Combining Local and Global Information for Enhanced Deep Classification

Heung-Seon Oh, Yoonjung Choi, Sung-Hyon Myaeng
KAIST, South Korea
Outline

- Introduction
- Past Research Direction
- Proposed Method
- Experiment Results
- Conclusion
Introduction

- **Traditional Text Classification**
  - Classifying a document to a small number of flat category set or a shallow hierarchy of categories

- **Large-scale Hierarchical Text Classification**
  - Text classification targeting on the large-scale web taxonomy such as ODP and Yahoo! Directory
  - Usefulness in rare query classification, contextual advertising, and web search improvement
Web Taxonomy

- ODP and Yahoo! Directory
- Hundreds of thousands of categories
- Millions of documents
  - But have short texts
  - Distributed in the skewed manner
- Deep depths
  - More than 10 levels from the root of a hierarchy
Past Research Directions

- Three different approaches: big-bang, top-down, and narrow-down

- Big-bang approach
  - Builds a single classifier over the entire hierarchy
  - Takes long training time and not effective

- Top-down approach
  - Builds a classifier at each level along with the path from root to leaf category
  - Error propagation from top level to low levels
Past Research Directions

- **Narrow-down approach**
  - Observed that most of categories are not relevant to an input document

- **Deep classification (SIGIR’08)**
  - Utilizes not only machine learning but also search technique
  - **Search stage**
    - Finds highly related categories with respect to an input document
  - **Classification stage**
    - Selects training data with respect to the category candidates
    - Assigns a category to an input document after building a classifier based on the categories and training data
Limitations of Deep Classification

- Data sparseness problem
  - Ancestor-assistant strategy for selecting training data
  - Still suffers from sparseness since common ancestor nodes appear close to category candidates

Pruned Hierarchy

Ancestor-assistant
Limitations of Deep Classification

- Utilizing only local information to classification
- Top level guidance will help to better classification
Proposed Algorithm

- **Enhanced deep classification**
  - Neighbor-assistant strategy for selecting training data
    - Includes the data of descendants of ancestors

- **Naïve Bayes combination**
  - Utilizes global information in terms of top-level categories to classification with combination of local information
Overview of Proposed Method

Stage 1

Training Data → Inverted Index Construction → Inverted Index → Computing Relevance → Category Candidates → Given Document

Stage 2

Feature Selection → Global Model Generation

Training Data Selection → Local Model Generation → Classification → Category
Search Stage

- Finds highly related categories to an input document

Two search strategies

- Document-based
  - Retrieves top-k documents and takes the categories with respect to the documents as category candidates

- Category-based
  - Represents categories as a set of documents under each category
  - Retrieves top-k categories as category candidates

Implementation based on Lucene open source IR engine
Classification: Training Data Selection

- **Neighbor-assistant strategy**
  - Collects training data from the descendants of ancestors and a category candidate
Classification: Naïve Bayes Combination Classifier

- **Naïve Bayes Classifier**

\[
c^* = \arg \max_{c_i \in C} \{ P(c_i \mid d) \}
\]

\[
= \arg \max_{c_i \in C} \{ P(d \mid c_i) P(c_i) \}
\]

\[
= \arg \max_{c_i \in C} \{ P(c_i) \prod_{j=1}^{N} P(t_j \mid c_i)^{v_j} \} \quad (2)
\]

- **Prior and likelihood by combining local and global models**

\[
P(t_j \mid c_i) = \lambda P(t_j \mid c_i^{local}) + (1 - \lambda) P(t_j \mid c_i^{global}) \quad (3)
\]

\[
P(c_i) = \lambda P(c_i^{local}) + (1 - \lambda) P(c_i^{global}) \quad (4)
\]
Classification: Global Model

- Focuses on parameters highly relevant to top level categories
- Use 15,000 terms for each top level category after chi-square feature selection
- Apply add-one smoothing
- Prior and likelihoods are calculated:

\[
P(c_i^{global}) = \frac{|D_i|}{|D|} \quad (5)
\]

\[
P(t_j | c_i^{global}) = \frac{\sum_{d_k \in c_i^{global}} tf_{jk} + 1}{\sum_{t_u \in V^{global}} \sum_{d_k \in c_i^{global}} tf_{uk} + |V^{global}|} \quad (6)
\]
Classification: Local Model

- Focuses on parameters of category candidates
- No feature selection & smoothing
  - Should reflect terms relevant to category candidates
  - Time burden
- Prior and likelihoods are calculated:

\[
P(c_{i_{local}}) = \frac{|D_i|}{|D|} \quad (7)
\]

\[
P(t_j \mid c_{i_{local}}) = \frac{\sum_{d_k \in c_{i_{local}}} tf_{jk}}{\sum_{t_u \in V_{local}} \sum_{d_k \in c_{i_{local}}} tf_{uk}} \quad (8)
\]
Experiments: Dataset

- ODP taxonomy after filtering
  - 1,427,426 docs for training
  - 13,000 docs for testing
Experiments: Evaluation Metric

- Evaluation in general text classification
  - Evaluations for each category
  - Costly due to many categories

- Evaluation for each level with Micro-F1
  - E.g.,
    - For A/B/C/D, evaluate A, A/B, A/B/C, A/B/C/D whether it is correctly classified or not
### Experiments:
#### Algorithm Comparison

<table>
<thead>
<tr>
<th></th>
<th>Deep Classification</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top-k</strong></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td><strong>Search</strong></td>
<td></td>
<td>Category-based</td>
</tr>
<tr>
<td><strong>Training Data</strong></td>
<td>Ancestor-assistant</td>
<td>Neighbor-assistant</td>
</tr>
<tr>
<td><strong>Selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Classifier</strong></td>
<td>Trigram Language Model</td>
<td>Naive Bayes Combination of Local and Global Information</td>
</tr>
</tbody>
</table>
Experiments: Results

- Overall Comparison

Figure 8. Comparison between the Deep Classification and the Proposed Method
Experiments:
Results

- Effectiveness of combining local and global information

Figure 9. Roles of Local vs. Global Information
Experiments: Results

- Comparison between ancestor and neighbor-assistant strategy
Conclusion

- This paper proposed an method for large-scale hierarchical text classification
  - Neighbor-assistant strategy for selecting training data
  - Naïve Bayes combination classifier utilizing both local and global information

- A series of experiments show the effectiveness of our proposed method