavy	
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
<b>Column Totals</b>	61	45	62	25	193

- ✎ 5 row points in the 4-dimensional space. The distance between the points summarize the similarities between the rows
- ✎ CA tries to find a lower dimensional space which retains yet almost all the information with regard to the differences between the rows/columns

(\*source of the example: Greenacre, 1984)

# An example (Cont.)



Eigenvalues and Inertia for all Dimensions					
Input Table (Rows x Columns): 5 x 4					
Total Inertia = .08519 Chi <sup>2</sup> = 16.442					
No. of Dims	Singular Values	Eigen-Values	Perc. of Inertia	Cumulative Percent	Chi Squares
1	.273421	.074759	87.75587	87.7559	14.42851
2	.100086	.010017	11.75865	99.5145	1.93332
3	.020337	.000414	.48547	100.0000	.07982

# Correspondence analysis

## ✎ Terminology

- **Mass**: the row and column totals of the matrix of relative frequencies are called the row mass and column mass, respectively
- **Inertia**: total Pearson *Chi-Square* for the two-way divided by the total sum:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

## ✎ Objective

- To identify the  $K^*$ -dimensional optimal subspace which minimizes the weighted sum of squared distances to the origin.
- Method: singular value decomposition (SVD)

# Notations & Calculation of CA

Datamatrix :  $\mathbf{N}(I \times J) = [n_{ij}], n_{ij} \geq 0; i = 1 \dots I$  and  $j = 1 \dots J$

Correspondence matrix :  $\mathbf{P} = (1/n_{..})\mathbf{N}$ , where  $n_{..} = \mathbf{1}^T \mathbf{N} \mathbf{1}$

Row and column sums :  $\mathbf{r} = \mathbf{P} \mathbf{1}$  and  $\mathbf{c} = \mathbf{P}^T \mathbf{1}$

Diagonal matrices of the sums :  $\mathbf{D}_r = \text{diag}(r)$  and  $\mathbf{D}_c = \text{diag}(c)$

Matrices of row and column profiles:

$$\mathbf{R} = \mathbf{D}_r^{-1} \mathbf{P} = \begin{bmatrix} \tilde{r}_1^T \\ \vdots \\ \tilde{r}_I^T \end{bmatrix} \text{ and } \mathbf{C} = \mathbf{D}_c^{-1} \mathbf{P}^T = \begin{bmatrix} \tilde{c}_1^T \\ \vdots \\ \tilde{c}_J^T \end{bmatrix}$$

$$in(I) = \text{trace}[\mathbf{D}_r (\mathbf{R} - \mathbf{1} \mathbf{c}^T) \mathbf{D}_c^{-1} (\mathbf{R} - \mathbf{1} \mathbf{c}^T)^T]$$

$$in(J) = \text{trace}[\mathbf{D}_c (\mathbf{C} - \mathbf{1} \mathbf{r}^T) \mathbf{D}_r^{-1} (\mathbf{C} - \mathbf{1} \mathbf{r}^T)^T]$$

$$in(I) = in(J) = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \chi^2 / n_{..}$$

$$\mathbf{P} - \mathbf{r} \mathbf{c}^T = \mathbf{A} \mathbf{D}_\mu \mathbf{B}^T \text{ where } \mathbf{A}^T \mathbf{D}_r^{-1} \mathbf{A} = \mathbf{B}^T \mathbf{D}_c^{-1} \mathbf{B} = \mathbf{I}$$

# A glance of the table (Domain – Requests (weekly))

✂ Categories of the number of requests:

- Low: <25% percentile
- Medium: 25-75% percentile
- High: >75% percentile

	1	2	3	4	
	Domain Name	Week	Level	Request	
56	.de		5 High	3614	
57	.de		6 High	4299	
58	.de		7 High	4625	
59	.de		8 High	4628	
60	.de		9 High	3101	
61	.dk		0 Medium	332	
62	.dk		1 Medium	674	
63	.dk		2 Medium	663	
64	.dk		3 Medium	658	
65	.dk		4 Medium	632	
66	.dk		5 High	1196	
67	.dk		6 High	1610	
68	.dk		7 High	1928	
69	.dk		8 Medium	721	
70	.dk		9 High	1109	
71	.do		1 Medium	78	
72	.do		2 Medium	39	
73	.do		3 Medium	47	
74	.do		4 Medium	29	
75	.do		5 Medium	53	

