Agglomerative clustering of a search engine query log

Doug Beeferman

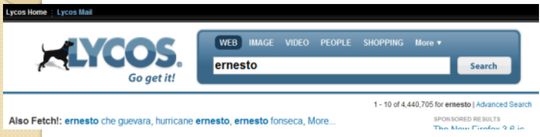
Lycos

Adam Berger

CMU

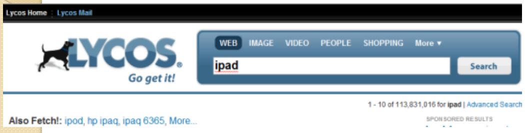
KDD 2000

Presented by: Cangming & Ronan



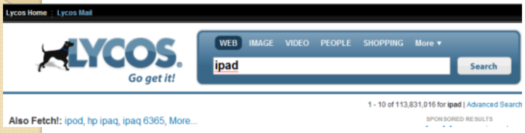
1 - 10 of 4,440,705 for **ernesto** | Advanced Search

Also Fetch!: **ernesto** che guevara, hurricane **ernesto**, **ernesto** fonseca, More...



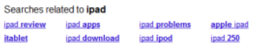
1 - 10 of 113,831,016 for **ipad** | Advanced Search

Also Fetch!: ipod, hp ipaq, ipaq 6365, More...



1 - 10 of 113,831,016 for **ipad** | Advanced Search

Also Fetch!: ipod, hp ipaq, ipaq 6365, More...



Searches related to **ipad**

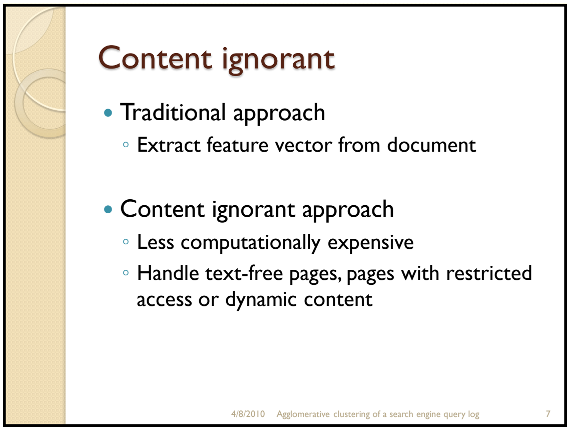
ipad **review**	ipad **apps**	ipad **problems**	**apple** ipad
itablet	ipad **download**	ipad **ipod**	ipad **250**

Disclaimer

	2000	2010
Workstation speed	266 MHz	> 3 GHz
Lycos used	Yes	?
Queries / day	10 million	400 million
Size of the internet	?	5 million TB

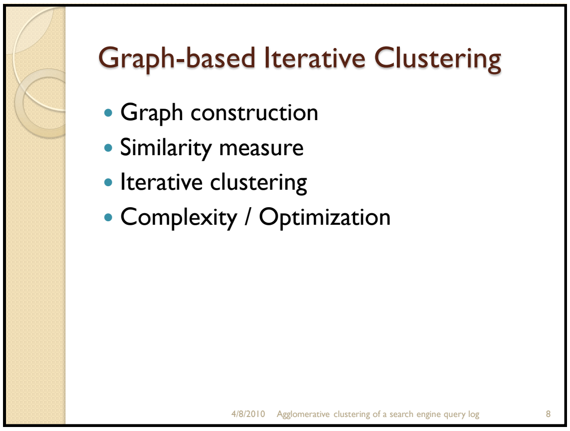
Motivation

- Query log analysis
  - Identifying late-breaking trends
- Clustering URLs
  - Generating ontology, organizing bookmarks, grouping search results
- Clustering queries
  - Query recommendations



