Network of Epidemiology Digital Objects

Naren Sundar, Kui Xu
Client: Sandeep Gupta, S.M. Shamimul
CS6604 Class Project
Outline

• Problem Statement
• Requirement Specification
• Related work
• What a human would do?
• Modeling
• Evaluation
Client Problem Statement

• CINET
  – Computational and analytic environment for network science research and education.

• Goal: A RDF graph building service
  – Web crawling for contents related to epidemiology
  – RDF representation to interconnect digital objects
Client Problem Statement

• Requirement: Build a connected network of:
  – Papers
  – Wiki pages
  – Websites
  – Videos
  – Other digital objects pertaining to epidemiology

• Representing using RDF for future graph analysis.
Refined Problem Statement

• Strongly connected digital objects & metadata
  – Include metadata to model the connection
  – Better represent the underlying correlation among digital objects

• Given a search request, provide resulting DO network:
  – Include all related digital objects
  – Strongly connected DOs are more related
Refined Problem Statement

- Automated process
  - Web exploration
  - Network construction
- Restricted digital objects
  - Research papers
- Web crawling (search engine from dblp)

**CompleteSearch DBLP**

A DBLP mirror with extended search capabilities maintained by Hannah Bast, University of Freiburg (formerly MPII Saarbrücken).

Zoomed in on 283 documents ... NEW: get these search results as XML, JSON, JSONP.

<table>
<thead>
<tr>
<th>Year</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>Joshua Tan, Khanh Nguyen, Michael Theodorides, Heidi Negrón-Arroyo, Christopher Thompson, Serge Egelman, David Wagner: The effect of developer-specified explanations for permission requests on smartphone user behavior. CHI 2014:91-100</td>
</tr>
</tbody>
</table>
Meta-data for DO

- Authors, key words, publisher, year
- Given papers \{Paper1\} \{Paper2\} \{Paper3\}
  - P1: author1, author2, publisher 1, 2011, \{keywords\}
  - P2: author2, author3, publisher 2, 2012, \{keywords\}
  - P3: author 3, publisher 2, 2011, \{keywords\}
- The resulting network through the meta-data:
Requirement Specification

• Undirected crawling or normal search engine not sufficient:
  – Results un-organized
  – Little specialization
  – Ambiguity
  – Relations and connections not available
Requirement Specification

What’s available

What we want
Related Work

• Web crawling topics
  – Building efficient, robust and scalable crawler
  – Traversal order of the web graph
  – Re-visititation of previously crawled content
  – Avoid problematic and undesirable content
  – Crawling “deep web” content

Why can a human do this better?

A user is interested in “Disease propagation”

1. Set $N = \{\text{Disease propagation}\}$
2. Search using $N$ and gets $R$
3. Splits $R$ into $A$ and $B$
4. Selects $A$ as relevant; extracts terms and connections; augments $N$
5. Ignores or stashes away $B$
6. Repeat from 1 with updated network $N$
Desired Property 1: Connected Growth

- A relevant paper shares nodes and edges in the network N
- Not all of the paper becomes relevant at once
- What is shared changes over time
Desired Property 2: Grouping

- Only a small fraction of a document is shared with network N
- The rest is used to group papers

Paper1
Paper2
Paper3
Paper4
Paper5
Modeling: Digital Object

• Is represented as a set of edges
  – $D_i = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

• Connections are determined by predicates over kind of vertex
  – $(x, y)$: Paper $p$, authored by $x$, is published in $y$

• From a generative point-of-view, an edge in a paper comes from
  – A shared network
  – Or, a local network
Modeling: Split-View of Digital Object

- A paper’s edges can be split into two parts
  - $D_i = S_i + E_i$
  - Shared set: $S_i$
  - Local set: $E_i$
Modeling: Grouping

- Each $D_i$ belongs to a group
- Groups are determined w.r.t. unshared edges $E_i$
- No need to determine number of groups beforehand

$$P(k | \alpha) \prod_{(x, y) \in E_i} P(x, y | G_k)$$
Modeling: Network

• Connectivity
  – Allow \((x,y)\) only if \(x\) is \underline{reachable} from root nodes

• Smoothness
  – Allow \((x,y)\) only if path connecting to \(x\) from root has \underline{decreasing weights}

Edge is either shared or local based on

\[
\arg\max_{\varepsilon} P(x, y | \varepsilon)
\]

\(\varepsilon \in \{N, G_k\}\)

satisfying connectivity and smoothness constraints when shared.
# The Entire Process

<table>
<thead>
<tr>
<th>Model</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Start with root nodes, e.g., “Disease propagation” in N</td>
<td>1. Set N = {Disease propagation}</td>
</tr>
<tr>
<td>2. Generate search query from N</td>
<td>2. Search using N and gets R</td>
</tr>
<tr>
<td>3. Crawl and add new digital objects</td>
<td>3. Splits R into A and B</td>
</tr>
<tr>
<td>4. Learn random variables for sharing (Si, Ei) and grouping (Gi)</td>
<td>4. Selects A as relevant; extracts terms and connections; augments N</td>
</tr>
<tr>
<td>5. Goto step 2</td>
<td>5. Ignores or stashes away B</td>
</tr>
<tr>
<td></td>
<td>6. Repeat from 1 with updated network N</td>
</tr>
</tbody>
</table>
Evaluation: Against a Search Engine

• **Goal**: Relevance
• Set a starting topic
• Perform k related queries
  – Let \( R_i \) be the sets of query results ranked by the search engine
  – Learn network \( N \)
  – For each paper \( p \) selected by \( N \)
    • Get best rank of \( p \) according to \( R_i \)
    – Compare ranks against relevance determined by \( N \)
• Are there lowly ranked results that are higher in \( N \) and vice versa?
Evaluation: Against a Digital Library

• **Goal:** Completeness

• Pick a digital library for digital objects with a categorical browsing service

• Pick a starting point
  
  – Learn network N

  – Evaluate completeness of network N against the digital library
Thank You!

Questions?