

Mining Web Queries



Ricardo Baeza-Yates
Center for Web Research
www.cwr.cl
 Dept. of Computer
 Science
 Universidad de Chile,
 Santiago
ricardo@baeza.cl



Summary

- **Web Mining**
- **Queries: Faster Indices**
- **Queries and User Clicks: Website Design, Ranking, Recommendations**

(Parts of this talk are joint work with Carlos Castillo, Georges Dupret, Carlos Hurtado, Marcelo Mendoza,

Barbara Poblete, Cuauhtemoc Rivera, Felipe Saint-Jean, and Javier Velasco)



www.todo.cl (2000-2005)

The Multiple Faces of the Web **The Structure of the Web**

Web Mining

What?	Data type	Why?
Dynamics	Numeric, time	Scalability

	sequences	
Structure	Graph	Popularity, Communities
User behavior	Transactions (logs)	Interfaces, Web organization, Performance
Content	Text, multimedia	Information Scent, Semantics

Our Goals

- Exploit the semantic embedded on Web interactions
- Improve search engines: ranking, query recommendation, performance, etc.
- Understand the context of the search
 - Primary goal: information vs. navigation/transaction
 - Some results: [WWW2004](#), [SIGIR2003](#)
 - Building better baselines (WWW2004: [q1](#), [q2](#))
- Better [Web Design](#), better [Findability and Use](#)
- Generate pseudo-semantic resources

• Data from a Chilean Search Engine: TodoCL.com

Data	2000	2001	2002	2003	2004
Pages	529,159	794,218	2,214,253	3,153,089	3,254,1
Web Sites	7,483	21,207	38,965	38,277	53,52

Domains	6,288	18,032	35,390	33,981	47,40
----------------	--------------	---------------	---------------	---------------	--------------

- **Well formed collection: mostly one language, one culture**
- **Half a million queries per month**

Log Analysis: Web Queries

Chile and Spain: default use (AND) and low use of sentences (12 to 15%) [User Profile](#)

Term Frequencies in TodoCL: [2000](#), [2001](#), [2003](#)

What is People Asking? [View 1](#) [View 2](#) [View 3](#) [View 3a](#) [View 4](#) [View 5](#)
(Excite/Fast/ToDoCL)
Data for [Sessions](#)