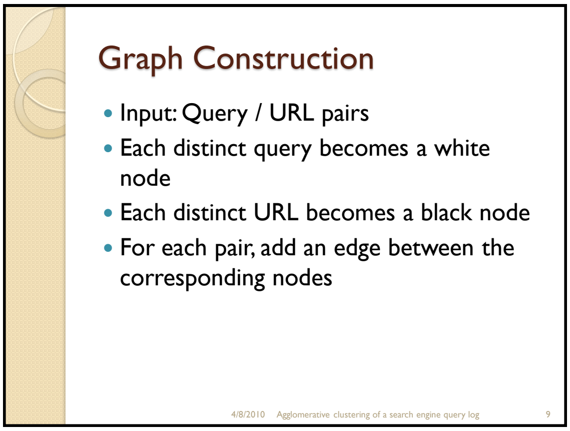
Content ignorant

- Traditional approach
  - Extract feature vector from document
- Content ignorant approach
  - Less computationally expensive
  - Handle text-free pages, pages with restricted access or dynamic content



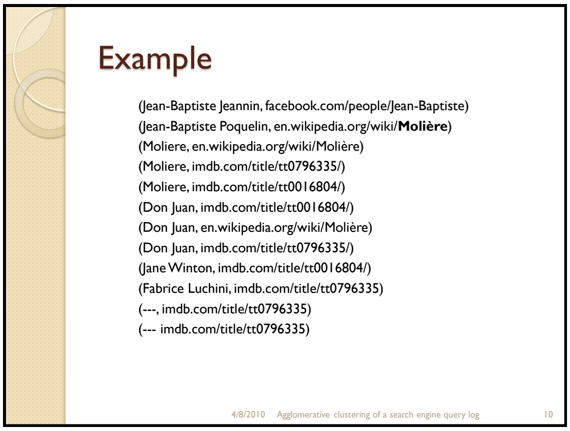
Graph-based Iterative Clustering

- Graph construction
- Similarity measure
- Iterative clustering
- Complexity / Optimization



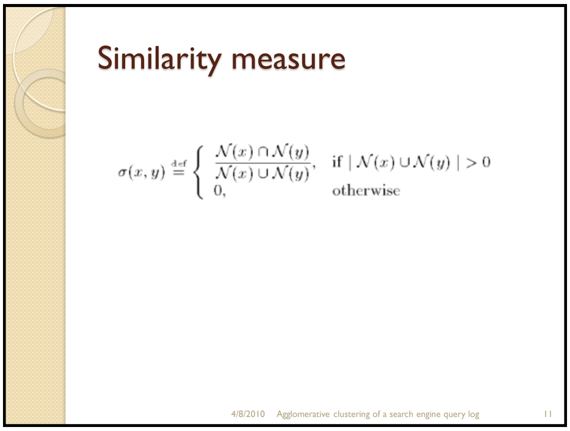
Graph Construction

- Input: Query / URL pairs
- Each distinct query becomes a white node
- Each distinct URL becomes a black node
- For each pair, add an edge between the corresponding nodes



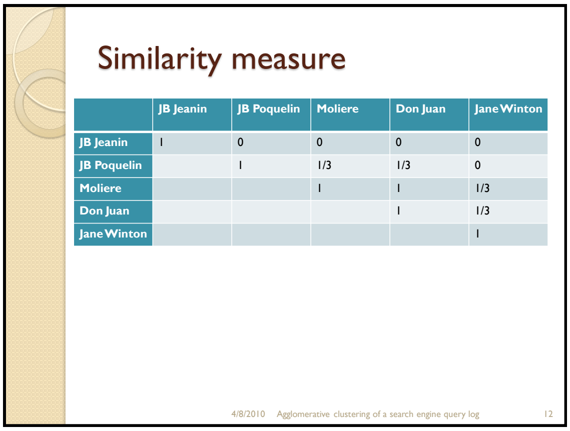
Example

(Jean-Baptiste Jeannin, facebook.com/people/Jean-Baptiste)
(Jean-Baptiste Poquelin, en.wikipedia.org/wiki/**Molière**)
(Moliere, en.wikipedia.org/wiki/Molière)
(Moliere, imdb.com/title/tt0796335/)
(Moliere, imdb.com/title/tt0016804/)
(Don Juan, imdb.com/title/tt0016804/)
(Don Juan, en.wikipedia.org/wiki/Molière)
(Don Juan, imdb.com/title/tt0796335/)
(Jane Winton, imdb.com/title/tt0016804/)
(Fabrice Luchini, imdb.com/title/tt0796335)
(---, imdb.com/title/tt0796335)
(--- imdb.com/title/tt0796335)



