Doozer
An Expand and Reduce approach to Automatic domain model creation

Christopher Thomas, Pankaj Mehra, Roger Brooks, Amit Sheth
Motivation

What will you want to know tomorrow?
Wisdom of the Crowds

- The most comprehensive and up to date account of the present state of knowledge is given by

  Everybody

  = The Web in general
  = Blogs
  = Wikipedia
Collecting Knowledge

- Wikipedia
  + Concise concept descriptions
  + Large, highly connected, sparsely annotated graph structure that connects named entities
  + Category hierarchy
Goal: Harness the Wisdom of the Crowds to Automatically define a domain with up-to-date concepts
Don’t overshoot

• Generating a **broad** topic hierarchy on the basis of Wikipedia is **easy**

• The Problem is to **restrict** it to the actual domain of interest
Simplifying assumptions

• We do **not** create a formal domain ontology
• Topic hierarchy will be used for classification and IR, not for reasoning
… Hence

- We can safely take advantage of existing (semi)structured knowledge sources
Find pertinent terms
Classify the terms in a topic hierarchy
Key Concepts

• Focus
  – What is the narrow domain of interest?
  – E.g. FTP Servers

• Domain
  – What is the broader domain of interest?
  – E.g. Networking, File Transfer

• World View
  – Where is our world view centered?
  – E.g. information sciences
  – Different World View ➔ Different connections
Overview - conceptual

- Expand and Reduce approach
  - Start with ‘high recall’ methods
    - Exploration - Full text search
    - Exploitation – Node Similarity Method
    - Category growth
  - End with “high precision” methods
    - Apply restrictions on the concepts found
    - Remove unwanted terms and categories
Links types

- “See Also” link
- Inverse “See Also” link
- Double link
- Link and common category
- Inverse link and common category
- Regular link
- Inverse link
Node Similarity computation

\[
s(a, b) = \frac{\sum_{\{i, j \mid \langle N_i(a), N_j(b) \rangle \in M\}} \text{avg}(w(N_i(a)), w(N_j(b)))}{\text{avg}\left(\sum_{i} w(N_i(a)), \sum_{j} w(N_j(b))\right)}
\]
Collecting Instances

Semantic Similarity Method

Full Text Search

Query: "server* compute* informati* internet www ftp"

FTS

root nameserver
DNN root zone
Server (computing)

NSD
AlterNIC
EDNS
DNS zone
ICANN
Internet

Wikipedia
Creating a Hierarchy

Seed Query --> Fulltext Concept Search (Simmath) --> Wikipedia

Graph Search (Denis) --> Graph Search (Denis) --> Graph Search (Denis)
Hierarchy Creation - summary

Knowledge Enabled Information and Services Science
Reduce steps

- Remove all terms that have low pertinence to the domain
- Intersect hierarchy with broader focus domain
- Reduce hierarchy depth
Remove unwanted individuals
Remove unwanted categories
Flatten categories
Snapshot of final Topic Hierarchy
### Evaluation wrt. Search and Set Expansion

<table>
<thead>
<tr>
<th></th>
<th>Top</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google Sets</strong></td>
<td>mortgage</td>
<td>overview</td>
<td>contact brochures</td>
<td>photos</td>
<td>personal finance</td>
</tr>
<tr>
<td><strong>Groxis Grokker</strong></td>
<td>Tort law</td>
<td>Lost, mislaid, and abandoned property</td>
<td>Mortgage loan</td>
<td>Law of Property Act 1925 (c.20...</td>
<td>Leasehold estate</td>
</tr>
<tr>
<td><strong>Doozer</strong></td>
<td>Mortgage Loan</td>
<td>Mortgage</td>
<td>Teachers' Building Society</td>
<td>Mortgage-backed Security</td>
<td>Agency Securities</td>
</tr>
<tr>
<td><strong>Powerset</strong></td>
<td>Mortgage loan</td>
<td>Mortgage broker</td>
<td>Mortgage underwriting</td>
<td>2007 Subprime mortgage financial crisis</td>
<td>Mortgage discrimination</td>
</tr>
<tr>
<td><strong>Wikimedia</strong></td>
<td>Mortgage</td>
<td>Mortgage loan</td>
<td>Mortgage-backed security</td>
<td>Mortgage discrimination</td>
<td>Mortgage discrimination</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google Sets</strong></td>
<td>foreclosures</td>
<td>houses for sale</td>
<td>real estate training</td>
<td>people search</td>
<td>mortgage lendingtree</td>
</tr>
<tr>
<td><strong>Groxis Grokker</strong></td>
<td>Non-possessory interest in land</td>
<td>Bank</td>
<td>Non-recourse debt</td>
<td>Lateral and subjacent support</td>
<td>Mortgage</td>
</tr>
<tr>
<td><strong>Doozer</strong></td>
<td>Uk Mortgage Terminology</td>
<td>Mortgage Note</td>
<td>Adjustable Rate Mortgage</td>
<td>Graduated Payment Mortgage Loan</td>
<td>Interest-only Loan</td>
</tr>
<tr>
<td><strong>Powerset</strong></td>
<td>Shared appreciation mortgage</td>
<td>Mortgage life insurance</td>
<td>Biweekly Mortgage</td>
<td>Mortgage</td>
<td>Foreign currency mortgage</td>
</tr>
<tr>
<td><strong>Wikimedia</strong></td>
<td>Mortgage broker</td>
<td>Mortgage bank</td>
<td>Adjustable rate mortgage</td>
<td>Lenders mortgage insurance</td>
<td>Ameriquest Mortgage</td>
</tr>
</tbody>
</table>
F-measures, computed against a reduced glossary, for the lists of terms generated by various mining tools.
Evaluation wrt. expert model

Evaluation wrt. MeSH versions of 2004 (04) and 2008 (08) for both the restricted and the full set of MeSH term
Thank you