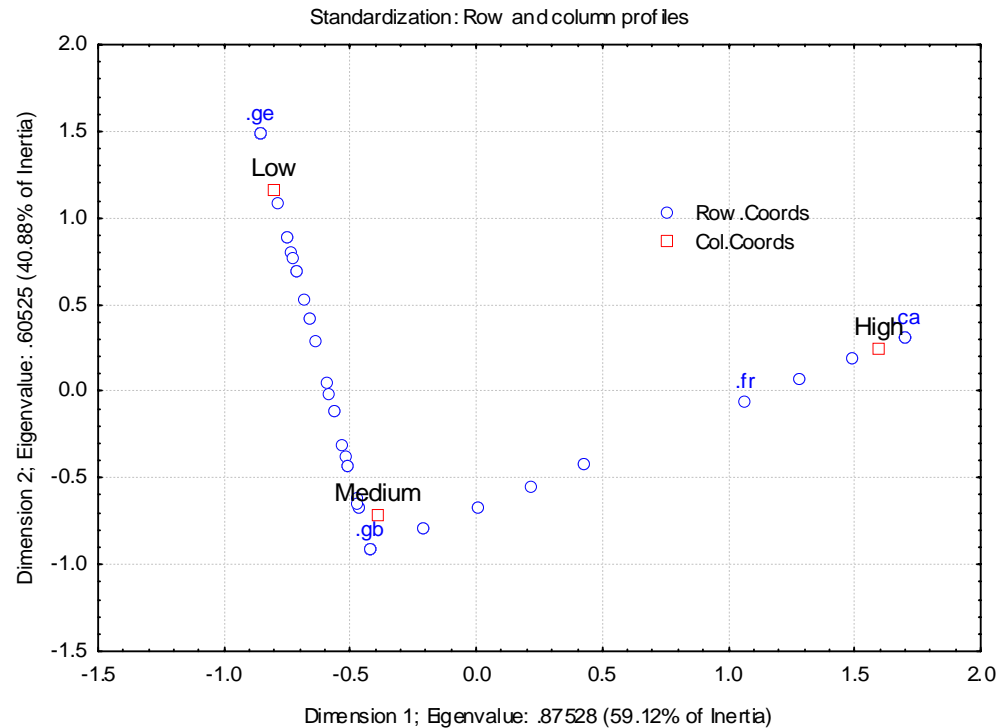


# Domain – Requests (weekly)

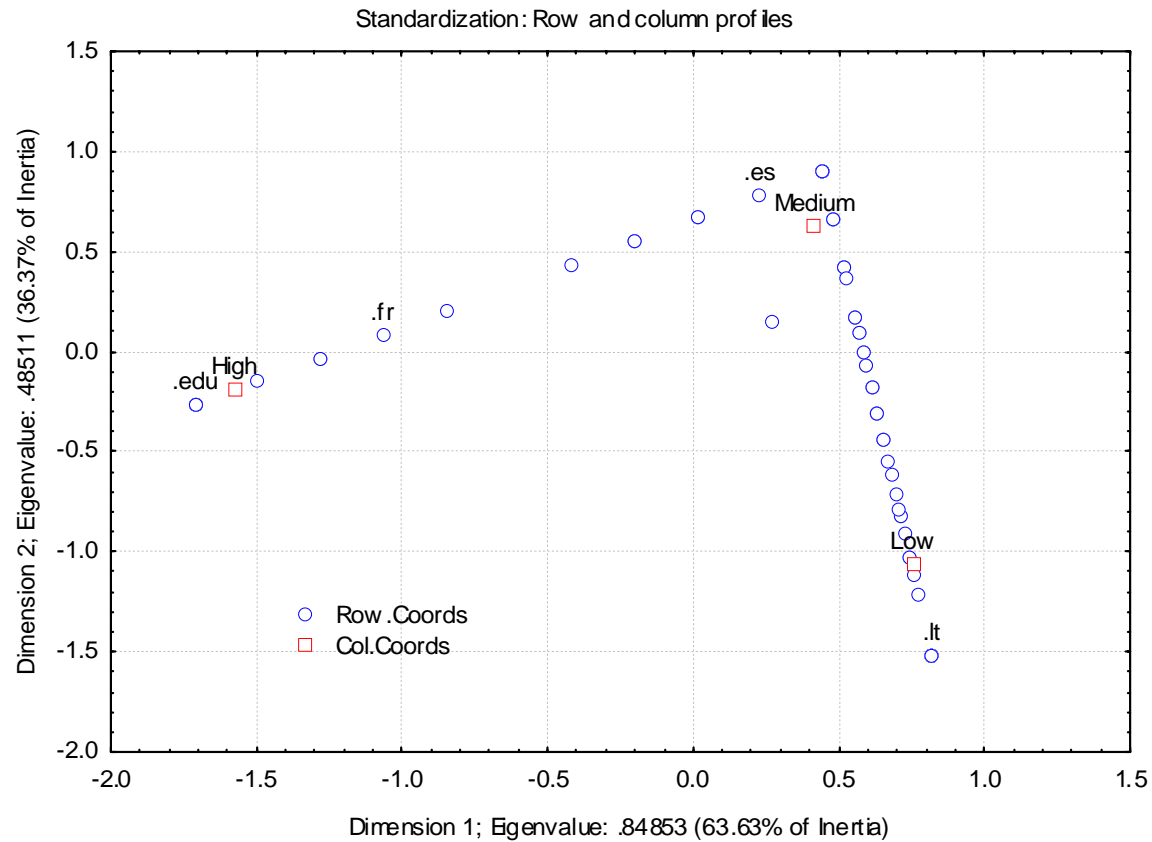
Column Name	Column Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine <sup>2</sup> Dim.1	Inertia Dim.2	Cosine <sup>2</sup> Dim.2
Medium	1	-0.393727	-0.711136	0.496855	1.000000	0.221738	0.087998	0.234619	0.415147	0.765381
High	2	1.592880	0.241267	0.250314	1.000000	0.438819	0.725612	0.977573	0.024074	0.022427
Low	3	-0.803288	1.158641	0.252830	1.000000	0.339443	0.186390	0.324629	0.560779	0.675371

