Hierarchical Classification of Web Content

Susan Dumais
Microsoft Research

Hao Chen
University of California at Berkeley

SIGIR 2000

Presented by Xudong He, Ailun Yi

University of Edinburgh
March, 2009
Outline

- Background
- Text Categorisation
- Hierarchical Classification
- Data Set and Feature Selection
- SVM
- Result and Evaluation
- Conclusion
Background

- Search Results Representation
Background

- Search Results Representation
- Flat ranked list. e.g. Google
Background

- Search Results Representation
- Flat ranked list. e.g. Google
- Clustering, Clusty
  - Automatic but hard to assign labels
  - Runtime cost, hard for large set of documents
Background

- Search Results Representation
- Flat ranked list. e.g. Google
- Clustering, Clusty
  - Automatic but hard to assign labels
  - Runtime cost, hard for large set of documents
- Hierarchical, Directory-style. ODP, Yahoo! Directory
Background

➤ Search Results Representation

➤ Flat ranked list. e.g. Google

➤ Clustering, Clusty

  ➤ Automatic but hard to assign labels
  ➤ Runtime cost, hard for large set of documents

➤ Hierarchical, Directory-style. ODP, Yahoo! Directory

➤ Motivation: Combine both?

  ➤ Need automatic text classifiers
  ➤ Efficiency Consideration – 16+ top level categories, thousands lower-level ones
Text Categorisation
Classify documents into broad semantic topics (e.g. Politics, Entertainment, Sports, etc.)

Democratic vice presidential candidate John Edwards on Sunday accused President Bush and Vice President Dick Cheney of misleading Americans by implying a link between deposed Iraqi President Saddam Hussein and the Sept. 11, 2001 terrorist attacks.

While No. 1 Southern California and No. 2 Oklahoma had no problems holding on to the top two spots with lopsided wins, four teams fell out of the rankings — Kansas State and Missouri from the Big 12 and Clemson from the Atlantic Coast Conference and Oregon from the Pac-10.
Classify documents into broad semantic topics (e.g. Politics, Entertainment, Sports, etc.)

Democratic vice presidential candidate John Edwards on Sunday accused President Bush and Vice President Dick Cheney of misleading Americans by implying a link between deposed Iraqi President Saddam Hussein and the Sept. 11, 2001 terrorist attacks.

While No. 1 Southern California and No. 2 Oklahoma had no problems holding on to the top two spots with lopsided wins, four teams fell out of the rankings — Kansas State and Missouri from the Big 12 and Clemson from the Atlantic Coast Conference and Oregon from the Pac-10.

Supervised Learning Problem

- Common algorithms: SVM, NB, kNN, NNet
Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
- News Stories / Academic Paper / Blogs / Tweets, by Subject
Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
- News Stories / Academic Paper / Blogs / Tweets, by Subject
- Selective Dissemination of Information to information consumers
Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
- News Stories / Academic Paper / Blogs / Tweets, by Subject
- Selective Dissemination of Information to information consumers
- Natural Language Processing (Word Sense Disambiguation, POS Tagging etc.)
Hierarchical Classification
Hierarchical Classification

- Multi-class problem
  - 1-against-rest, sort $p(c | d)$ into categories
Hierarchical Classification

- Multi-class problem
  - 1-against-rest, sort $p(c | d)$ into categories
- Multi-level problem
  - How?
Hierarchical Classification

- Multi-class problem
  - 1-against-rest, sort $p(c | d)$ into categories

- Multi-level problem
  - How?

- Our Goal
  - Classify into Hierarchical Categories (for Search Results)
The Approach: Flat vs. Hierarchical
The Approach: Flat vs. Hierarchical

A Subject Taxonomy

Comparison among
(1) $P(L2)$
(2) $P(L1) \&\& P(L2)$
(3) $P(L1) \ast P(L2)$

Confidence the document belongs to the class
Experiment Setting: Data Set

- **LookSmart directory (May, 1999)**
  - 370,597 web pages
  - 17,173 categories over 7-level hierarchy

- **Focus on top 2 levels**
  - 13 top-level & 150 second-level categories

- **Train / Test split**
  - Randomly select 16%, 50,078 pages for training
  - 10,024 pages for testing
Experiment Setting: Feature Preprocessing

- **Feature Vector**
  - Words (uni-gram vs. bigram)
  - Binary (vs. Tf-idf)

- **Feature Selection**
  - Top 1000 by Mutual Information:

\[
MI(F, C) = \sum_{F \in \{f, \bar{f}\}} \sum_{C \in \{c, \bar{c}\}} P(F, C) \log \frac{P(F, C)}{P(F)P(C)}
\]

- **Second-level Local Selection**
  - Second level feature sets narrowed by top-level categories, more balanced features
The SVM Model

- The equation of a hyperplane:
  \[ \mathbf{w}^T \mathbf{x} + b = 0 \]

