Geographically Focused Collaborative Crawling

Hyun Chul Lee
University of Toronto
&
Genieknows.com

Joint work with
Weizheng Gao (Genieknows.com)
Yingbo Miao (Genieknows.com)
Outline

Introduction/Motivation
  - Crawling Strategies
  - Evaluation Criteria
  - Experiments
  - Conclusion
Evolution of Search Engines

- Due to a large number of web users with different search needs, the current search engines are not sufficient to satisfy all needs.
- Search Engines are being evolved!
- Some possible evolution paths:
  - Personalization of search engines
    - Examples: MyAsk, Google personalized search, My Yahoo Search, etc.
  - Localization of search engines
    - Examples: Google Local, Yahoo Local, Citysearch, etc.
  - Specialization of search engines
    - Examples: Kosmix, IMDB, Scirus, Citeseer, etc.
  - Others (Web 2.0, multimedia, blog, etc)
Yahoo estimates that 20-25% of all search queries have a local component either stated explicitly (e.g. Home Depot Boston; Washington acupuncturist) or implicitly (e.g. flowers, doctors).

Source: SearchengineWatch Aug 3, 2004
Local Search Engine

- **Objective:** Allow the user to perform the search according to his/her keyword input as well as the geographical location of his/her interest.

- Location can be
  - City, State, Country.
    - E.g. Find restaurants in Los Angeles, CA, USA
  - Specific Address
    - E.g. Find starbucks near 100 millam street houston, tx
  - Point of Interest
    - E.g. Find restaurants near LAX.
Web Based Local Search Engine

- The precise definition of what Local Search Engine is not possible as there are numerous Internet YellowPages (IYP) that claim to be local search engines.
- Certainly, a true Local Search Engine should be **Web based**.
- Crawling of deep web data (geographically-relevant) is also desirable.
Motivation for geographically focused crawling

- 1st step toward building a local search engine is to collect/crawl geographically sensitive pages.
- There are two possible approaches

1. General Crawling
2. Filter out pages that are not geographically-sensitive.

1. Target those pages which are geographically-sensitive during crawling

We study this problem as part of building Genieknows local search engine.
Outline

- Introduction/Motivation
  - Crawling Strategies
  - Evaluation Criteria
  - Experiments
  - Conclusion
Problem Description

- **Geographically Sensitive Crawling:** Given a set of targeted locations (e.g. list of cities), collect as many pages as possible that are geographically-relevant to the given locations.

- For simplicity, for our experiment, we assume that the targeted locations are given in forms of city-state pairs.
Basic Assumption
(Example)

Pages about Boston

Non-relevant pages

Pages about Houston
Basic Strategy
(Single Crawling Node Case)

- Exploit features that might potentially lead to the desired geographically-sensitive pages
- Guide the behavior of crawling using such features.

- Note that similar ideas are used for topical focused crawling.

**Target Location: Boston**

- Crawl this URL
  - Extracted URL 1
    - (www.restaurant-boston.com)
  - Extracted URL 2
    - (www.restaurant.com)
Extension to the multiple crawling nodes case

WEB

<table>
<thead>
<tr>
<th>Boston</th>
<th>Chicago</th>
<th>Houston</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 Geographically-Sensitive Nodes</td>
<td>G2 General Crawling Nodes</td>
<td>G3 General Crawling Nodes</td>
</tr>
<tr>
<td>C1 General Crawling Nodes</td>
<td>C2 General Crawling Nodes</td>
<td></td>
</tr>
</tbody>
</table>
Extension to the multiple crawling nodes case (cont.)

URL 1 (about Chicago)

URL 2 (Not geographically-sensitive)

Geographically-Sensitive Crawling Nodes

Boston G1 Chicago G2 Houston G3

General Crawling Nodes

C1 C2
Extension to the multiple crawling nodes case (cont.)

Geographically-Sensitive Crawling Nodes
- Boston
- Chicago
- Houston

General Crawling Nodes
- C1
- C2

Extracted URL (Houston)
Crawling Strategies
(URL Based)

Does the considered URL contain the targeted city-pair A?

Yes

Assign the corresponding URL to the crawling node responsible for the city-pair A
Crawling Strategies
(Extended Anchor Text)

- Extended Anchor Text refers to the set of prefix and suffix tokens to the link text.
- When multiple findings of city-state pairs occur, then choose the closest one to the link text.

Does the considered Extended Anchor Text contain the targeted city-pair A?