Row Name	Row Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine <sup>2</sup> Dim.1	Inertia Dim.2	Cosine <sup>2</sup> Dim.2
.ar	1	-0.420844	-0.914083	0.012579	1.000000	0.008604	0.002545	0.174896	0.017365	0.825104
.arpa	2	-0.420844	-0.914083	0.012579	1.000000	0.008604	0.002545	0.174896	0.017365	0.825104
.at	3	-0.208501	-0.791663	0.012579	1.000000	0.005694	0.000625	0.064865	0.013025	0.935135
.au	4	1.702586	0.310121	0.012579	1.000000	0.025445	0.041658	0.967888	0.001999	0.032112
.be	5	-0.208501	-0.791663	0.012579	1.000000	0.005694	0.000625	0.064865	0.013025	0.935135
.bm	6	-0.639729	0.287607	0.012579	1.000000	0.004180	0.005881	0.831864	0.001719	0.168136
.br	7	1.490243	0.187700	0.012579	1.000000	0.019167	0.031915	0.984384	0.000732	0.015616
.ca	8	1.702586	0.310121	0.012579	1.000000	0.025445	0.041658	0.967888	0.001999	0.032112
.ch	9	1.277900	0.065280	0.012579	1.000000	0.013910	0.023468	0.997397	0.000089	0.002603
.cl	10	-0.420844	-0.914083	0.012579	1.000000	0.008604	0.002545	0.174896	0.017365	0.825104
.cn	11	-0.664049	0.421128	0.011321	1.000000	0.004728	0.005703	0.713171	0.003317	0.286829
.co	12	-0.508398	-0.433407	0.012579	1.000000	0.003792	0.003714	0.579123	0.003904	0.420877
.com	13	1.702586	0.310121	0.012579	1.000000	0.025445	0.041658	0.967888	0.001999	0.032112
.cr	14	-0.464621	-0.673745	0.012579	1.000000	0.005691	0.003102	0.322292	0.009434	0.677708
.cy	15	-0.858613	1.489298	0.001258	1.000000	0.002511	0.001059	0.249462	0.004610	0.750538
.cz	16	-0.420844	-0.914083	0.012579	1.000000	0.008604	0.002545	0.174896	0.017365	0.825104
.de	17	1.702586	0.310121	0.012579	1.000000	0.025445	0.041658	0.967888	0.001999	0.032112
.dk	18	0.428528	-0.424402	0.012579	1.000000	0.003090	0.002639	0.504838	0.003743	0.495162