- The separator property:
  \[ \mathbf{w}^T \mathbf{x}_i + b \geq 1 \text{ iff } y_i = +1 \]
  \[ \mathbf{w}^T \mathbf{x}_i + b \leq 1 \text{ iff } y_i = -1 \]

- Minimise \( ||\mathbf{w}||^2 \) subject to:
  \[ \forall_{i=1}^n : y_i \left[ \mathbf{w}^T \mathbf{x}_i + b \right] \geq 1 - \xi_i \]

- The decision function:
  \[ f(x) = \text{sign} \left( \sum_{i=1}^N \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b \right) \]

- Independent of the dimensionality of the feature space
  The normal of the decision surface is a linear combination of examples
Why SVMs are good for TC
Why SVMs are good for TC

- High-dimensional input space.
  Use overfitting protection not dependent on the number of features

- Few irrelevant features:
  almost all feature contain considerable information.
Why SVMs are good for TC

- High-dimensional input space.
  
  Use overfitting protection not dependent on the number of features

- Few irrelevant features:
  
  almost all feature contain considerable information.

- Sparse document vectors
  
  Each of the document vectors contain only a few non-zero element

- Most text categorisation problems are linearly separable
Train and Optimise the SVM
Train and Optimise the SVM

- Use the simplest linear SVM
  - Good classification accuracy
  - Fast to learn and apply

- Use Platt’s sequential minimal optimization (SMO) algorithm
  - Break large quadratic programming (QP) problem into small QP problems
Train and Optimise the SVM

- Use the simplest linear SVM
  - good classification accuracy
  - fast to learn and apply

- Use Platt’s sequential minimal optimization (SMO) algorithm
  - Break large quadratic programming (QP) problem into small QP problems

- New items are classified by using $w \cdot x$.

- Fit a sigmoid to output by maximum likelihood and produce posterior probabilities.
Train and Optimise the SVM

- Use the simplest linear SVM
  - good classification accuracy
  - fast to learn and apply

- Use Platt’s sequential minimal optimization (SMO) algorithm
  - Break large quadratic programming (QP) problem into small QP problems

- New items are classified by using $w \cdot x$.

- Fit a sigmoid to output by maximum likelihood and produce posterior probabilities.

- $c$, the penalty imposed on training instances.
  - Default value 0.01.

- $p$, the decision threshold to control precision and recall
  - Flat non-hierarchical models, $p=0.5$
  - Hierarchical models with multiplicative decision rule, $p=0.2$
  - For the Boolean decision rule, $p_1=0.2$ (first level) and $p_2=0.5$ (second level)
Experimental Results and Evaluation [1]
Experimental Results and Evaluation [1]

- **Accuracy**

  Measure: $F_1 = \frac{2PR}{P+R}$  
P: precision, R: recall
Experimental Results and Evaluation [1]

- **Accuracy**
  
  Measure: $F_1 = \frac{2PR}{P+R}$  
  P: precision, R: recall
  
  Top level  
  .572
  
  By optimizing the C parameter, 15%~20% improvement
  
  Non-hierarchical  
  .476  
  baseline
Experimental Results and Evaluation [1]

- **Accuracy**

  Measure: $F_1 = \frac{2PR}{P+R}$  
  P: precision, R: recall

  Top level  
  0.572

  By optimizing the C parameter, 15%~20% improvement

  Non-hierarchical  
  0.476  
  baseline

  $P(L_1) \times P(L_2)$  
  0.495  
  4% improvement

  $P(L_1) \& \& P(L_2)$  
  0.497  
  4.4% improvement

  No difference between Boolean rule and multiplicative scoring rule
Experimental Results and Evaluation [1]

Accuracy

Measure: $F1 = \frac{2PR}{P+R}$  
P: precision, R: recall

Top level: .572

By optimizing the C parameter, 15%~20% improvement

Non-hierarchical: .476 baseline

$P(L1) \cdot P(L2)$: .495 4% improvement

$P(L1) \& \& P(L2)$: .497 4.4% improvement

No difference between Boolean rule and multiplicative scoring rule

Assuming $P(L1)=1$: .711 49.34% !!!
Best and Worst F1 Score by Category
<table>
<thead>
<tr>
<th>Category Name</th>
<th>Best F1</th>
<th>Worst F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health &amp; Fitness/Drugs &amp; Medicines</td>
<td>0.841</td>
<td></td>
</tr>
<tr>
<td>Home &amp; Family/Real Estate</td>
<td>0.797</td>
<td></td>
</tr>
<tr>
<td>Reference &amp; Education/K-12 Education</td>
<td>0.781</td>
<td></td>
</tr>
<tr>
<td>Sports &amp; Recreation/Fishing</td>
<td>0.750</td>
<td></td>
</tr>
<tr>
<td>Reference &amp; Education/Higher &amp; Cont. Ed.</td>
<td>0.741</td>
<td></td>
</tr>
<tr>
<td>Society &amp; Politics/World Cultures</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>Home &amp; Family/For Kids</td>
<td>0.088</td>
<td></td>
</tr>
<tr>
<td>Computers &amp; Internet/News &amp; Magazines</td>
<td>0.122</td>
<td></td>
</tr>
<tr>
<td>Computers &amp; Internet/Internet &amp; the Web</td>
<td>0.131</td>
<td></td>
</tr>
<tr>
<td>Business &amp; Finance/Business Professions</td>
<td>0.133</td>
<td></td>
</tr>
</tbody>
</table>
Performance varies widely across categories

- Difficult categories are based on non-content distinctions.

