Benchmarking Current Deep Web Crawlers
Amruta Chavan (agc9066@rit.edu)  Advisor: Dr. Carlos Rivero
Department of Computer Science ,Rochester Institute of Technology

Introduction

- A web page shows up in the results of a search engine only if it can be indexed. Crawlers are used for traversing to different web pages and indexing them.
- Traditionally, crawlers have only targeted a portion of the Web called the publicly indexable Web (PIW).
- However, a large amount of data available on is accessible only by filling out request forms. This data is called hidden data. Since there are no static links to the Hidden Web pages, search engines cannot discover and index such pages and thus do not return them in the results.
- The aim of this project is to implement a crawler which can retrieve the hidden data in an automated manner.

Background

- The hidden web crawlers employ a mechanism for automatic form filling, navigation and result classification. Since these web crawlers can access the hidden web content, this content can be indexed by the web engine and hence shows up in the search result.

Design

- The diagram illustrates the process of indexing the hidden Web data using a Random Keyword Selection crawler.
- The crawler extracts links and adds them to the list, visits each page, and fills the forms to access the hidden data.
- The retrieved data is then stored for future use.

Result

- The table shows the coverage of different domains.
- The coverage is calculated as links extracted / total number of links.

<table>
<thead>
<tr>
<th>Sample keyword classification</th>
<th>High Coverage</th>
<th>Medium Coverage</th>
<th>Low Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>diabetes</td>
<td>(390/1000)</td>
<td>endocrinology (49/1000)</td>
<td>dentistry (3/1000)</td>
</tr>
<tr>
<td>transplantation</td>
<td>(166/1000)</td>
<td>myocardial (29/1000)</td>
<td>neurology (2/1000)</td>
</tr>
<tr>
<td>cardiovascular</td>
<td>(191/1000)</td>
<td>flu (40/1000)</td>
<td>radiotherapy (2/1000)</td>
</tr>
</tbody>
</table>

Conclusion

- Implemented a Random Keyword selection crawler which provided a mechanism to crawl hidden data in websites in an automated manner.
- New features like checking robots.txt, politeness policy and option for link classification method were added to the original crawler design.
- Keywords were classified as high coverage, medium coverage and low coverage depending on the results links they returned.

Future Work

- Merge the current web crawler implementation and CALA to auto-generate the URL patterns.
- Include more fields in the automated form filling phase.

References

- Stephen W. Liddle, David W. Embrey, Del T. Scott "Extracing Data Behind Web Forms".
- CALA is...