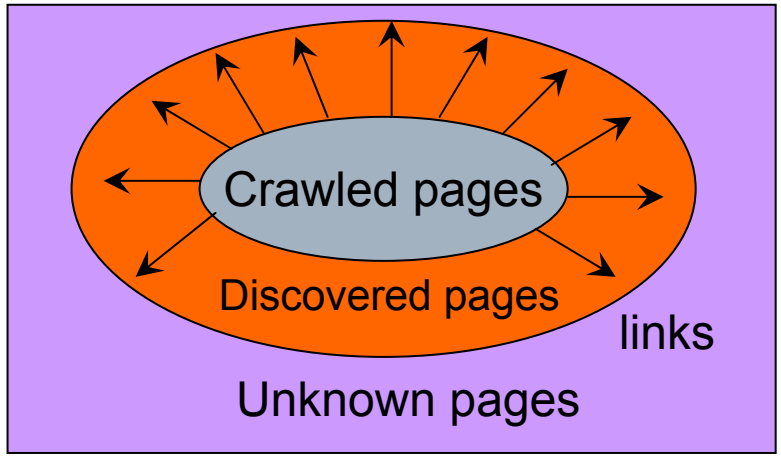
# Crawl Ordering by Search Impact

Sandeep Pandey
Christopher Olston

# Selecting pages to crawl next

Goal: Crawl discovered pages



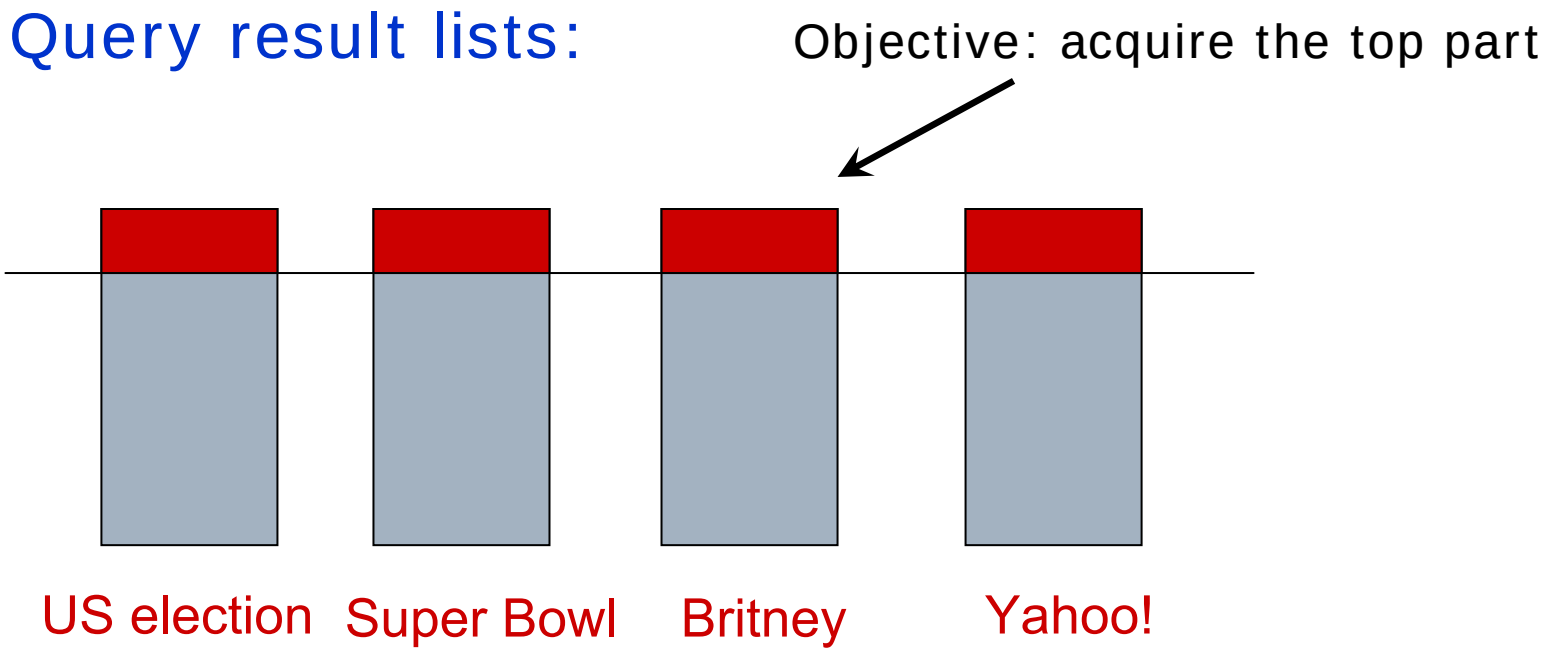

Challenges:

- Huge number of pages
- Varying quality
- Quality is hard to judge beforehand

# Crawling Objective

- acquire pages that show up in query results (**impact**)

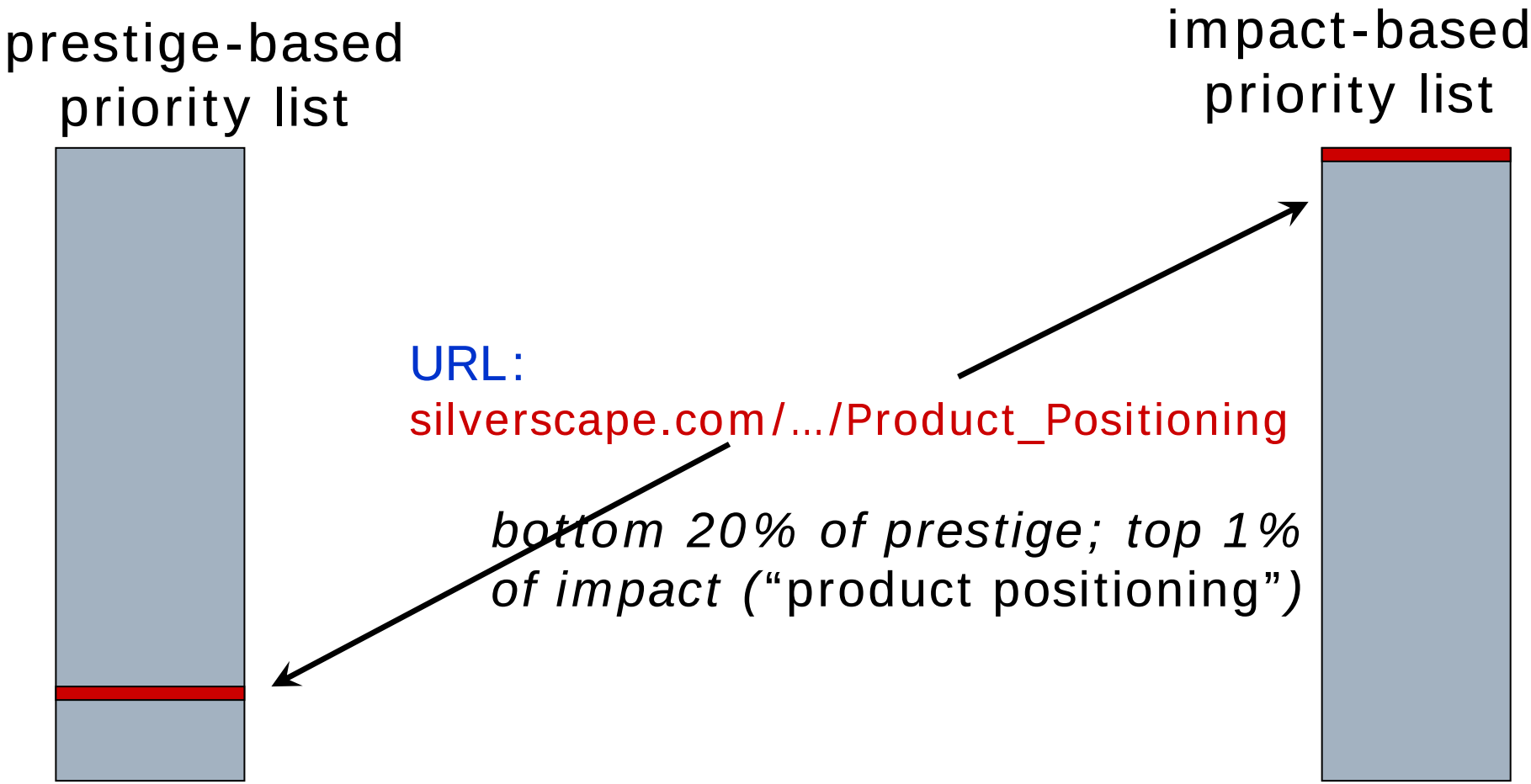
Query result lists:



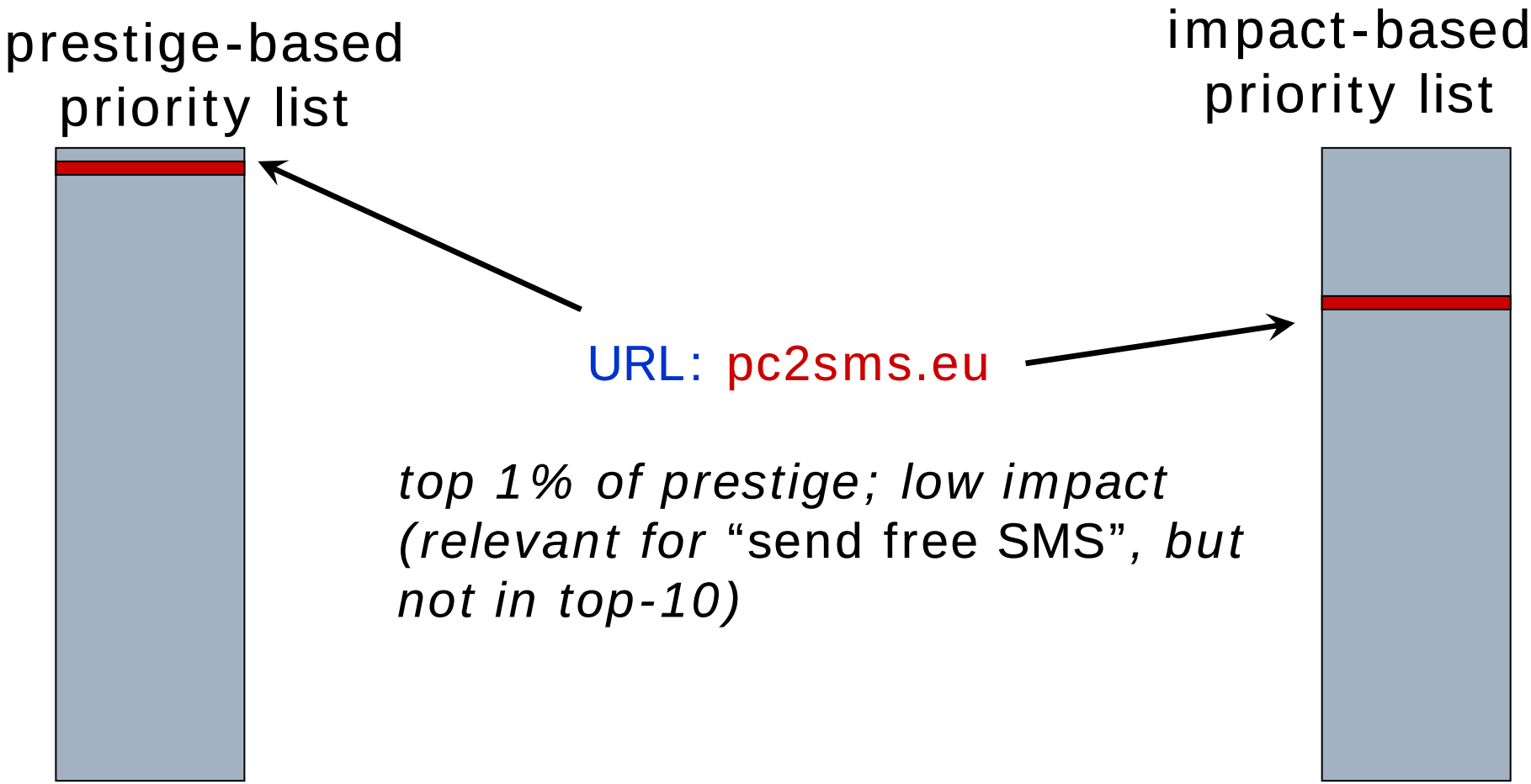
# Impact of Crawling Page p

- Impact(p) = $\sum_{\text{queries } q}$ freq(q) * top-K(p,q)
- top-K(p,q) = { 1 if p is in top-K results of q, 0 otherwise
- Ideal approach: Crawl high impact pages
- Standard approach: Crawl high prestige pages
  - e.g., Pagerank or approximation thereof
    [Najork et. al. WWW'01; Abiteboul et. al. WWW'03]