# Domain – Requests (weekly)



# Domain – Bytes (weekly)





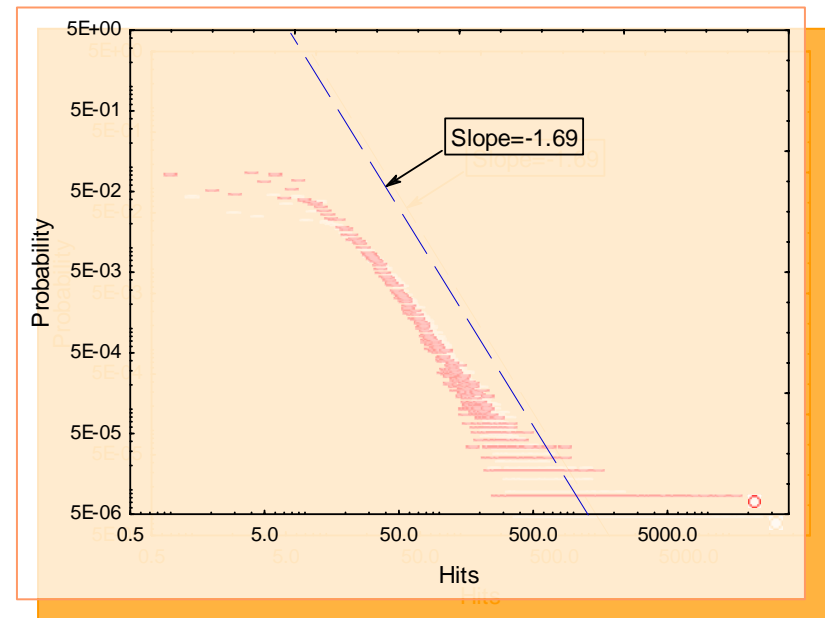
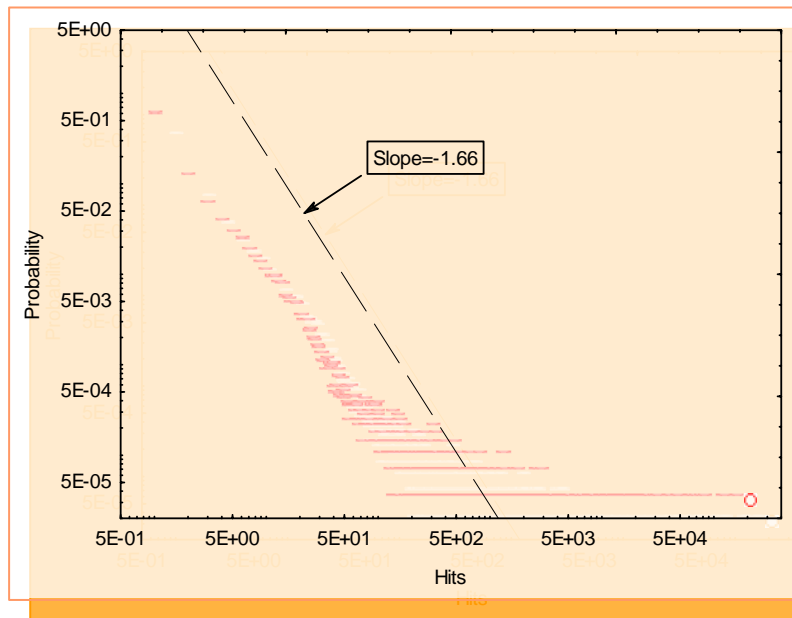
# Other criteria

---

- ✗ Domain-Request (daily)
- ✗ Domain-Byte (daily)
- ✗ Country-Request (weekly, daily)
- ✗ Country-Byte (weekly, daily)
- ✗ URL-Request
- ✗ ...

# Dominant distributions

- ✎ Power law distribution  $P(k) \sim K^{-\gamma}$ 
  - L: number of requests to the documents
  - R: number of requests from the hosts



# Consequence of power law distribution

## ✎ Statistical moments

$$\langle k^n \rangle \sim \int_{k_0}^{\infty} k^n k^{-\gamma} dk$$

Do not exist for  $n = [r] - 1, [r], \dots$  where  $[r]$  is the smallest integer greater than  $r$ : networks have no characteristic scales (**Scale-Free Networks**)

Examples of SFN: (1) **WWW**,  $r(\text{in}) \sim 2.1$ ,  $r(\text{out}) \sim 2.4$ ;  
(2) **Interent** ( $r \sim 2.5$ ) (3) **Network of movie actors** ( $r \sim 2.3$ ); (4) **Electrical power-grid of western US** ( $r \sim 4$ ) (5) **Scientific citation network** ( $r \sim 3.0$ )



## In conclusion...

---

- ✂ Correspondence analysis is a powerful yet simple tool to boost the web log data visualization & to identify the visit pattern and trend.
- ✂ Ongoing research
  - Real-time log data analysis
  - Combined with web content mining
  - Customized logging techniques



# Q & A

---

Thanks!

