The Structure of Broad Topics on the Web

Chakrabarti, Joshi, Punera, Pennock

Presented by: Xiaoguang Qi

What Are We Trying to Find Out

- Convergence of topic distribution on undirected random walks
- Degree distribution restricted to topics
- How topic-biased are breadth-first crawls?
- How representative are web directories of topics on the web?
- Topic convergence on directed walks
- Link-based vs. content-based Web communities

How do we do it

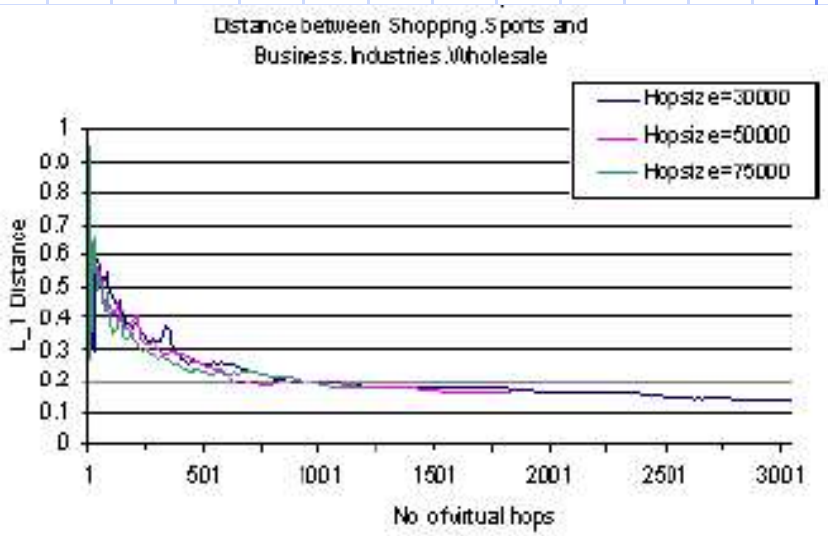
- Start from web directories
  - Open Directory
  - Prune the dmoz hierarchy to 482 topics and 144,859 URLs
- Train text classifier
  - Rainbow, Naïve Bayes
- New documents are topic vectors
  - d = (0.5, 0.3, 0.2)

Sampling Web Pages

- PageRank-based random walk
- Wander Walk
  - The same as PageRank walk except d=0
- The Bar-Yossef random walk
  - Make graph undirected
  - Make graph regular: add self loops
- Sampling walk
  - Bar-Yossef walk with random jump

Topic Convergence

- Start from two different topics
- Perform Sampling walk from each of them
- Measure the topic distance between two sampled page sets



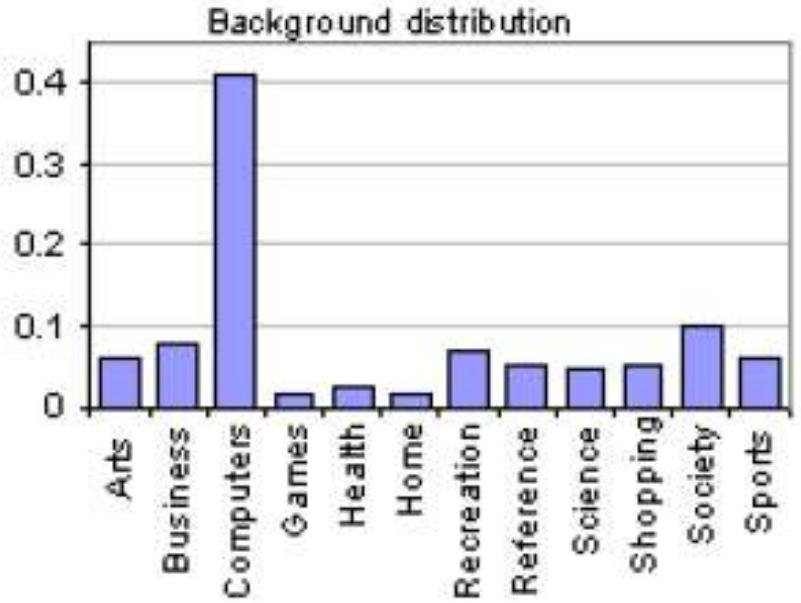
$$\vec{p}(D) = \frac{1}{|D|}\sum_{d \in D} \vec{p}(d)$$