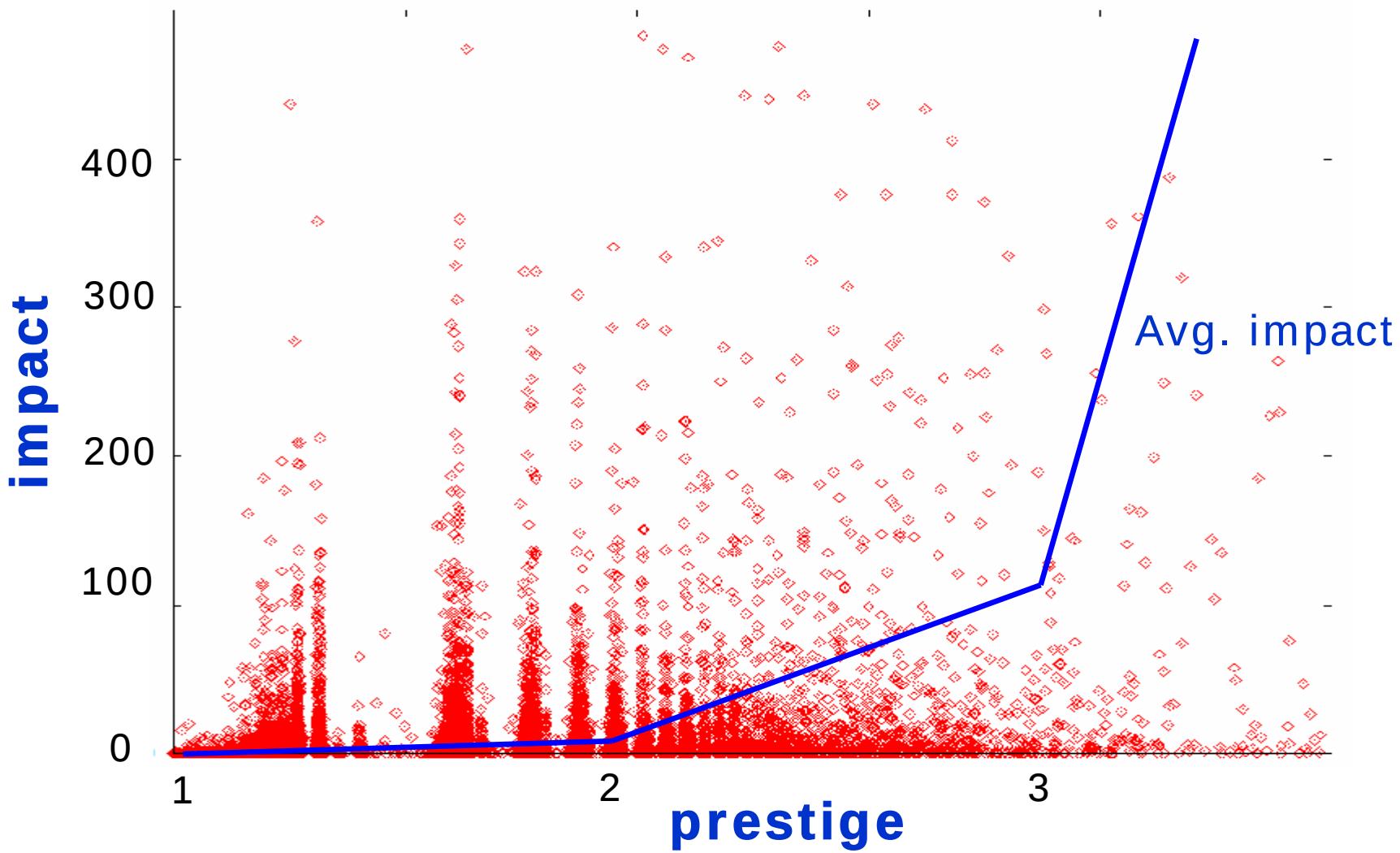
# prestige ≠ impact



# prestige ≠ impact



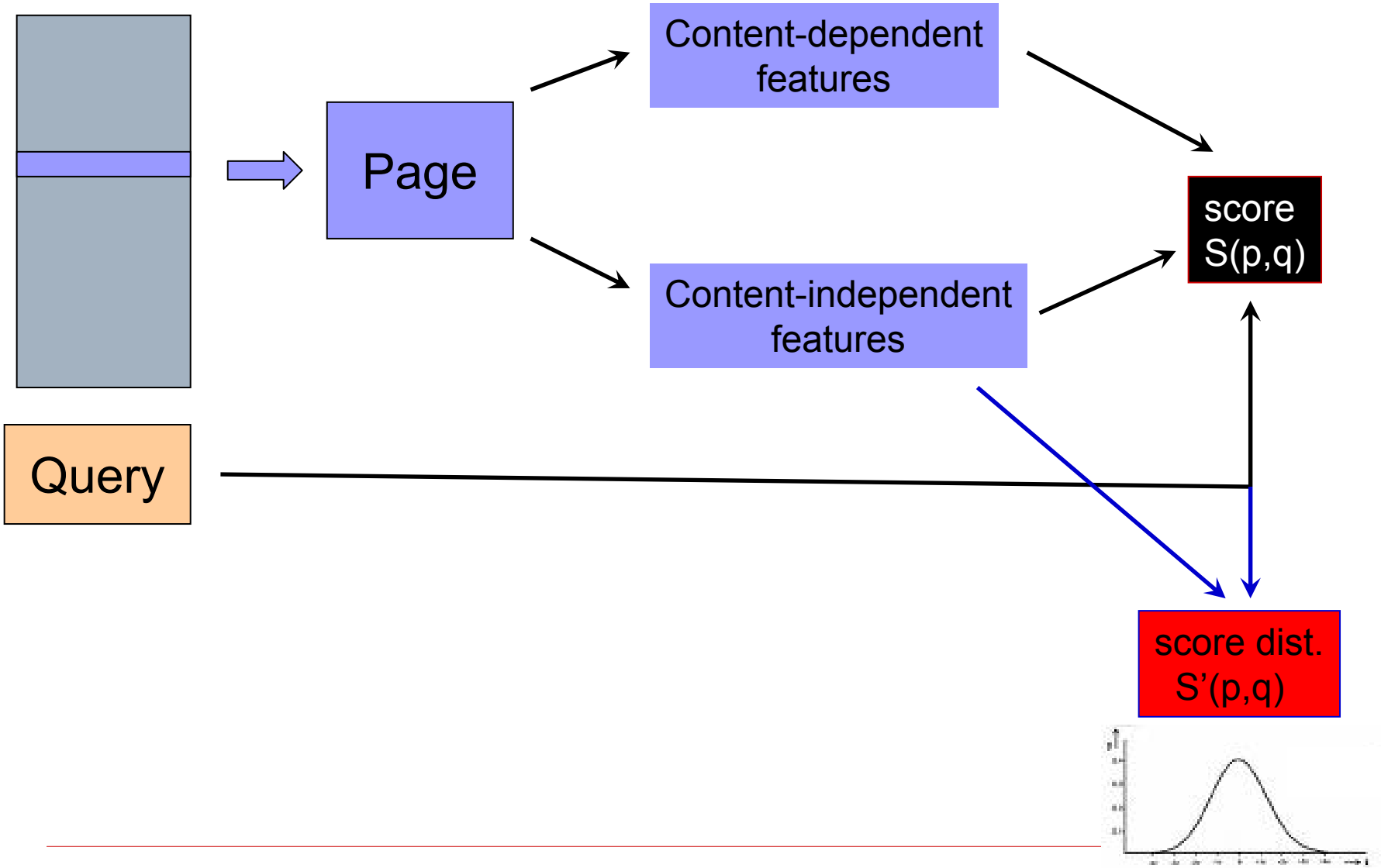
# Poor Correlation Between Prestige and Impact



# Outline

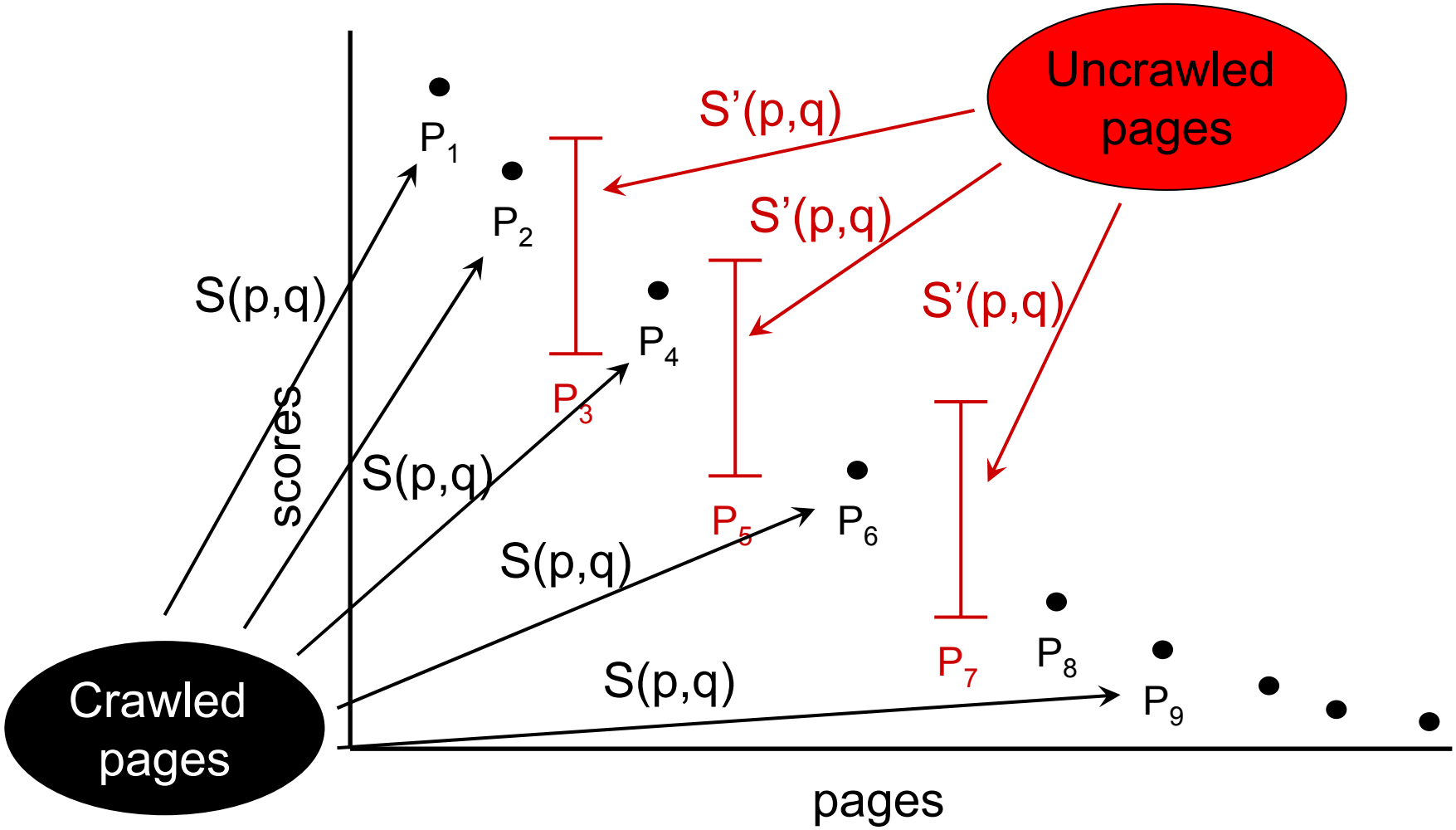
- Introduction
- Problem formulation and Complexity
- Our Approach
- Experiments

# Ranking Crawled Pages



# Ranking Crawled & Uncrawled Pages

- "Query sketch" for query q:

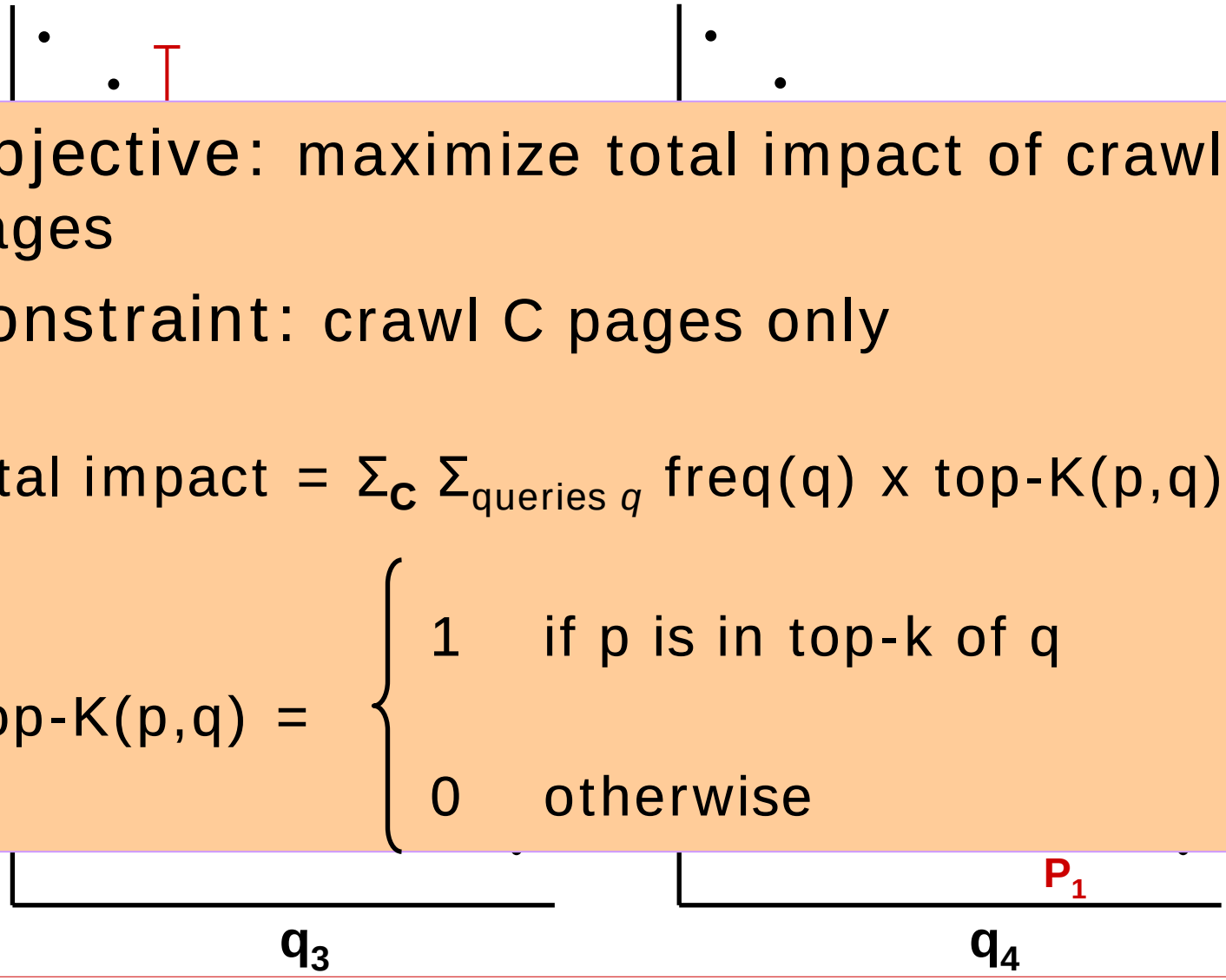


# Selecting Pages to Crawl

- Objective: maximize total impact of crawled pages
- Constraint: crawl C pages only

total impact = $\Sigma_{\mathbf{C}} \Sigma_{\text{queries } q}$ freq(q) x top-K(p,q)

$$\text{top-K}(p,q) = \begin{cases} 1 & \text{if p is in top-k of q} \\ 0 & \text{otherwise} \end{cases}$$



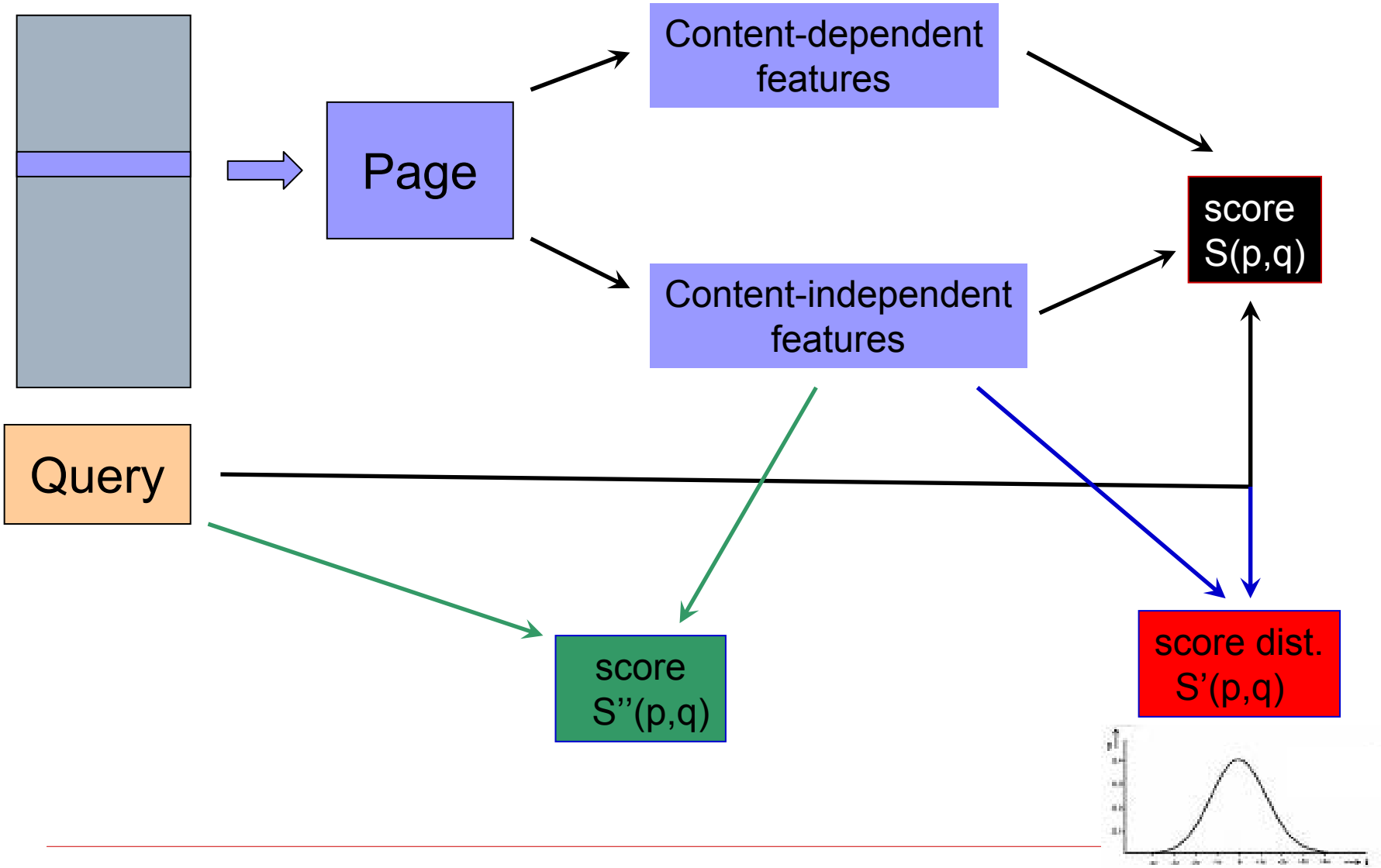
# Complexity

- Maximize worst-case impact:
  - NP-hard.
  - Reduction from densest k-vertex sub-hypergraph problem
- Maximize expected impact:
  - Polynomial but expensive

# Outline

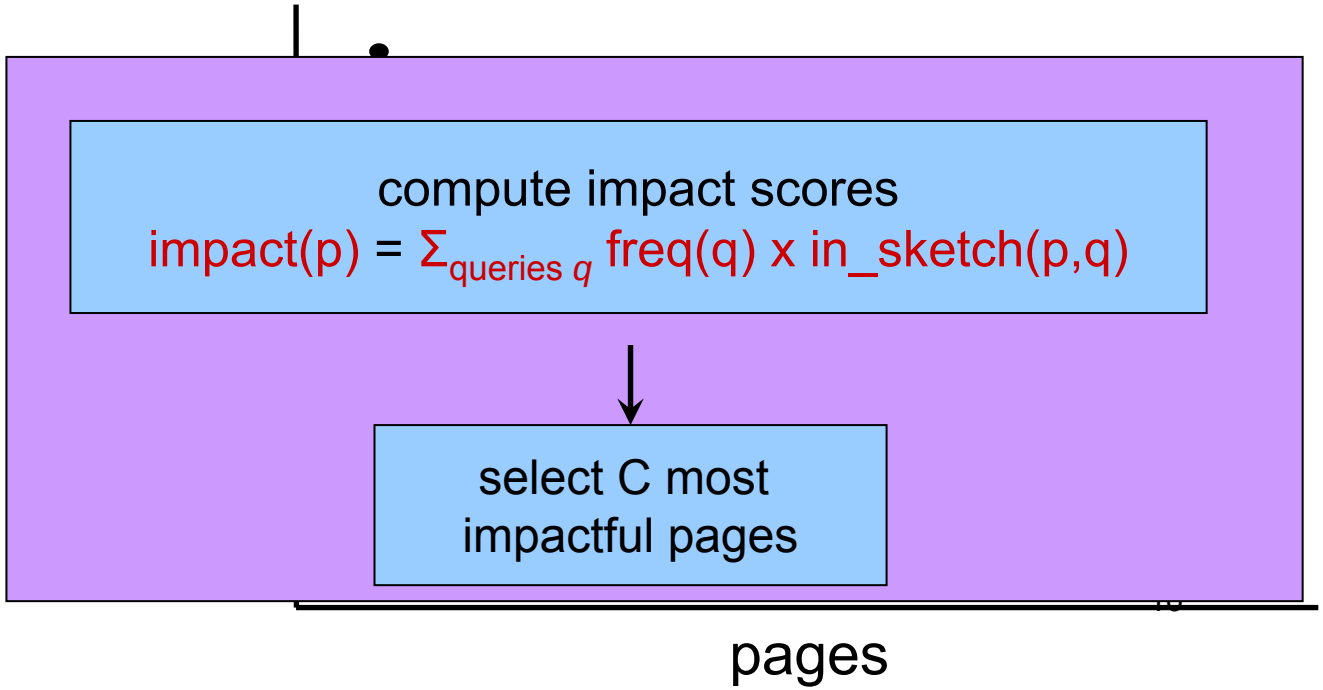
- Introduction
- Problem Formulation and Complexity
- Our Approach
- Experiments

