Category Hierarchy Maintenance: a Data-Driven Approach

Quan Yuan¹, Gao Cong¹, Aixin Sun¹, Chin-Yew Lin², Nadia Magnenat-Thalmann¹

¹Nanyang Technological University
²Microsoft Research Asia

Presenter: Quan YUAN
1.1 Category Hierarchy

- Category hierarchies have been widely adopted to organize large volume of documents.
  - Community-based QA: Yahoo! Answers (YA), AnswerBag (AB)
  - Web directory: Open Directory Project (ODP)
  - Online retailers: Amazon, eBay
1.2 Category Hierarchy Maintenance

- Hierarchy evolves slower than its documents:
  - Structural Irrelevance
  - Semantic Irrelevance
1.2 Category Hierarchy Maintenance

- Hierarchy evolves slower than its documents:
  - Structural Irrelevance
  - Semantic Irrelevance

Hierarchy Diagram:

- Original Hierarchy:
  - Root
    - Geography
    - Politics
      - Petroleum

- Modified Hierarchy (after oil spill event):
  - Root
    - Geography
    - Petroleum (renamed to Petroleum Politics)

BP Oil Spill Image
1.2 Category Hierarchy Maintenance

- Hierarchy evolves slower than its documents:
  - Structural Irrelevance
  - Semantic Irrelevance

- Consequences:
  - Poor experiences in accessing information.
  - Poor classification accuracy.

Category Hierarchy Maintenance

Modify a hierarchy to make it better reflect the topics of documents.

- eBay modifies its hierarchy more frequently.
2.1 Hierarchy Generation

Creating a hierarchy structure from a set of categories.

- C. C. Aggarwal et al. KDD 1999.
- X. Qi et al. CIKM 2011.

Input: a set of categories

Output: a hierarchical structure

Do not change the given categories

→ cannot solve structural irrelevance

(e.g., multi-functional devices)
2.2 Hierarchy Modification

- Modify a given hierarchy for better classification accuracy.
  - L. Tang et al. KDD 2006.

- Only modify structure.
- Do not change leaf categories.
  - cannot solve semantic irrelevance.
3.1 Overview

- **Structure**
  category

- **Semantics**
  hidden topics

- **Content**
  documents
  collection

- Hidden topics of a category should be similar.
- Hidden topics of different categories should be separate.
3.1 Overview

- **Structure category**
- **Semantics hidden topics**
- **Content documents collection**

- Discover hidden topics from documents **Sprout**
- Merge similar hidden topics to form new categories **Merge**
- Rebuild the hierarchical structure **Assign**
3.1 Overview

- A data-driven approach to modify a given (original) hierarchy with reference to an auxiliary hierarchy.
- Auxiliary hierarchy contains similar topics with the original hierarchy.
- We can get the auxiliary hierarchy from the Internet (e.g., ODP→YA)

---

**Original Hierarchy**
- Hardware
  - Desktop
  - Laptop

**Auxiliary Hierarchy**
- Hardware
  - Mainframe
  - Video Camera
  - Mouse
3.2 Sprout

- Generate hidden topics.
- Sprout the original category.
3.2 Sprout

- Generate hidden topics.
- Sprout the original category.

- Q: How to use desktop camera to record video?
- Q: What kind of video camera is perfect for Skype on desktop?
3.2 Sprout

- Generate hidden topics.
- Sprout the original category.
3.2 Sprout

Criteria for good hidden topics:

- Suitable granularity.
- No semantic overlap.
- Each topic: 
  cohesive; different topics: separated.
3.2 Sprout

- Criteria for good hidden topics:
  - Suitable granularity.
  - Few semantic overlap.
  - Each topic: cohesive; different topics: separated.
3.2 Sprout

Criteria for good hidden topics:

- Suitable granularity.
- Few semantic overlap.
- Each topic: cohesive; different topics: separated.
3.2 Sprout

- Generate hidden topics.
- Sprout the original category.
3.2 Sprout

- Generate hidden topics.
- Sprout the original category.
3.3 Merge

- Group similar hidden topics (document collections) to form a new set of categories.

Original Hierarchy

- Hardware
  - Desktop
  - Laptop

Auxiliary Hierarchy

- Hardware
  - Mainframe
  - Video Camera
  - Mouse

Modified Hierarchy

- Hardware
  - Desktop
  - Laptop
  - Peripherals
### 3.3 Merge

#### Similarity between two hidden topics:
- Content similarity (document containment)
- Path similarity (inverted path length)

\[
s_{i,j} = \frac{1}{1 + \text{path}(C_i, C_j)}
\]

\[
\frac{1}{1+2} = 0.33
\]

