Deep Classification in Large-scale Web Hierarchies

Gui-Rong Xue\textsuperscript{1}
Joint work with Dikan Xing\textsuperscript{1}, Qiang Yang\textsuperscript{2}, Yong Yu\textsuperscript{1}

\textsuperscript{1}, Shanghai Jiao-Tong University
\textsuperscript{2}, Hong Kong University of Science and Technology
Outline

• Motivation
• Related Work
• Deep Classification
• Experiments
• Conclusion
Outline

• Motivation
• Related Work
• Deep Classification Algorithm
• Experiments
• Conclusion
Motivation

• Text classification

• Web application:
  ✓ Web content organization
  ✓ Personalized search
  ✓ Web advertising
Motivation

• Current classification algorithms confine themselves to classifying a document into two or a few predefined categories. It limits the power of text classification.
Motivation

• Benefits of classification on large taxonomies by assigning very detailed topic-related class information
  – Browsing Web search results
  – Semantic Page-Ads matching
  – User modeling for personalized search
Motivation

• Three difficulties for directing applying traditional algorithms
  – The large size of the taxonomy of categories
  – A very long training time by traditional methods
  – How to efficiently consider that those categories are with a hierarchical structure
Our Solutions

• Deep Classifier on Large-Scale Hierarchies (SIGIR 2008)

• Deep Classifier: Automatically Categorizing Search Results into Large-scale Hierarchies (WSDM 2008)

• Advertising Keyword Suggestion Based on Concept Hierarchy (WSDM 2008)

• Exploring Naïve Bayesian for Large Scale Categories
Motivation

• We propose a novel effective algorithm for large-scale hierarchical classification: *deep classification*
  – Utilizing the advantages of search engine and classification algorithm
Outline

• Motivation
• Related Work
• Deep Classification
• Experiments
• Conclusion
Related Work

• Flat classification
  – Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Naive Bayes (NB)
    • Higher performance on small categories
    • Lower performance on large-scale categories

• Hierarchical text classification
  – Big-bang approach
  – Top-down approach
Related Work

• Big-bang approaches: a single classifier is used by considering the hierarchical structure of the categories
  – A modified version of SVM [Cai and Hofmann 2004]
  – EM based shrinkage classification [McCallum and Rosenfeld 1998]
  – Rocchio-like classifiers [Labrou and Finin 1999]

• Most of experiments were conducted with at most a few hundreds categories
  infeasible to directly build a classifier for a large-scale hierarchy
Related Work

• Top-down approaches
  – Hierarchical SVM
  – TAPER uses the NB to build hierarchical classification

• Previous work [Liu et al. 2005] evaluates the performance of hierarchical SVM
  – The performance is about 40% lower on measures of Micro-F1 at the 5th level on Yahoo! Directory.

• One main reason for lower performance
  – Misclassification at a parent category may force a document to be excluded from the child categories
Outline

• Motivation
• Related Work
• Deep Classification
• Experiments
• Conclusion
Deep Classification

• Observations:
  – For a very large scale hierarchy, the number of related categories for a document is much less than the number of the unrelated categories
  – The classification performance on a few categories will be better than that on a larger set of categories
Deep Classification

Our approach:
Stage 1: Search Stage
Stage 2: Classification Stage
Deep Classification

• Stage 1: Search Stage
  – A search based approach is used to find the category candidates for the given document

  – Two Strategies for building training set
    • Document based Strategy
    • Category based Strategy
Deep Classification

• Document based Strategy
  – Indexing documents with a search system
• Inverted Indexing
  – Performing search
• Query: the given document
• Output: ranked categories
  – Merging the categories information of the relevant documents
  – Sorting
Deep Classification

• **Category based strategy**
  – Merging the documents in a category into a category-document
  – Indexing category-documents with a search system
    • Inverted Indexing
  – Performing search
    • Query: Given document
    • Output: Top-K most similar categories
Deep Classification

- The large hierarchies are pruned into a narrow one
Deep Classification

• Stage 2: Classification Stage

![Diagram showing stages of deep classification](image-url)
Deep Classification

• Training Strategies
  – Flat strategy
  – Pruned top-down strategy
  – Ancestor-assistant strategy
Deep Classification

• Flat strategy
Deep Classification

- Pruned top-down strategy
Deep Classification

• Ancestor-assistant strategy

1, combining the training data from the category candidate itself and the training data from its ancestors

2, Don’t share the common ancestors of other category candidates.
Deep Classification

