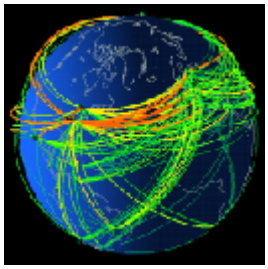
Understanding and Mining the Web

Ricardo Baeza-Yates
Center for Web Research
Dpto. de Ciencias de la Computación
Universidad de Chile
www.baeza.cl



Summary

- Characteristics and Models for the Web
- Web Mining
- Comparing Countries: Brazil, Chile, and Spain
- Relating Web Characteristics
- Queries: Faster Indices
- User Choices: Ranking
- Links and Web Dynamics: Ranking

(Part of this talk is joint work with Carlos Castillo and Felipe Saint Jean)

todobr O BRASIL NA INTERNET **(April 2000)** todocl TODO CHILE EN INTERNET **(May 2000, July 2001)** inicia BUSCOPIO **(January 2001)**

The Multiple Faces of the Web

Web Mining

The best case for data mining:

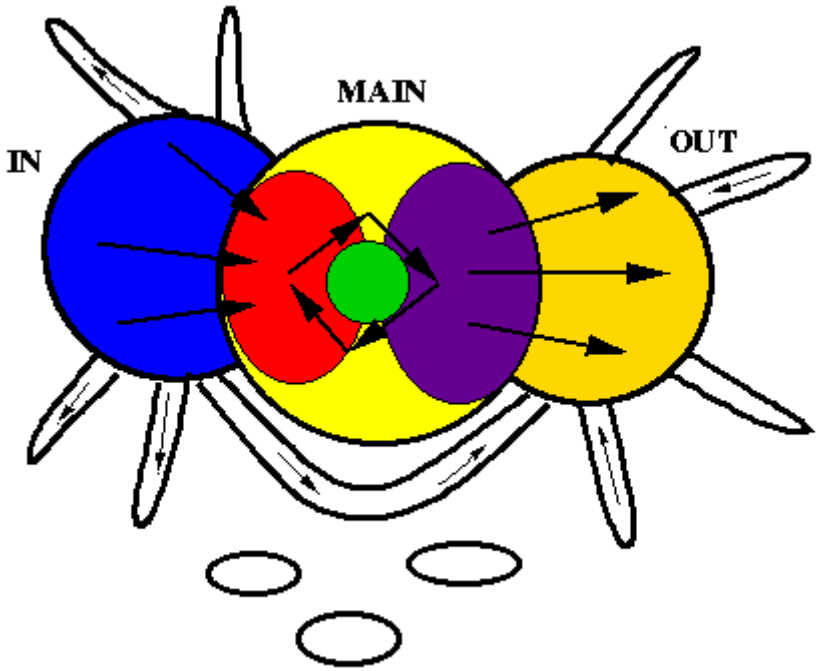
What?	Data type	Why?
Dynamics	Numeric, time sequences	Scalability
Structure	Graph	Popularity, clusters
User behavior	Transactions (logs)	Interfaces, Web organization, performance
Content	Text, multimedia	Semantics

In a search engine Web mining improves ranking, interfaces and performance (indices, etc.) as we will see later

---

Statistical Comparison

Towards Web Graph Mining: Refined Web Structure



Details for Chile (diameter of Main is 13 domains)

Structure

Country	Main and parts (in, out, main, norm)	Out	In	Others: tentacles, islands
Spain	71 (0,3,0,68)	10	12	0, 7
Chile	23 (5,8,2,11)	45	15	13, 4

Web Characteristics vs. Structure (Chile)

Sites and Pages Pages per Site

Raw Text and Tagged Text

Connectivity and Depth (Organization)

Pages (Brazil): 6.7 links on average (22% no links, 63% one link)

Top Websites: Chile and Spain

All graphs

Editor's and User's Choices: 3.100 Websites, 18.000 Searches (2000)

Usage vs. Structure Detailed View

Over 6.000 Websites and 700.000 Searches (2001)

Detailed View More ... (without SII)

Log Analysis: Web Queries

Opción	% de uso	frecuencia de uso
AND	99.9%	777.342
OR	0%	8
FRASE	0%	1
con acentos	0.1%	525
sin acentos	99.9%	777.021

Frequency: Chile (TodoCL, 2000, 2001) Spain (Buscopio, 2001)

Variance: Spain, Chile (Variance vs. Frequency)

