The Semantic Growbag – Automatically Organizing Topic Facets

Jörg Diederich, Uwe Thaden, Wolf-Tilo Balke,

Faceted Search Workshop at SIGIR’06
Seattle, Washington, USA
Outline

- Motivation
- Semantic Growbag
  - Algorithm
- Experimental Setup
- Demonstrator
- Summary and Future Work
Motivation

- Faceted search groups objects (Web pages, documents, etc.) into independent categories for navigational access
  - Users might focus on or remember different aspects when trying to locate information
  - Keywords of categories do not always exist in documents
  - Example: author, publication year, journal,...

- Problem in faceted search:
  - How to organize highly dynamic facets with many instances?
  - Example: “Topic” in a faceted browser for publications
Motivation

- Hierarchies in individual facets
  - Taxonomies
  - But: facet-specific, user-specific, time-variant,...

- Idea: Semantic GrowBag Algorithm
  - Organize a facet comprising the tags from a folksonomy-based corpus of objects
    - Example: Publications annotated with keywords
    - Derive community-driven topic hierarchies automatically
    - Use them to better organize the topic hierarchy
  - Organization dependent on *when* the tagging took place
    - Show development over time (if tagging associated with a date)
Example: Topic Hierarchy

 Legacy topic facet:
- Search Engine (10)
- Vector Space Model (3)
- Link Analysis (4)
- Crawler (3)
- [+ 100's more....]

 Hierarchically organized topic facet:
- **Search Engine (10)**
  - Vector Space Model (3)
  - Link Analysis (4)
  - Crawler (3)

Problem: How to find such hierarchies automatically?
Main Idea

- Example for tagged folksonomy corpus
  - Publications manually tagged with keywords (ACM classification)
  - Social tagging activities (flickr, del.icio.us, ...)
- Input data: \((<\text{nodeid}>,<\text{keyword}>) \& (<\text{nodeid}>,<\text{year}>)\)
- Query:
  - Topic hierarchy of \(<\text{keyword k}>\) for \(<\text{startyear}> - <\text{endyear}>\)?
- Semantic GrowBag Output: Graph with
  - Keywords as nodes (tagged with their rank)
  - Edges denoting the subsumption hierarchy
    - Different weights for the confidence
    - Graphically displayed in different classes
      - Currently two: ‘dashed’ vs. ‘double-headed’
Example Topic Hierarchy from GrowBag

- Keyword: ‘search engine’
- Time span: 2003-2004
The Algorithm: Overview

- 3 basic steps
  1. Determine most closely related keywords for a given keyword $k$ (top-$X$ keywords of $k$) based on co-occurrences + compute a ranking
     - Using tag-co-occurrence matrix, ‘TF-IDF’ and Biased PageRank
  2. Find super-topic / sub-topic relations between $k$ and its top-$X$ keywords
  3. Combine the relations and the top-X tags into a single graph, the topic facet of $k$
Part I: **Determine ranking of keywords**

- Given query: keyword k, time period p

1. Compute a list of top-X related keywords
   - Based on weighted co-occurrences with other keywords during p
   - Take the set of keywords that accounts for 20% of TF-IDF mass
     → Called ‘main keywords of k’
2. Compute a ranking using Biased PageRank, biasing on these main keywords of k
3. Output: Ranked list of keywords related to k (L_k)
4. → Do this for all keywords
Example: Step 1 for keyword ‘IR’

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tag</th>
<th>TFxIDF</th>
<th>TF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IR</td>
<td>669.0</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>Search Engine</td>
<td>67.0</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>Language Model</td>
<td>60.2</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>WWW</td>
<td>51.1</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>Web search</td>
<td>41.4</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Query expansion</td>
<td>39.9</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Text mining</td>
<td>36.2</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>Indexing</td>
<td>29.1</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>NLP</td>
<td>28.3</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>Question Answer</td>
<td>27.0</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>22</td>
<td>Inf. Extraction</td>
<td>18.7</td>
<td>4</td>
</tr>
</tbody>
</table>

Co-occurrence analysis

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tag</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IR</td>
<td>148.3</td>
</tr>
<tr>
<td>2</td>
<td>WWW</td>
<td>103.8</td>
</tr>
<tr>
<td>3</td>
<td>Machine Learning</td>
<td>94.3</td>
</tr>
<tr>
<td>4</td>
<td>Ontology</td>
<td>91.1</td>
</tr>
<tr>
<td>5</td>
<td>Search Engine</td>
<td>90.4</td>
</tr>
<tr>
<td>6</td>
<td>Semantic Web</td>
<td>88.9</td>
</tr>
<tr>
<td>7</td>
<td>Digital Library</td>
<td>85.6</td>
</tr>
<tr>
<td>8</td>
<td>Web search</td>
<td>79.2</td>
</tr>
<tr>
<td>9</td>
<td>Knowledge Management</td>
<td>77.3</td>
</tr>
<tr>
<td>10</td>
<td>Query expansion</td>
<td>76.2</td>
</tr>
</tbody>
</table>

Biased PageRank: L_k(‘IR’)
### Step 1 for keyword ‘Query expansion’

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tag</th>
<th>TFxIDF</th>
<th>TF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Query expansion</td>
<td>130.0</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>IR</td>
<td>23.4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Probabilistic Model</td>
<td>16.9</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Web search</td>
<td>15.5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Search Engine</td>
<td>13.4</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>Page Segmentation</td>
<td>12.6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Web information retrieval</td>
<td>12.2</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