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 Growth

Impact of Web

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:**

$X = \{x_1, \dots, x_p\}$
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

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

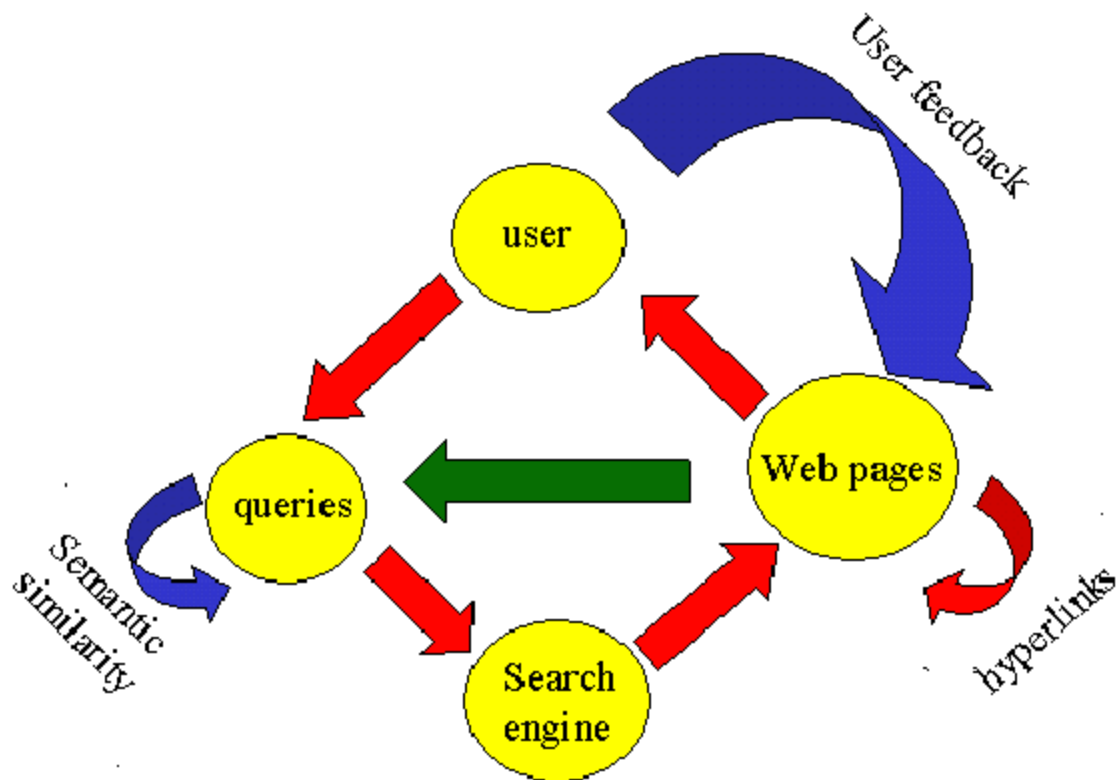
Issues: replication, on-line, adaptive

Log Analysis: Navigation and URLs clicked after each query

User Navigation State Diagram

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

Application II: Query Clustering for Ranking Boosting and Query Recommendation



Vector Model for Queries:

$$q[i] = \sum_{URLu} \frac{\text{Pop}(q, u) \times \text{Tf}(t_i, u)}{\max_t \text{Tf}(t, u)}$$

Problem: User clicks are [biased by the ranking](#)

Unbiasing using a power law distribution with parameter b

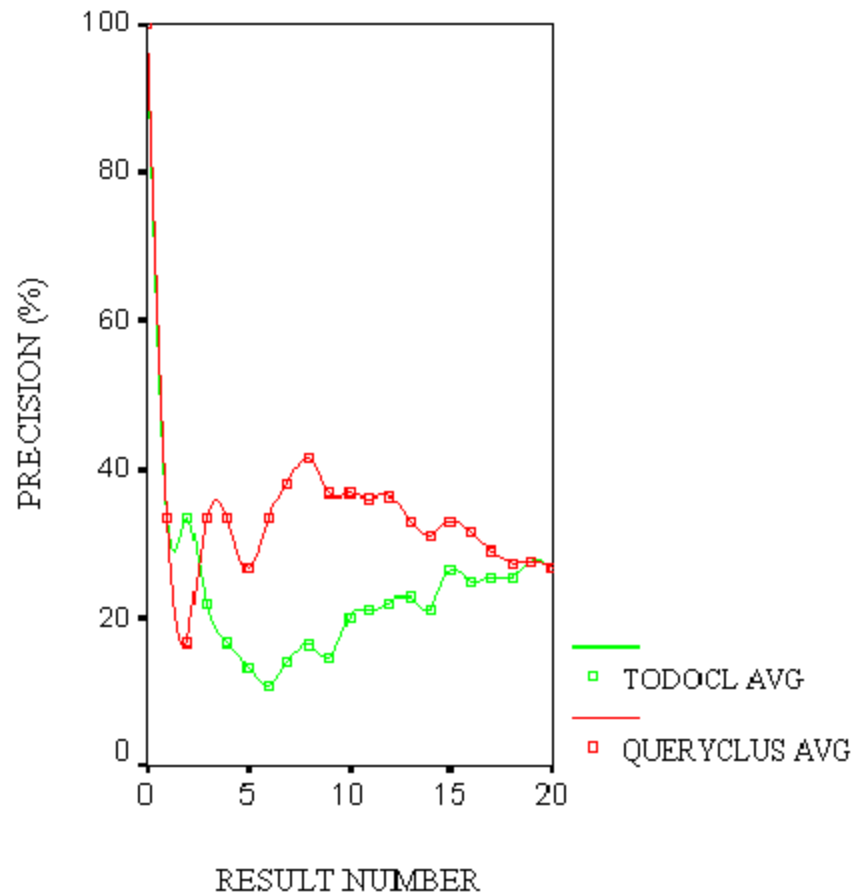
[Clustering](#) using a k-means algorithm on the text of the clicked URLs ([Histogram](#))

Examples:

Q	Cluster Rank	ISim	ESim	Queries in Cluster
q_1	252	0,447	0,007	car sales, cars Iquique, cars used, diesel, new cars,
q_2	497	0,313	0,009	stamp, serigraph inputs, ink reload, cartridge
q_3	84	0,697	0,015	office rental, rentals in Santiago, real state, apartment rental