(Soft counting)

$$L_1(D_1, D_2) = \sum_c |\, p_c(D_1) - p_c(D_2)\,|$$

Background Distribution

- An estimation of the background distribution
- 12 top-level topics
- “Computers” accounts for more than 40%



Faithful Representation of Topics in web directories

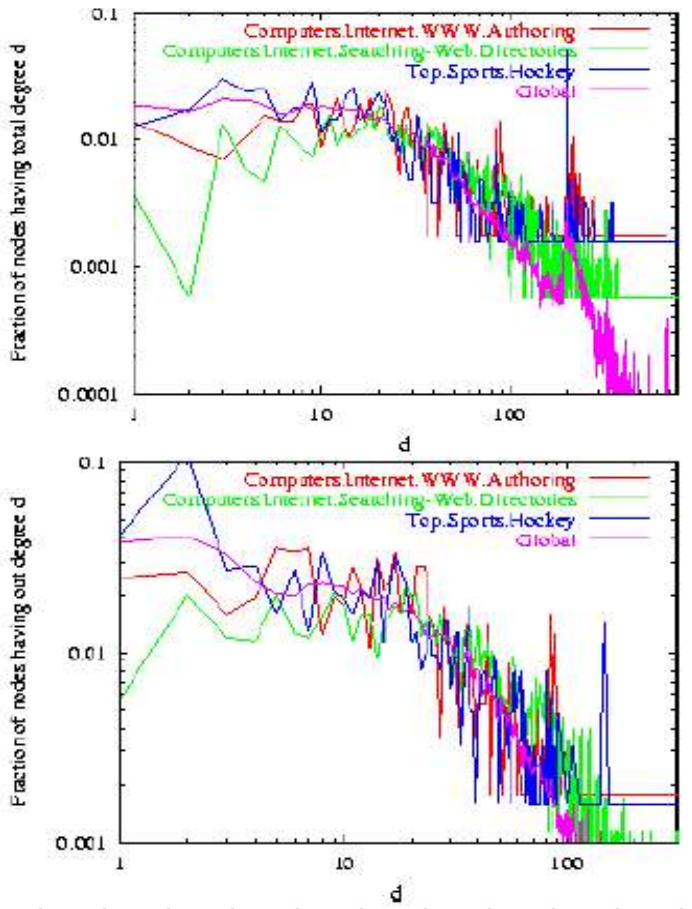
- Many web users implicitly expect topic directories to be a microcosm of the web itself
- Our sample of dmoz is highly topic biased
  - L1 distance between dmoz and background distribution is high (1.43)

Topics OVER-represented in Dmoz compared to the background
Games.Video_Games.Genres
Society.People
Arts.Celebrities
Reference.Education.Colleges_and_Universities.North_America...
Recreation.Travel.Reservations.Lodging
Society.Religion_and_Spirituality.Christianity.Others
Arts.Music.Others
Reference.Others
Topics UNDER-represented in Dmoz compared to the background
Computers.Data_Formats.Markup_Languages
Computers.Internet.WWW.Searching_the_Web.Directories
Sports.Hockey.Others
Society.Philosophy.Philosophers
Shopping.Entertainment.Recordings
Reference.Education.K_through_12.Others
Recreation.Outdoors.Camping

Figure 7: Some of the largest discrepancies between the Web's background topic distribution and our selection from Dmoz.