Similarity measure

$$\sigma(x,y) \overset{\text{def}}{=} \begin{cases} \dfrac{\mathcal{N}(x) \cap \mathcal{N}(y)}{\mathcal{N}(x) \cup \mathcal{N}(y)}, & \text{if } | \mathcal{N}(x) \cup \mathcal{N}(y) | > 0 \\ 0, & \text{otherwise} \end{cases}$$



Similarity measure

	JB Jeanin	JB Poquelin	Moliere	Don Juan	Jane Winton
JB Jeanin	1	0	0	0	0
JB Poquelin		1	1/3	1/3	0
Moliere			1	1	1/3
Don Juan				1	1/3
Jane Winton					1

Iterative Clustering

- Repeatedly merge the most similar pair, alternating between queries and URLs
- Until some stopping criterion is met
- Iterative approach helps to group queries (or URLs) that are otherwise uncorrelated

Complexity

- Naïve analysis: $\Theta(n_w^2 + n_b^2)$r iteration
- Incremental computation of distances
  - Compute only non-zero distances

$$\underbrace{(n_w + n_b)|\mathcal{N}|_{\max}{}^2}_{One\text{-}time\ computation}$$

  - Re-compute only distances that are susceptible to have changed

$$\underbrace{m(4|\mathcal{N}|_{\max})}_{per\ iteration}$$

Experiment: Query Recommendations

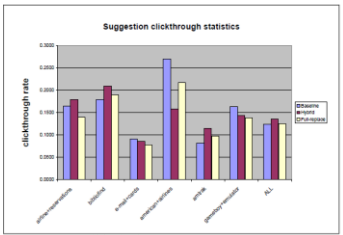
- Comparing 3 methods for building suggestion lists:
  - Baseline
  - Full-replacement
  - Hybrid
- Measure performance by clickthrough rate

Data

- Learning data: 500,000 query/URL pairs
- Test data: around 6 million impressions for each method

Results

period	strategy	impressions	clicks	clickthrough rate
April 7-8	baseline	6,120,943	71138	1.16%
April 14-15	hybrid	6,058,757	79515	1.31%
April 21-22	full replacement	5,985997	61377	1.03%



Discussion

- Similarity metric limitation
  - Two URLs shared should be better than one
  - More clicks should mean better correlation
  - Sensitivity to noisy clickthroughs

➜ Adding weights to edges

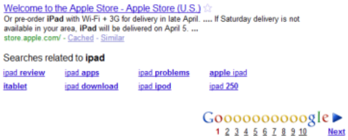
"Clustering Search Engine Query log containing noisy clickthroughs", Wing Shun Chan et al., 2004.

Discussion

- Quality of the search engine
  - Clustering only as good as the search engine itself
  - Perhaps, it might possibly improve the search experience, and the IR.

Discussion

- Limitation of experiments
  - Clickthrough does not reflect the quality of clustering
  - Users are likely to prefer search refinements over related searches



Welcome to the Apple Store - Apple Store (U.S.) ☆
Or pre-order **iPad** with Wi-Fi + 3G for delivery in late April. .... If Saturday delivery is not available in your area, **iPad** will be delivered on April 5. ...
store.apple.com/ - Cached - Similar

Searches related to **ipad**

ipad review	ipad apps	ipad problems	apple ipad
itablet	ipad download	ipad ipod	ipad 250