#### Co-occurrence analysis

#### Biased PageRank (L_k(‘Query expansion’))

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tag</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IR</td>
<td>541.3</td>
</tr>
<tr>
<td>2</td>
<td>Query expansion</td>
<td>490.5</td>
</tr>
<tr>
<td>3</td>
<td>Probabilistic Models</td>
<td>476.4</td>
</tr>
<tr>
<td>4</td>
<td>Search Engines</td>
<td>74.6</td>
</tr>
<tr>
<td>5</td>
<td>Web Search</td>
<td>45.3</td>
</tr>
<tr>
<td>6</td>
<td>WWW</td>
<td>41.4</td>
</tr>
<tr>
<td>7</td>
<td>Data mining</td>
<td>37.5</td>
</tr>
<tr>
<td>8</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Part II: Find super-topic / sub-topic relations

- Given: keyword k, ranked lists L_k’ for all keywords k’ (for period p)

1. Extract the top-X keywords from L_k (the ranked_list of k)
   - Top-X the same as in Part I
2. For all keywords k* in top-X keywords
   1. Get the scores of k and k* in L_k and L_k*
   2. If both scores of k are larger than the ones of k*: k is a super-topic of k*
   3. Analogously for “smaller than” → “sub-topic”
3. Do it for all keywords k
→ Output: all super-topic / sub-topic relations
→ Example: ‘IR’ super-topic of ‘query expansion’, ‘query expansion’ sub-topic of ‘IR’
Part III: Create the topic hierarchy

1. Use the main keywords as found in step 1 as ‘seed’
2. Get all keywords $k^*$ which are sub-topics of the main keywords (or of which the main keywords a super-topics) (output of step 2)
   - Repeat recursively until no additional sub-topics found
   - ‘Growing’ the set of nodes $N$ of the topic hierarchy
3. Add the immediate super-topics of the main keywords
4. Find all super-topic/sub-topic relations where nodes in $N$ are involved in
   → The set of links $L$ in the topic hierarchy
5. Determine the ‘strength’ of a relation $k_1$ to $k_2$:
   - $k_1$ super-topic of $k_2$ and $k_2$ sub-topic of $k_1$: high
   - $K2$ sub-topic of $k_1$: low
Experimental Setup

- DBLP data (as of July/August 2005)
- Extracted manually added keywords + abstracts using the links provided in DBLP (limited to Springer, ACM DL, IEEE DL)
  - About 53,000 documents stored in mysql database
  - About 180,000 distinct tags
- Postprocessing:
  - Replacing well-known acronyms (e.g. XML, ) by the most popular full version in the database
  - Stemming of keywords
- Creating a list of most popular keywords (8622, about 60%)
  - Actually ‘key phrases...’, not only keywords...
Development of Hierarchies over Time (1)

1998-1999

World Wide Web (WWW) (4)

TCP/IP (30)

virtual communities (187)

search engine (3)

2001-2002

self-adaptation (4)

user log (5)

World Wide Web (WWW) (2)

web search (21)

user profile (160)

web design (212)

crawling (53)

Libraries (81)

information retrieval (3)

sketching (516)
Development over Time (2)

2003-2004
The Semantic GrowBag Demonstrator
for Tagged Computer Science Publications

Available Topic facets for 2003-2004 with at least one strong edge:

Graphs with no strong edge  Graphs without edges

(In the reduced version those subgraphs, that are not connected to the graph with the start tag, are folded into the participating top-X node. Please note, that quite some graphs do not contain edges because of the power-law nature of the co-occurrence distribution of tags in our collection. The `top-X` values are an indicator for the size of the `community` around a tag (limited to 10 for visualization reasons). The `nodes` and the `edge` values are shown to have an indication of the size of the graph.

<table>
<thead>
<tr>
<th>Topic</th>
<th>top-X</th>
<th>Nodes</th>
<th>Edges</th>
<th>Link to Pictures</th>
<th>Development over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstract interpretation</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>Full version</td>
<td>Full version</td>
</tr>
<tr>
<td>abstraction</td>
<td>7</td>
<td>19</td>
<td>14</td>
<td>Full version</td>
<td>Full version</td>
</tr>
<tr>
<td>access control</td>
<td>8</td>
<td>19</td>
<td>14</td>
<td>Full version</td>
<td>Full version</td>
</tr>
<tr>
<td>active objects</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>Full version</td>
<td>Full version</td>
</tr>
<tr>
<td>adaptation</td>
<td>10</td>
<td>20</td>
<td>10</td>
<td>Full version</td>
<td>Full version</td>
</tr>
<tr>
<td>ad hoc</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>Full version</td>
<td>Full version</td>
</tr>
<tr>
<td>ad hoc networks</td>
<td>9</td>
<td>21</td>
<td>20</td>
<td>Full version</td>
<td>Full version</td>
</tr>
</tbody>
</table>
Summary

- Problem:
  Sensible hierarchical organization of dynamic facets

- Idea:
  Use GrowBag algorithm to automatically create topic hierarchies + organize topic facets accordingly

- Applicable to any tagged set of objects
  - medline, bibsonomy,...
Future Work

- Evaluation of hierarchies difficult...
- Use different input data sets, such as the Medline Database
- Find ‘interesting’ topics (changing over time) automatically (change rate of the top-X over time)
- Detect keywords which are used as synonyms in two non-connected communities
- Allow for more than one start tag
Questions?

balke@l3s.de