**Original Hierarchy**
- Hardware
  - Desktop
  - Laptop

**Auxiliary Hierarchy**
- Hardware
  - Mainframe
  - Video Camera
  - Mouse

**Similar topics**
- \(t_1\): Desktop mainframe
- \(t_2\): Laptop video camera

\[
sim(t_1, t_2) = 0.6 \times 0.33 \times 0.7 +
\]
3.3 Merge

- **Similarity between two hidden topics:**
  - Content similarity (document containment)
  - Path similarity (inverted path length) \( s_{i,j} = \frac{1}{1 + \text{path}(C_i, C_j)} \)

- **Example:**
  - **Original Hierarchy:**
    - Desktop
    - Laptop
    - Mainframe
  - **Auxiliary Hierarchy:**
    - Hardware
    - Video Camera
    - Mouse

- **Equation:**
  \[
  \text{sim}(t_1, t_2) = 0.6 \times 0.33 \times 0.7 + 0.9 \times 0.33 \times 0.8 = 0.38
  \]
3.4 Assign

- Rebuild the parent-children relations.

Original Hierarchy
- Hardware
  - Desktop
  - Laptop
  - Dell

Modified Hierarchy
- Hardware
  - Desktop
  - Laptop
  - Peripherals
  - Dell

Documents of Dell

Modified Level

Unmodified Level
3.4 Assign

Rebuild the parent-children relations.

Original Hierarchy

Modified Hierarchy

Documents of Dell
4.1 Experimental Setup

- **Evaluation aspects:**
  - Classification accuracy
  - Semantic quality

- **Datasets:** Yahoo! Answers ($H_{YA}$), AnswerBag ($H_{AB}$) and Open Directory Project ($H_{ODP}$).
  - Two general categories: Sports and Computers.
  - Pre-processing: truncate level, remove stopwords, stemming.

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>$H_{YA}$</th>
<th>$H_{AB}$</th>
<th>$H_{ODP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#document</td>
<td>421,163</td>
<td>148,822</td>
<td>203,448</td>
</tr>
<tr>
<td>#nodes</td>
<td>75</td>
<td>195</td>
<td>460</td>
</tr>
<tr>
<td>#non-leaf nodes</td>
<td>40</td>
<td>70</td>
<td>98</td>
</tr>
<tr>
<td>height</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
4.2.1 Evaluation by Classification

- Modify the YA hierarchy to get a new hierarchy $H_{nYA}$.
- Classifiers: Naive Bays (NB) and Support Vector Machine (SVM).
- Metrics: Macro F1, Micro F1 (the higher the better).
- Baseline Methods:
  - Bottom Up Clustering (BUC): Cluster categories using K-means in a bottom-up manner.
- Manually annotate 500 questions as testing instances.
4.2.2 Classification Results

- SMA achieves substantial improvement in classification accuracy.

(a) Micro-$F_1$

(b) Macro-$F_1$
4.2.2 Classification Results

- SMA achieves substantial improvement in classification accuracy.
- Using original hierarchy as auxiliary hierarchy is effective for modification.

![Graphs showing Micro-F1 and Macro-F1 for different methods and hierarchies](image)
4.3 Evaluation by User Study

- **Category Labels** (should be easy to understand)

<table>
<thead>
<tr>
<th>Judgment</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H_nY_A) is better than (Y_A)</td>
<td>12</td>
</tr>
<tr>
<td>(H_nY_A) is not as good as (Y_A)</td>
<td>1</td>
</tr>
<tr>
<td>Both are equally good</td>
<td>81</td>
</tr>
<tr>
<td>Neither is good</td>
<td>6</td>
</tr>
</tbody>
</table>

- The generated labels well reflect the content of categories

- **Category Structure** (should reflect the topics of data)

- The modified hierarchy is of high quality.

Q: What corrupts files in my USB Flash Disk?

**PATH(\(Y_A\)):**
- Computer → Hardware → Add ons

**PATH(\(H_nY_A\)):**
- Computer → Hardware → Storage
4.4 Case Study

- Select three categories **Software**, **Internet** and **Hardware** under **Computer & Internet** from Yahoo! Answers.

**Part of the Yahoo! Answers hierarchy**

**Part of the modified hierarchy**
4.4 Case Study

Select three categories **Software, Internet** and **Hardware** under **Computer & Internet** from Yahoo! Answers.

**Part of the Yahoo! Answers hierarchy**

**Part of the modified hierarchy**
4.4 Case Study

- Select three categories *Software*, *Internet* and *Hardware* under *Computer & Internet* from Yahoo! Answers.
4.4 Case Study

- Select three categories **Software**, **Internet** and **Hardware** under **Computer & Internet** from Yahoo! Answers.
5 Conclusion

- Category Hierarchy has been widely used to organize large volume of documents.
- With continuously added documents, hierarchies suffer from the problems of structural irrelevance and semantic irrelevance.
- We propose a method SMA to modify the hierarchy, which comprises three operations.
- Experimental results demonstrate that SMA is able to generate semantically meaningful hierarchies with better classification accuracy.
Q & A?

Thank You!