# Relaxed Model



# Relaxed Model

- Revised query sketch (just top-K points):



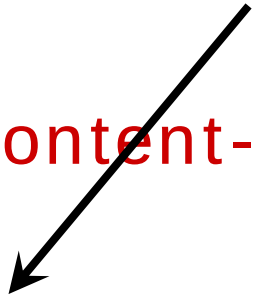
# Three Hiccups

1. Large number of query sketches

Solution: focus on queries where most impact can be had from crawling

2. Hard to anticipate exact query workload

3. Low recall from content-independent features



Solution:

1. estimate impact based on past workload
2. supplement impact estimation with prestige-based approach

# Solution 1: limit number of sketches

- Only create sketches for queries which could benefit from crawling additional pages (needy queries)
- 0.7% of queries -> most of benefit
- Depends on:
  - Current answer quality
    - Σ(score of current top-k results for q)
  - Quality of uncrawled relevant pages
    - estimate based on last crawl cycle

# Three Hiccups

1. Large number of query sketches

Solution: focus on queries where most impact can be had from crawling

2. Hard to anticipate exact query workload

3. Low recall from content-independent features

Solution:

1. estimate impact based on past workload
2. supplement impact estimation with prestige-based approach

# Solution 2: hybrid impact estimation

- 2 ways to estimate impact
  - Using past workload — Workload-based expert
  - Using prestige — Prestige-based expert
- Combine their estimations
  - linear weighted combination
  - Impact-based = 0.9 ; prestige-based = 0.1

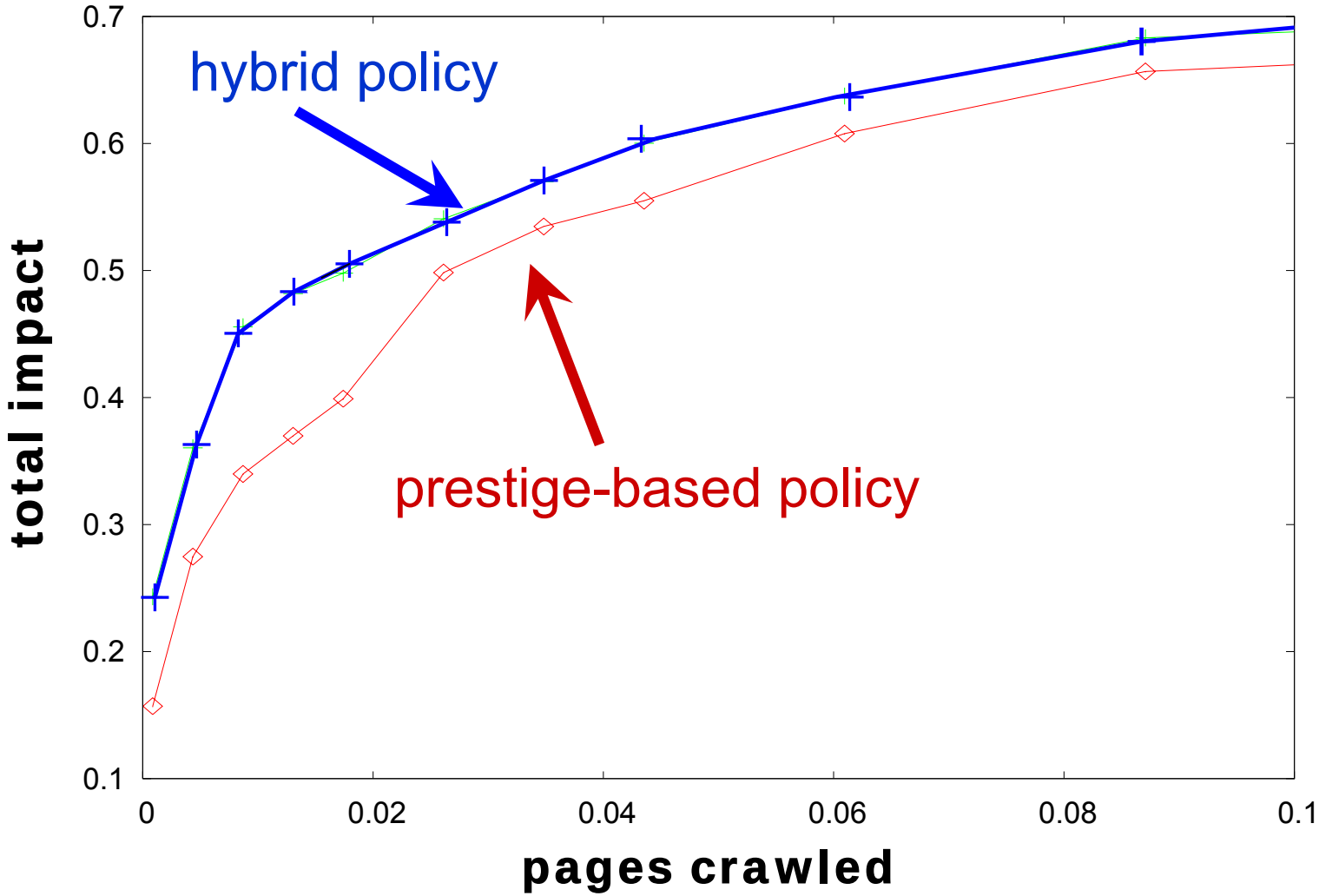
# Experiments

- Query workload: 5 day query log of a major search engine
- Scoring function: function used by that search engine

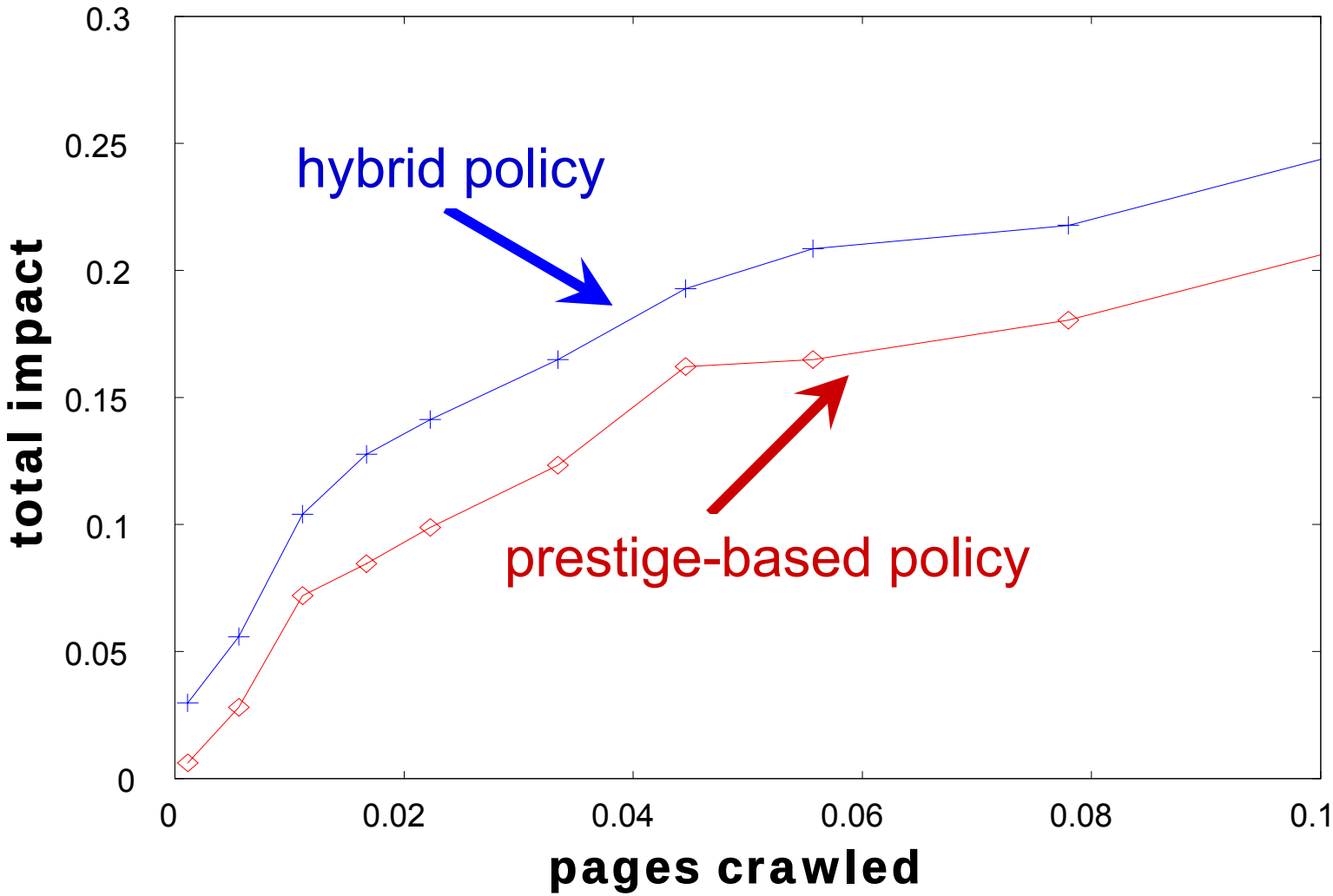
- Web page dataset 1:
  - Uncrawled pages:
    - Random sample of 110,000 pages
  - Crawled pages:
    - All other pages