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Best F1</th>
<th>Worst F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health &amp; Fitness/Drugs &amp; Medicines</td>
<td>0.841</td>
<td>0.034</td>
</tr>
<tr>
<td>Home &amp; Family/Real Estate</td>
<td>0.797</td>
<td>0.088</td>
</tr>
<tr>
<td>Reference &amp; Education/K-12 Education</td>
<td>0.781</td>
<td>0.122</td>
</tr>
<tr>
<td>Sports &amp; Recreation/Fishing</td>
<td>0.750</td>
<td>0.131</td>
</tr>
<tr>
<td>Reference &amp; Education/Higher &amp; Cont. Ed.</td>
<td>0.741</td>
<td>0.133</td>
</tr>
<tr>
<td>Society &amp; Politics/World Cultures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home &amp; Family/For Kids</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computers &amp; Internet/News &amp; Magazines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computers &amp; Internet/Internet &amp; the Web</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business &amp; Finance/Business Professions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results and Evaluation [2]

Efficiency

Offline training

Measured by the CPU seconds of the standard 266MHz Pentium II PC running WinNT

The total training time for all the models in the paper was only about 30 minutes

<table>
<thead>
<tr>
<th>Description</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>For all the non-hierarchical models on 50078 examples</td>
<td>729</td>
</tr>
<tr>
<td>For all 150 hierarchical second-level models</td>
<td>128</td>
</tr>
<tr>
<td>For the 13 top-level categories</td>
<td>1258</td>
</tr>
</tbody>
</table>
Online evaluation

focus on the cost of operation on the dot product of eight vector and the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>Each instance Compared to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-hierarchical</td>
<td>150 second-level categories</td>
</tr>
<tr>
<td>Hierarchical, the multiplicative scoring</td>
<td>13 first-level and all 150 second-level categories</td>
</tr>
<tr>
<td>rule</td>
<td></td>
</tr>
<tr>
<td>Hierarchical, the Boolean scoring rule</td>
<td>all 13 first-level but only the second-level categories pass</td>
</tr>
<tr>
<td></td>
<td>the first-level criterion</td>
</tr>
</tbody>
</table>

More speed gain in the Boolean decision rule, overall 14.8% of the multiplicative case, 16% of the non-hierarchical case.
Yet another Hierarchical SVM (vs. Flat)
Y Yang et al (2005) investigated the efficiency and effectiveness of SVMs over very large-scale taxonomies:

- Yahoo! Directory: classify 792,601 documents into 292,216 categories within 16 levels;
- Cluster of SVMs
  10 x 3GHz CPU and 4GB Ram
Y Yang et al (2005) investigated the efficiency and effectiveness of SVMs over very large-scale taxonomies:

- Yahoo! Directory: classify 792,601 documents into 292,216 categories within 16 levels;
- Cluster of SVMs
  10 x 3GHz CPU and 4GB Ram

Effectiveness

- Benefit balanced local feature sets
Y Yang et al (2005) investigated the efficiency and effectiveness of SVMs over very large-scale taxonomies:

- Yahoo! Directory: classify 792,601 documents into 292,216 categories within 16 levels;
- Cluster of SVMs
  - 10 x 3GHz CPU and 4GB Ram

**Effectiveness**
- Benefit balanced local feature sets

**Efficiency**
- Save 90% training time

### Time complexity of SVMs ($h$)

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Training without SCut</th>
<th>Training with SCut</th>
<th>Testing with SCut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical SVMs</td>
<td>0.42</td>
<td>2.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Flat SVMs</td>
<td>310.87</td>
<td>1304.64</td>
<td>53.63</td>
</tr>
</tbody>
</table>
Conclusion

- Novel approach, much more efficient on large taxomony
- Balance feature sets, better effectiveness
- Future works: generalised sequential model in hierarchical classification
Conclusion

- Novel approach, much more efficient on large taxonomy
- Balance feature sets, better effectiveness
- Future works: generalised sequential model in hierarchical classification
- Questions?
Reference

- Y Yang et al, Support vector machines classification with a very large-scale taxonomy, ACM SIGKDD Explorations Newsletter, 2005
- T Joachims, Text categorization with support vector machines: learning with many relevant features, ECML-98, 1998
Thank you