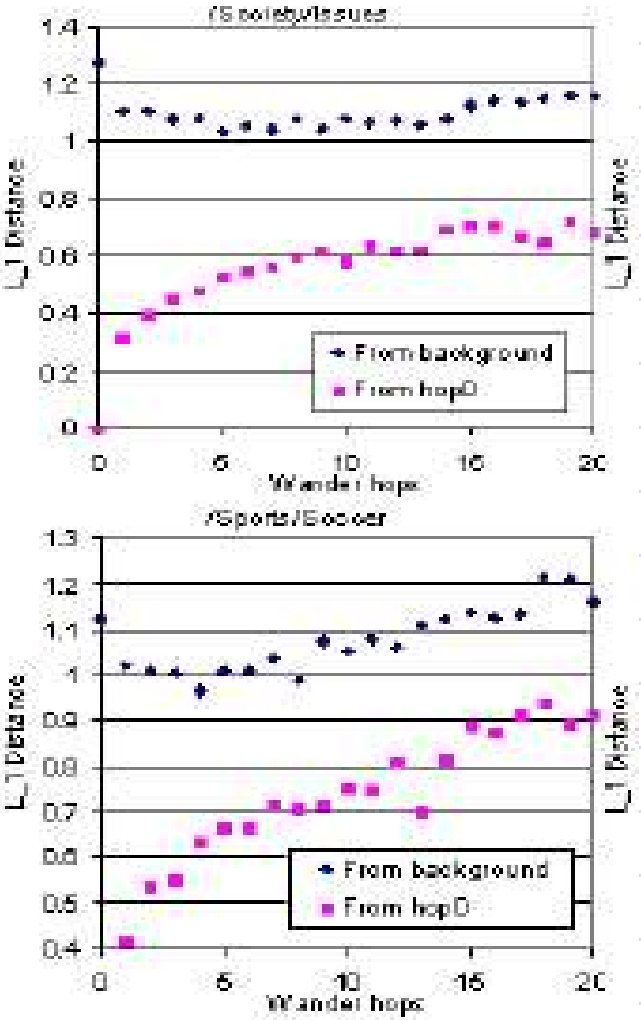
Topic-Specific Degree Distributions

- Degrees of web pages in general follow a power law distribution
  - The probability that a randomly picked node has degree i is proportional to 1/i^x, for some constant power x>1
- Does power law still hold in fixed topics?
  - Yes!



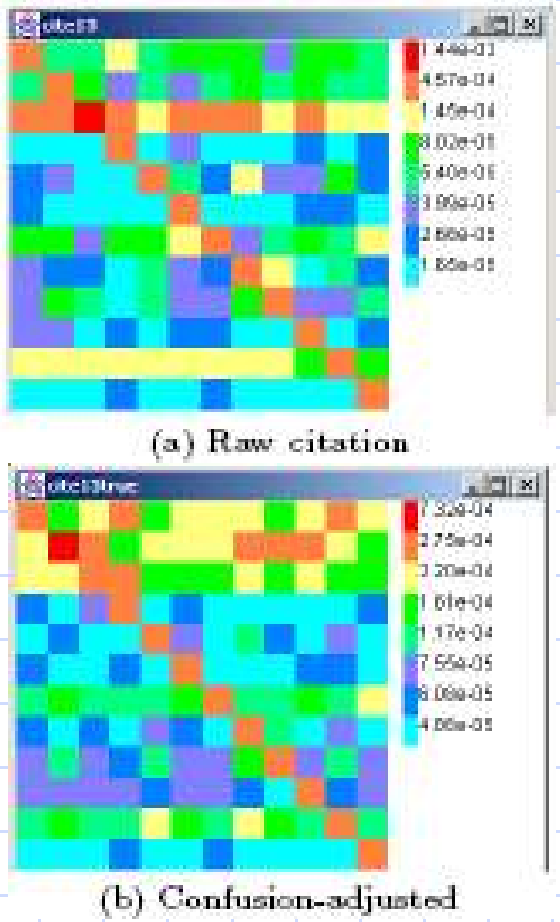
Topical Locality

- How?
  - Wander walk
    - Like PageRank walk, no jumping
  - Start from a page related to a specific topic
  - Collect the pages D_i found at distance i
  - Find soft classification histogram $p(D_i)$
  - Calculate the L_1 distance between $p(D_i)$ and background distribution, and the distance between $p(D_i)$ and $p(D_0)$



Relations Between Topics

- An N*N matrix, C(i, j) is the probability that a page about topic i links to a page about topic j
- Soft counting:
  - $C(i, j) = C(i, j) + p_i(u) * p_j(v)$


(a) Raw citation

(b) Confusion-adjusted

Concluding Remarks

- What we have shown
- Possible future work
  - How to set PageRank jump parameter?
  - Topic stability of distillation algorithms
  - Better crawling