• Strategies for Classifier Selection
  – Classifiers
    • Support vector Machine (SVM)
    • Naive Bayesian (NB) – efficient but a litter low performance than SVM
  – Feature Representation
    • Unigram
      \[ P(t_i \mid t_1 \cdots t_{i-1}) = P(t_i) \]
    • 3-Gram Language Models for Classifiers:
      \[ P(t_i \mid t_1 \cdots t_{i-1}) = P(t_i \mid t_{i-n+1} \cdots t_{i-1}) \]
Outline

• Motivation
• Related Work
• Deep Classification
• Experiments
• Conclusion
Experiments

• Dataset
  – Open Directory Projects
    • 4,800,870 Web pages
    • 712,548 categories
    • 17 top-categories
      – Arts, Business and Economy, Computers and Internet, Games, Health, Home, Kids and Teens, News, Recreation, Reference, Science, Shopping, Society, Sports, Regional, Adult and World

• After downloading, we get 130,000 categories and 1.3M Web pages
• 130,000 documents from 1.3 million documents as the testing data
• Documents Distribution on Different Level

![Bar chart showing the distribution of documents across different levels. The highest level accounts for 68.6% of the total documents.]

- Level 1: 20,5199 documents
- Level 2: 101,689 documents
- Level 3: 280,896 documents
- Level 4: 322,661 documents
- Level 5: 293,249 documents
- Level 6: 156,795 documents
- Level 7: 99,098 documents
- Level 8: 37,231 documents
- Level 9: 8,552 documents
- Level 10: 1,139 documents
- Level 11: 57 documents

The dataset shows a significant majority of documents are concentrated in the earlier levels, with Level 5 contributing the most (68.6% of the total documents).
Dataset

- Categories Distribution on Different Level

<table>
<thead>
<tr>
<th>Level</th>
<th># of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>467</td>
</tr>
<tr>
<td>3</td>
<td>5408</td>
</tr>
<tr>
<td>4</td>
<td>22359</td>
</tr>
<tr>
<td>5</td>
<td>43896</td>
</tr>
<tr>
<td>6</td>
<td>35026</td>
</tr>
<tr>
<td>7</td>
<td>25553</td>
</tr>
<tr>
<td>8</td>
<td>17225</td>
</tr>
<tr>
<td>9</td>
<td>5967</td>
</tr>
<tr>
<td>10</td>
<td>1641</td>
</tr>
<tr>
<td>11</td>
<td>247</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
</tr>
</tbody>
</table>

64.1%
Dataset

- About 93.46% of the documents belong to one category. Only 6.54% of the documents have two or more categories.

<table>
<thead>
<tr>
<th>Number of Categories</th>
<th>Number of Documents</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1214977</td>
<td>93.460%</td>
</tr>
<tr>
<td>2</td>
<td>74237</td>
<td>5.711%</td>
</tr>
<tr>
<td>3</td>
<td>2410</td>
<td>0.185%</td>
</tr>
<tr>
<td>&gt;=4</td>
<td>195</td>
<td>0.015%</td>
</tr>
</tbody>
</table>
Comparison Algorithms

• Hierarchical SVM
  – Top-Down based SVM

• Search based Strategy
  – Just taking the most similar category as the category for the given document,

• Deep Classification
General Performance

- Overall performance

![Bar chart showing Mi-F1 performance across different levels with three strategies: Search based Strategy, Hierarchical SVM, and Deep Classification.](chart.png)

- Search based Strategy: 51.8%
- Hierarchical SVM: 51.8% (Level 5)
- Deep Classification: 29.2% (Level 5)

Level 5 has the highest Mi-F1 score for both hierarchical SVM and deep classification methods.
Strategies Selection

- Search Strategies

Category-based strategy is used in deep classification.
Ancestor-assistant strategy is used in deep classification.
Strategies Selection

• Classifier Selection

3-Gram NB is used as the classifier in deep classification
Outline

• Motivation
• Related Work
• Deep Classification
• Experiments
• Conclusion
Conclusion

• A novel deep classification algorithm for Web classification on a large scale text hierarchy is proposed.

• A two-stage algorithm is presented, consisting of a search stage and a classification stage.

• Experimental results show that the proposed deep classification algorithm can achieve highest performance.

• Future work
  – the deep classification algorithm for different kinds of applications, such as online advertisement classification
  – Improving the efficiency of the search stage algorithm of deep classification
Q/A

Thank you!
References