- **Ranking boosting:**

$$\text{NewRank}(u) = \beta \text{OrigRank}(u) + (1 - \beta) R$$



- **Query Recommendation:**

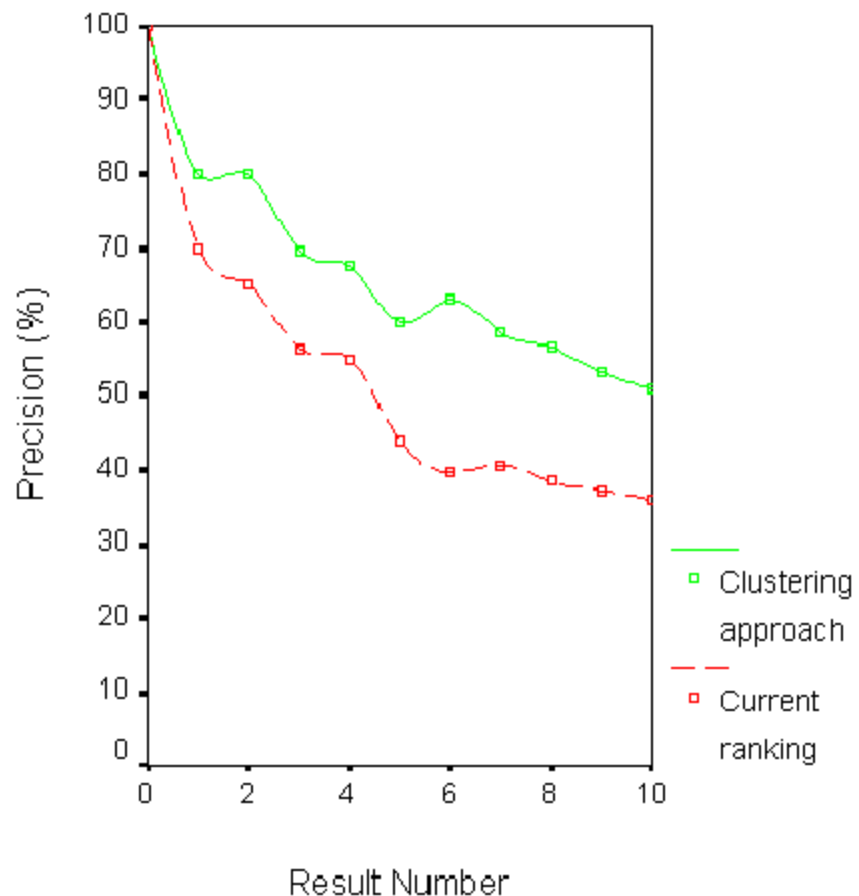
$$\text{Rank}(q) = \gamma \times \text{Sup}(q, q_{ini}) + (1 - \gamma) \times \text{Cl}$$