What is People Asking? View 1 View 2

Application I: Two level index and cache of precomputed answers

Inverted File Two Levels and Cache

Minimize total answer time under memory constraints:

$$kW + V + 8 \sum_{i=k+1}^{p+k} L_i \leq M$$

$$E(L_k) = \frac{T}{N}$$

$$M = kW + V + 8p\frac{T}{N}$$

$$p = \frac{N(M - kW - V)}{8T} = \frac{N(M-V)}{8T} - \frac{NW}{8T}k$$

$$k \leq (M - V)/W$$

Analysis

Frequency vs. Random order

Cache of precomputed answers

Cache improvement Impact on search time

This can also be used for load balancing in distributed inverted indices

Given a global inverted index,
how to distribute the word lists among *p* processors:

$X = \{x_1, ..., x_p\}$
partition of p subsets
of the set of words

$$\min_X \left\{ \sum_{i=1}^{p} \left(\frac{T_{tot}}{p} - \sum_{j \in x_i} T(f_j) \right)^2 \right\}$$

under the restriction of reasonable memory balancing for all i

$$\alpha M \leq \sum_{j \in x_i} S(j) \leq M$$

Issues: replication, on-line, adaptive

Log Analysis: Navigation

User Navigation State Diagram

Average answers seen: 1.15 paginas (Chile) vs over 2 (Spain)

Should be used for *user-driven Web design*:
organization, links and anchor texts (from sitesearch)

Application II: Ranking can be modified based in user choices

Users depend on queries, queries on answers, and answers on

users

For a given query *q and a time period *T*:**

$$u_i = \sum_{j=1}^{n} a_{ij} q_j, \qquad a_{ij} = \begin{cases} num(u_i, q_j) & q_j = q^* \\ \alpha num(u_i, q_j) & q_j \neq q^* \end{cases}$$

$$q_j = \sum_{k=1}^{s} b_{jk} r_k, \qquad b_{jk} = \begin{cases} sim(q_j, r_k) & q_j = q^* \\ \beta sim(q_j, r_k) & q_j \neq q^* \end{cases}$$

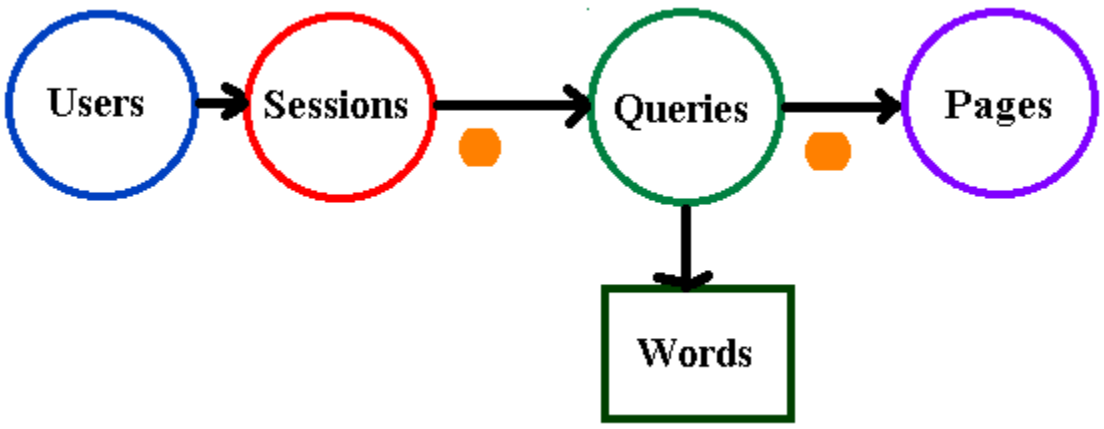
$$r_k = \sum_{i=1}^{m} c_{ki} u_i, \qquad c_{ki} = hit(r_k, u_i, \{q^*\}) + \gamma hit(r_k, u_i, Q - \{q*\})$$

$$u = Aq, \quad q = Br, \quad r = Cu$$

$$u = Aq = A(Br) = A(B(Cu)) = (ABC)u$$

Precision depends on *T*

Improvement: add sessions and a graph of relations



Ranking Algorithms based on Links

- Link Analysis
- Page Distribution

Ranking Web Sites: Average, Max or Sum?

- Rankings vs. Structure
- Site Distribution

Application III: Ranking based on Links & Age

- Web Dynamics, Age and Page Quality