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:

Objective: acquire the top part

US election  Super Bowl  Britney  Yahoo!
Impact of Crawling Page $p$

- **Impact**($p$) = $\sum_{\text{queries } q} \text{freq}(q) \times \text{top-K}(p,q)$

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

- **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-based priority list

impact-based priority list

URL: silverscape.com/.../Product_Positioning

bottom 20% of prestige; top 1% of impact ("product positioning")
prestige ≠ impact

prestige-based priority list

impact-based priority list

URL: pc2sms.eu

top 1% of prestige; low impact (relevant for “send free SMS”, but not in top-10)
Poor Correlation Between Prestige and Impact
Outline

- Introduction
- Problem formulation and Complexity
- Our Approach
- Experiments
Ranking Crawled Pages

Query

Page

Content-dependent features

Content-independent features

score $S(p,q)$

score dist. $S'(p,q)$
Ranking Crawled & Uncrawled Pages

“Query sketch” for query q:

- Crawled pages
- Uncrawled pages

Scores: S(p,q)

Pages: P1, P2, P3, P4, P5, P6, P7, P8, P9

S'(p,q)
Objective: maximize total impact of crawled pages

Constraint: crawl C pages only

\[
\text{total impact} = \sum_c \sum_{\text{queries } q} \text{freq}(q) \times \text{top-K}(p,q)
\]

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

Page

Content-dependent features

Content-independent features

Score $S(p,q)$

Score dist. $S'(p,q)$

Score $S''(p,q)$
Relaxed Model

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

\[
\text{compute impact scores} \\
\text{impact}(p) = \sum_{\text{queries } q} \text{freq}(q) \times \text{in\_sketch}(p,q)
\]

select C most impactful pages
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
  - Quality of uncrawled relevant pages

\[ \Sigma (\text{score of current top-k results for } q) \]

\[ \text{estimate based on last crawl cycle} \]
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
  - Using prestige

- 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)

- hybrid policy
- prestige-based policy
Dataset 2 (w/all query sketches)

- Linkflux-based policy
- Prestige-based policy
- Hybrid policy
Example 1

1. YotaTech
   YotaTech is a Toyota truck and SUV discussion forum powered by Bulletin. ... YotaTech Knowledge Base Forums > Toyota SUV & Truck Tech > Offroad Tech & Fab Shop > ... www.yotaech.com - 49k - Cached - More from this site

2. Toyota Forum - Home
   Mambbo - the dynamic portal engine and content management system ... Download NFSI Firkot ryt udeende til Toyota Hiace • Stærke dieselmotorener • 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=US11 - 39k - Cached - More from this site

4. Toyota Forums - Topic
   Toyota Forum. Forums and message boards for Toyota, Toyota, News. Forum. Wire ... happening on all Topix forums. Toyota News: Fee considered for dinner ... www.topix.net/forum/autos/toyota - 61k - Cached - More from this site

hybrid policy

prestige-based policy
Example 2

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. Bakor Cemetery, Bethel Cemetery, Goodwell ... Osso 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, 9th-12th. Town of Texhoma ... www.texhomaisd.net - 5k - Cached - More from this site

4. Texhoma’s Location and History

Hybrid policy

Prestige-based policy

So Big It Takes Two States To Hold Us!
We’re on the Oklahoma-Texas State line

Centennial Year 1901-2001
Now 104 Years Old
Our Location and Our History

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 1500 people in the town, with Sherman County Texas now having grown to 3,000 and T...

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.
Dataset 2 (w/0.7% of sketches)

- all sketches
- prestige-based policy
- 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

search queries

user

crawled pages

link extractor

crawler

WWW

query selection

needy queries
THE END