- Web page dataset 2:
  - Move top 20% prestige pages to “crawled set”

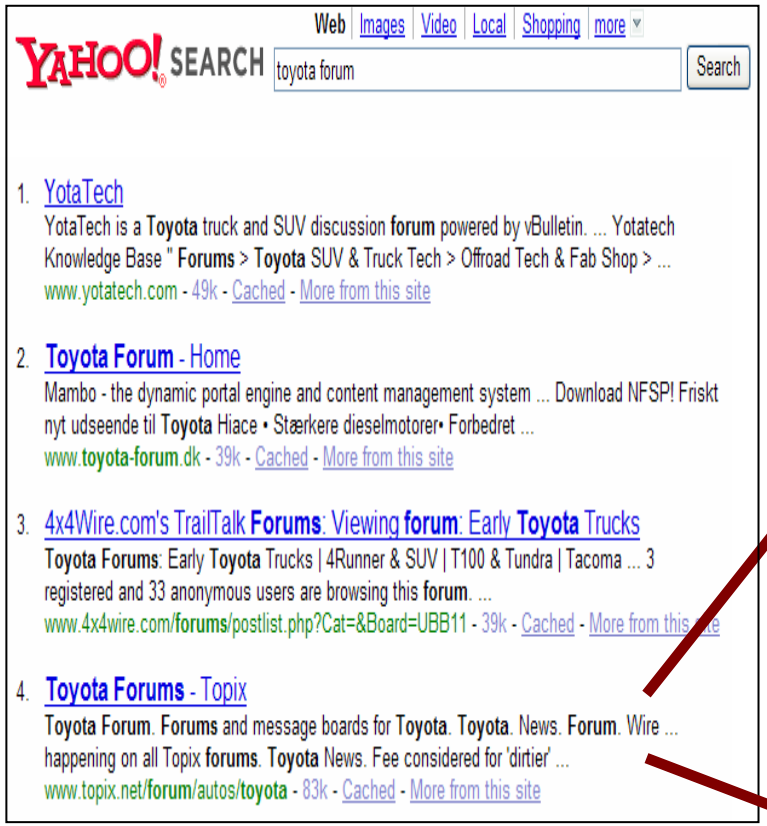
# Dataset 1 (w/all query sketches)

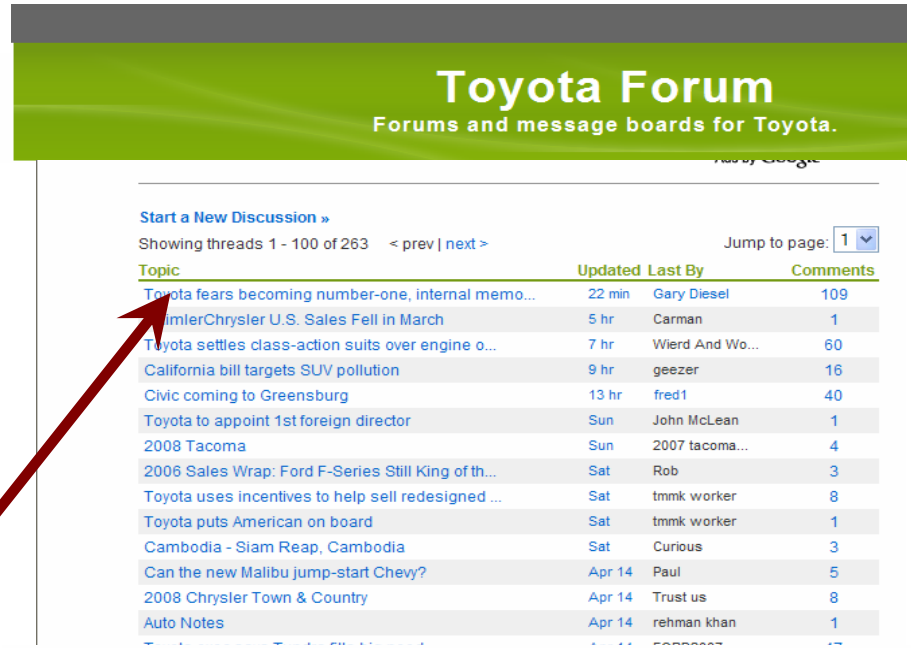


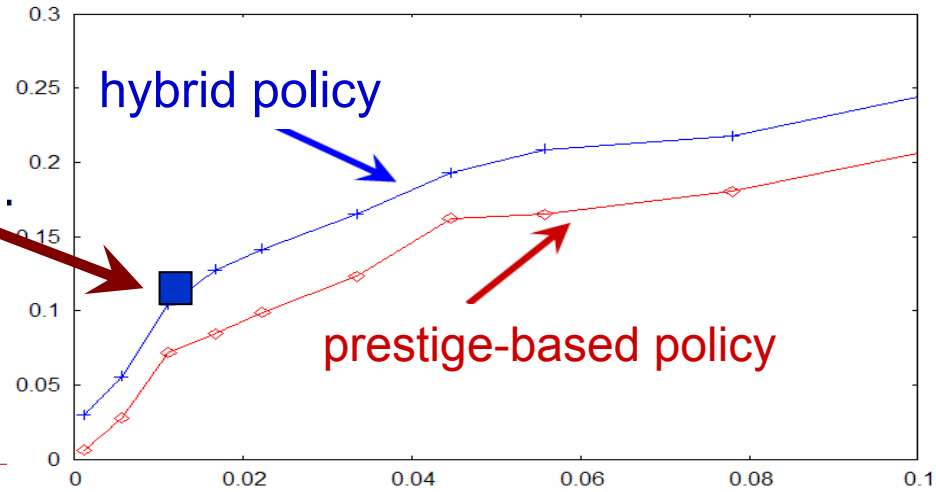
# Dataset 2 (w/all query sketches)