Closedness: cosine similarity over query model

Support: Fraction of documents returned and clicked for the query

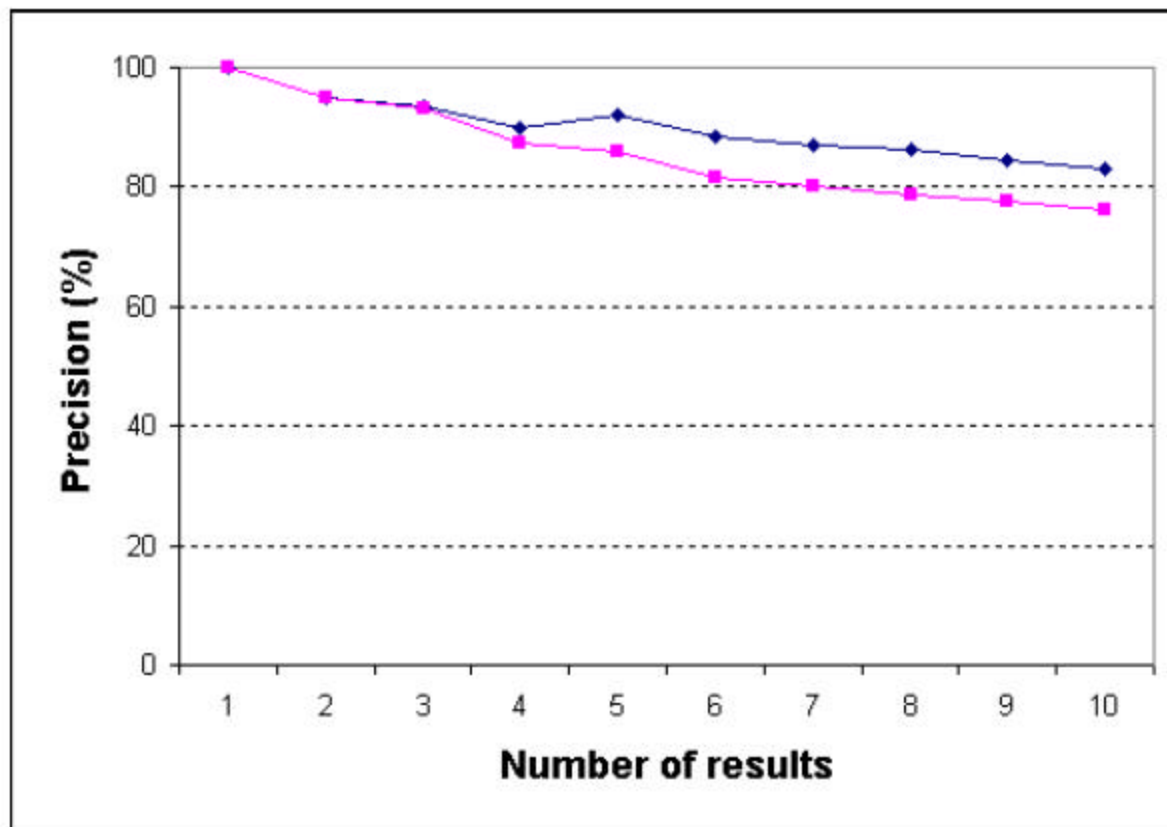
Example: summer rental

Query	Popularity	Support	Clos
rentals apartments viña del mar owners	2	0,133	0,40
rentals apartments viña del mar	10	0,2	0,25
viel properties	4	0,1	0,31
rental house viña del mar	2	0,166	0,12
house leasing rancagua	8	0,166	0,03
quintero	2	0,166	0,02
rentals apartments cheap vina del mar	3	0,033	0,15
subsidize renovation urban	5	0,133	0,00
houses being sold in pucon	10	0	0,11
apartments selling pucon villarrica	2	0,066	0,01
portal sell properties	3	0,033	0,02
sell house	2	0,033	0,01
sell lots pirque	2	0,033	0,00
canete hotels	1	0	0,01



Application III: Generating Pseudo-Semantic Resources (ongoing work)

- **Query dominance: automatic generation of pseudo-taxonomies of semantic relations based on query relations ([ex1](#), [ex2](#), [ex3](#))**
- **[Document Classification](#) based on queries: preliminary results improves upon human judgments**



Application IV: Capturing Information Scent

User-driven Web design:

organization (hotlinks) and anchor texts (from query analysis)

Query centered mining model

Classes of queries/pages: Navigation (DRWS) vs. Search (DRQ) No Clicks/Answers

Class	Semantic exists	Answer	Clicks	Document type	Interest
A	yes	yes	yes	$DRQ \cap DRWS$	low
B	yes	yes	yes	$DRQ - DRWS$	medium
C	yes	yes	no	—	low
C'	yes	no	no	—	medium
D	it should	no	no	—	high
E	no	no	no	—	none

Tool that analyzes correlations of queries, clicks & navigation ([Demo](#), [Examples](#))

Also: Text clustering and correlation with internal links

Other Applications

- **Web Structure and Link Page Ranking: Pagerank, Authorities, and Hubs**

Pagerank based on link age and [biased to new pages](#)

- **Web Dynamics: Structure Composition [Evolution](#) (%)**

What is [happening?](#): (2003: [All cases](#), [Stable cases](#)) (2004: [A1](#) [A2](#) [S1](#) [S2](#))

Looks like duplication: 115% growth, 25% death

- **Correlations with user behavior: choices of ODP editor's and TodoCL user's are correlated with [Web structure](#) and hence page quality**

Final Remarks

- **Plenty of data mining to do**
 - **First, understanding queries (different types of segmentations)**
 - **Dynamics of queries imply [social behavior](#) (social Web Mining)**
 - **Optimality of static caching with changes of the query distribution**
 - **More on Web structure dynamics, including death**
 - **Clustering user actions/topics?**
 - **Model for [information mining](#) ([Example](#))**
 - **Link-based quality against real quality?**
-