Yes

Assign the corresponding URL to the crawling node responsible for the city-pair A.
Crawling Strategies (Full Content Based)

1. Compute the probability that a page is about a city-state pair using the full content

2. Assign all extracted URLs to the crawling node responsible for the most probable city-state pair

- Consider the number of times that the city-state pairs is found as part of the content
- Consider the number of times that only city-name is found as part of the content
Crawling Strategies (Classification Based)

1. Classifier determines whether a page is relevant to the given city-state pair

2. Assign all extracted URLs to the crawling node responsible for the most probable city-state pair

- Shown to be useful for the topical collaborative crawling strategy (Chung et al., 2002)
- Naïve-Bayes classifier was used for simplicity.
- Training data were obtained from DMOZ.
Crawling Strategies (IP-Address Based)

1. Associate the IP-address of the web host service with the corresponding city-state pair

2. Assign all extracted URLs to the crawling node responsible for the obtained city-state pair

- For the IP-address mapping tool, hostip.info (API) was employed.
Normalization and Disambiguation of city names

- From previous researches (Amitay et al. 2004, Ding et al. 2000)
  - **Aliasing:** Problem of different names for the same city.
    - United States Postal Service (USPS) was used for this purpose.
  - **Ambiguity:** City names with different meanings.
    - For the full content based strategy, solve it through the analysis of other city-state pairs found within the page
    - For the rest of strategy, simply assume that it is city name with the largest population size.
Outline

- Introduction/Motivation
- Crawling Strategies
- Evaluation Criteria
- Experiments
- Conclusion
Evaluation Model

- **Standard Metrics (Cho et al. 2002)**
  - **Overlap**: $N-I/N$ where $N$ refers to the total number of downloaded pages by the overall crawler and $I$ denotes the number of unique downloaded pages by the overall crawler.
  - **Diversity**: $S/N$ where $S$ denotes the number of unique domain names of downloaded pages by the overall crawler and $N$ denotes the total number of downloaded pages by the overall crawler.
  - **Communication Overlap**: Exchanged URLs per downloaded page.
Geographically Sensitive Metrics

- Use extracted geo-entities (address information) to evaluate.
- **Geo-Coverage:** Number of pages with at least one geo-entity.
- **Geo-Focus:** Number of retrieved pages that contain at least one geographic entity to the assigned city-state pairs of the crawling node.
- **Geo-Centrality:** How central are the retrieved nodes relative to those pages that contain geographic entities.
Outline

- Introduction/Motivation
- Crawling Strategies
- Evaluation Criteria
- Experiment
- Conclusion
Experiment

- Objective: Crawl pages pertinent to the top 100 US cities
- Crawling Nodes: 5 Geographically Sensitive Crawling nodes and 1 General node
- 2 Servers (3.2 GHz dual P4s, 1 GB RAM)
- Around 10 million pages for each crawling strategy
- Standard Hash-Based Crawling was also considered for the purpose of comparison
Result
(Geo-Coverage)
Result
(Geo-Focus)
Result
(Communication-Overhead)

![Bar chart showing different drawing strategies and their communication overhead.]

- URL-Hash Based
- URL-Based
- Extended Anchor Text Based
- Full Content Based
- Classification Based
- IP-Address Based
Outline

- Introduction/Motivation
- Crawling Strategies
- Evaluation Criteria
- Experiment

Geographic Locality
Geographic Locality

- **Question:** How close (graph based distance) are those geographically-sensitive pages?
- \( \delta(\Omega) \): the probability that a pair of linked pages, chosen uniformly at random, is pertinent to the same city-state pair under the considered collaborative strategy.
- \( \overline{\delta}(\Omega) \): the probability that a pair of un-linked pages, chosen uniformly at random, is pertinent to the same city-state pair under the considered collaborative strategy.
# Results

<table>
<thead>
<tr>
<th>Crawling Strategy</th>
<th>$\delta(\Omega)$</th>
<th>$\bar{\delta}(\Omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL Based</td>
<td>0.41559</td>
<td>0.02582</td>
</tr>
<tr>
<td>Classification Based</td>
<td>0.044495</td>
<td>0.008923</td>
</tr>
<tr>
<td>Full Content Based</td>
<td>0.26325</td>
<td>0.01157</td>
</tr>
</tbody>
</table>
Conclusion

- We showed the feasibility of geographically focused collaborative crawling approach to target those pages which are geographically-sensitive.
- We proposed several evaluation criteria for geographically focused collaborative crawling.
- Extended anchor text and URL are valuable features to be exploited for this particular type of crawling.
- It would be interesting to look at other more sophisticated features.
- There are lots of problems related to local search including ranking, indexing, retrieval, crawling