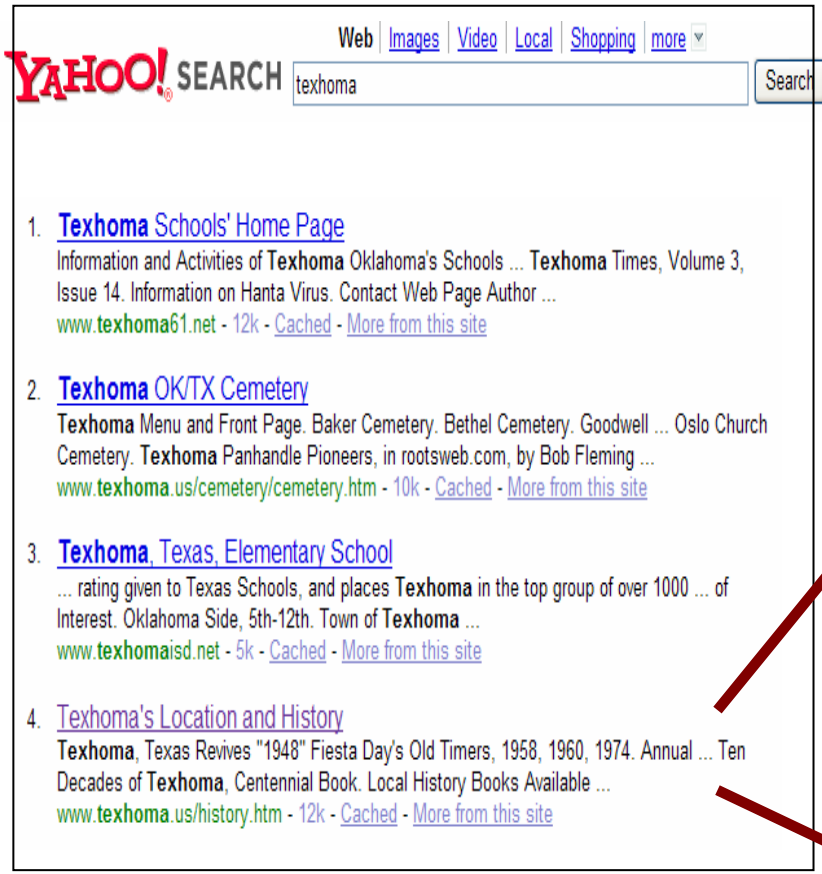
# Example 1

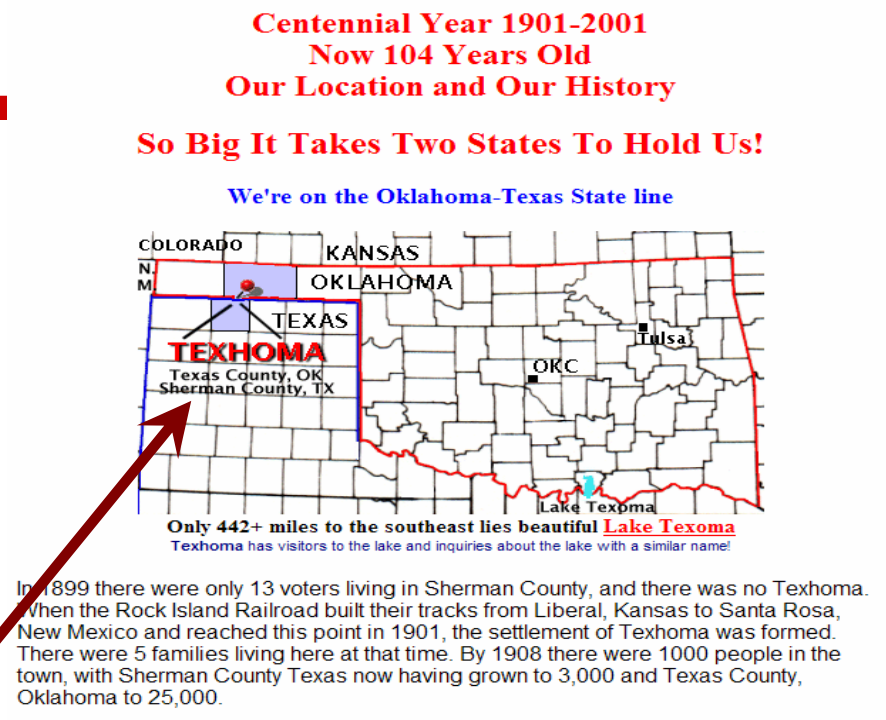

Web | Images | Video | Local | Shopping | more
YAHOO! SEARCH
toyota forum
Search
1. YotaTech
YotaTech is a Toyota truck and SUV discussion forum powered by vBulletin. ... Yotatech Knowledge Base " Forums > Toyota SUV & Truck Tech > Offroad Tech & Fab Shop > ...
www.yotatech.com - 49k - Cached - More from this site
2. Toyota Forum - Home
Mambo - the dynamic portal engine and content management system ... Download NFSP! Friskt nyt udseende til Toyota Hiace • Stærkere dieselmotorer• Forbedret ...
www.toyota-forum.dk - 39k - Cached - More from this site
3. 4x4Wire.com's TrailTalk Forums: Viewing forum: Early Toyota Trucks
Toyota Forums: Early Toyota Trucks | 4Runner & SUV | T100 & Tundra | Tacoma ... 3 registered and 33 anonymous users are browsing this forum. ...
www.4x4wire.com/forums/postlist.php?Cat=&Board=UBB11 - 39k - Cached - More from this site
4. Toyota Forums - Topix
Toyota Forum. Forums and message boards for Toyota. Toyota. News. Forum. Wire ... happening on all Topix forums. Toyota News. Fee considered for 'dirtier' ...
www.topix.net/forum/autos/toyota - 83k - Cached - More from this site

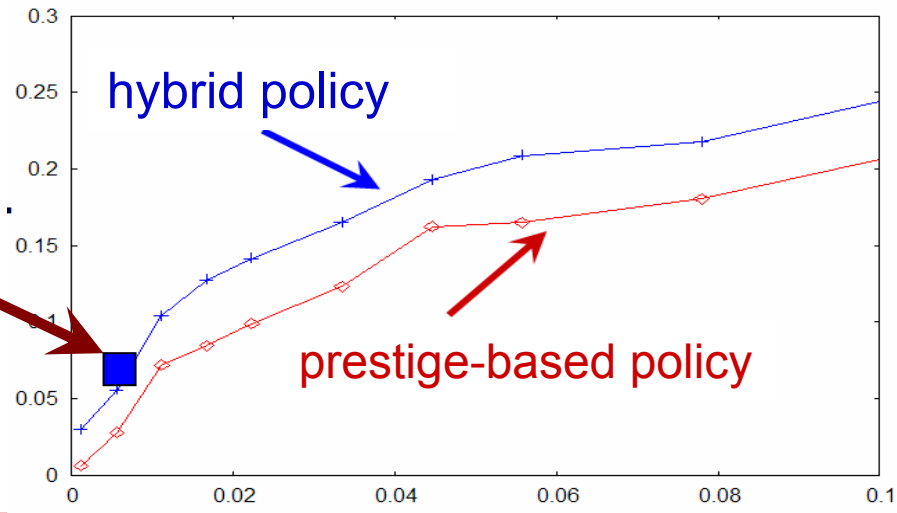

Toyota Forum
Forums and message boards for Toyota.
Start a New Discussion »
Showing threads 1 - 100 of 263 < prev | next > Jump to page: 1
Topic | Updated | Last By | Comments
Toyota fears becoming number-one, internal memo... | 22 min | Gary Diesel | 109
DaimlerChrysler U.S. Sales Fell in March | 5 hr | Carman | 1
Toyota settles class-action suits over engine o... | 7 hr | Wierd And Wo... | 60
California bill targets SUV pollution | 9 hr | geezer | 16
Civic coming to Greensburg | 13 hr | fred1 | 40
Toyota to appoint 1st foreign director | Sun | John McLean | 1
2008 Tacoma | Sun | 2007 tacoma... | 4
2006 Sales Wrap: Ford F-Series Still King of th... | Sat | Rob | 3
Toyota uses incentives to help sell redesigned ... | Sat | tmmk worker | 8
Toyota puts American on board | Sat | tmmk worker | 1
Cambodia - Siam Reap, Cambodia | Sat | Curious | 3
Can the new Malibu jump-start Chevy? | Apr 14 | Paul | 5
2008 Chrysler Town & Country | Apr 14 | Trust us | 8
Auto Notes | Apr 14 | rehman khan | 1


hybrid policy
prestige-based policy
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.02
0.04
0.06
0.08
0.1

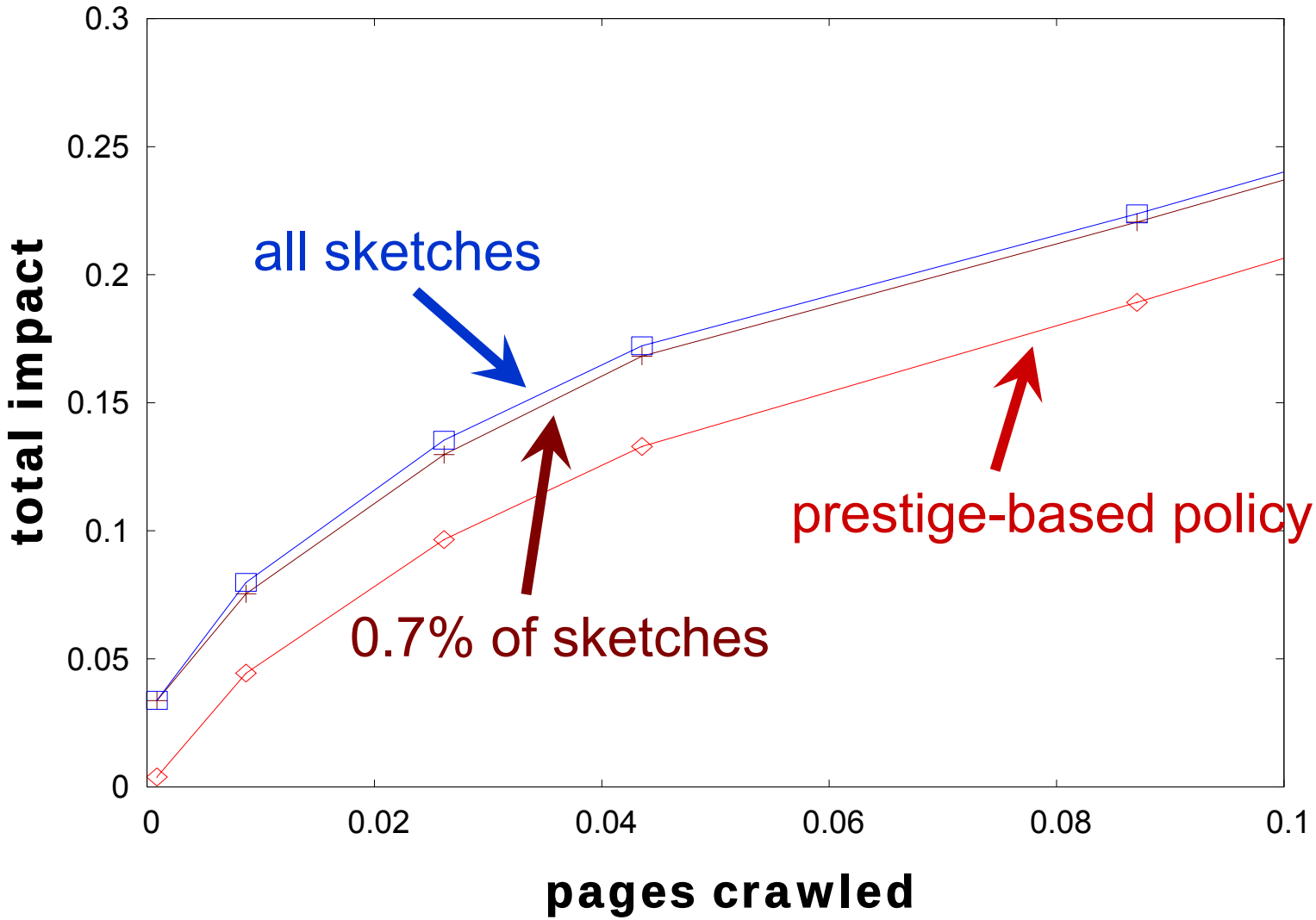
# Example 2


Web | Images | Video | Local | Shopping | more
YAHOO! SEARCH
texhoma
Search
1. Texhoma Schools' Home Page
Information and Activities of Texhoma Oklahoma's Schools ... Texhoma Times, Volume 3, Issue 14. Information on Hanta Virus. Contact Web Page Author ...
www.texhoma61.net - 12k - Cached - More from this site
2. Texhoma OK/TX Cemetery
Texhoma Menu and Front Page. Baker Cemetery. Bethel Cemetery. Goodwell ... Oslo Church Cemetery. Texhoma Panhandle Pioneers, in rootsweb.com, by Bob Fleming ...
www.texhoma.us/cemetery/cemetery.htm - 10k - Cached - More from this site
3. Texhoma, Texas, Elementary School
... rating given to Texas Schools, and places Texhoma in the top group of over 1000 ... of Interest. Oklahoma Side, 5th-12th. Town of Texhoma ...
www.texhomaisd.net - 5k - Cached - More from this site
4. Texhoma's Location and History
Texhoma, Texas Revives "1948" Fiesta Day's Old Timers, 1958, 1960, 1974. Annual ... Ten Decades of Texhoma, Centennial Book. Local History Books Available ...
www.texhoma.us/history.htm - 12k - Cached - More from this site


Centennial Year 1901-2001
Now 104 Years Old
Our Location and Our History
So Big It Takes Two States To Hold Us!
We're on the Oklahoma-Texas State line
COLORADO
N. M.
KANSAS
OKLAHOMA
TEXAS
TEXHOMA
Texas County, OK
Sherman County, TX
Tulsa
OKC
Lake Texoma
Only 442+ miles to the southeast lies beautiful Lake Texoma
Texhoma has visitors to the lake and inquiries about the lake with a similar name!
In 1899 there were only 13 voters living in Sherman County, and there was no Texhoma. When the Rock Island Railroad built their tracks from Liberal, Kansas to Santa Rosa, New Mexico and reached this point in 1901, the settlement of Texhoma was formed. There were 5 families living here at that time. By 1908 there were 1000 people in the town, with Sherman County Texas now having grown to 3,000 and Texas County, Oklahoma to 25,000.


hybrid policy
prestige-based policy
0.3
0.25
0.2
0.15
0.05
0
0
0.02
0.04
0.06
0.08
0.1

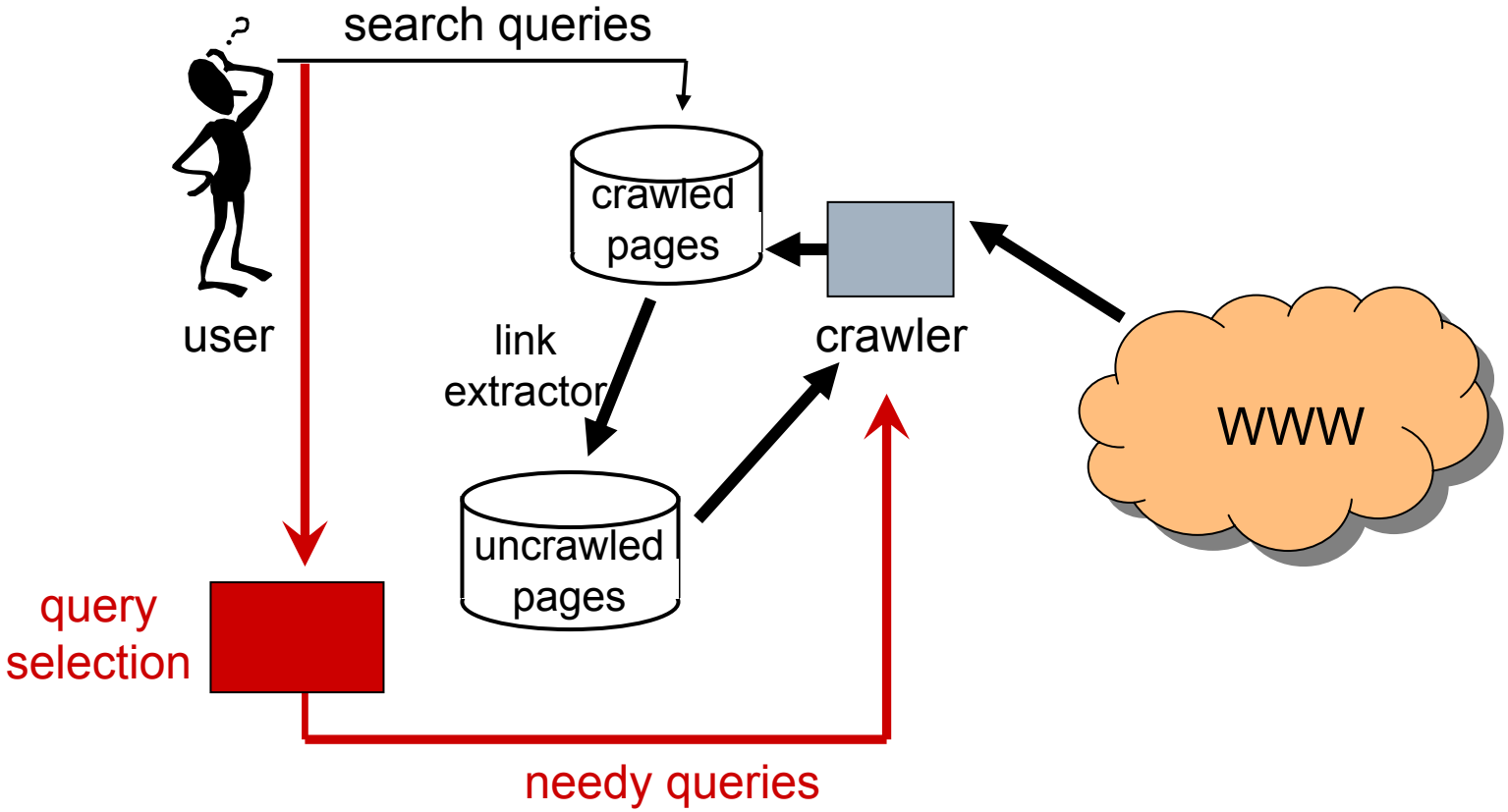
# Dataset 2 (w/0.7% of sketches)



# Related Work

- Discovering Unknown pages
  - Growth of the Web [Douglis et. al. USENIX'97, Fetterly et. al. WWW'03, Ntoulas et. al. WWW'04]
  - Discoverability [Dasgupta et. al. WWW'07]
- Crawling newly discovered pages
  - Breadth-first [Najork et. al. WWW'01], OPIC [Abiteboul et. al. WWW'03], PageRank [Cho et. al. WWW'98; Eiron et. al. WWW'04]
  - Focused Crawling [Chakrabarti et. al. WWW'99]
- Recrawling
  - Staleness-based [Cho et. al. SIGMOD'00], Embarassment-based [Wolf et. al. WWW'02], User-centric [Pandey et. al. WWW'05]

# The Big Picture



# THE END