Categorization: Information and Misinformation

Paul Thompson

7 November 2001
Outline

• Supervised machine learning in an industry setting
  – Categorization of case law
  – Categorization of statutes
• Categorization applied to misinformation
Categorization of Case Law (ICAIL 2001)

- Query-Based or k-Nearest Neighbor-like
- Knowledge-Based, Topical View Queries
- Machine Learning
  - Decision Trees: C4.5, C5.0
  - Rule-based: Ripper
Categorization of Case Law (cont.)

- Routing incoming case to 40 broad topical areas, e.g., bankruptcy
- Query-based Approach: Treat incoming case as NL query by extracting terms (10-300)
- Machine Learning Approach: C4.5
  - training on 7,535 labeled example cases
Categorization of Case Law – Issues with data

• Started with over 11,000 marked cases, not 7,535
• Cases categorized by several means
  – By attorney/editors
  – By topical view queries
  – Other
• Only want cases categorized by editors
Cross-validation for 8 Categories with the most data: C4.5

<table>
<thead>
<tr>
<th>Category</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bankruptcy</td>
<td>66.14</td>
<td>87.86</td>
</tr>
<tr>
<td>Constitutional Law &amp; Theory</td>
<td>46.19</td>
<td>55.61</td>
</tr>
<tr>
<td>Criminal Justice</td>
<td>78.56</td>
<td>79.50</td>
</tr>
<tr>
<td>Labor &amp; Employment</td>
<td>61.87</td>
<td>72.46</td>
</tr>
<tr>
<td>Legal Ethics &amp; Professional</td>
<td>17.06</td>
<td>69.29</td>
</tr>
<tr>
<td>Responsibility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Americans Law</td>
<td>82.05</td>
<td>77.84</td>
</tr>
<tr>
<td>Taxation</td>
<td>82.92</td>
<td>82.71</td>
</tr>
<tr>
<td>Transportation</td>
<td>41.52</td>
<td>55.58</td>
</tr>
<tr>
<td><strong>Averages</strong></td>
<td>59.54</td>
<td>72.61</td>
</tr>
<tr>
<td><strong>Averages on all 40 Categories</strong></td>
<td>42.36</td>
<td>67.49</td>
</tr>
</tbody>
</table>
Comparison: C4.5, Ripper, and Topical View Queries

<table>
<thead>
<tr>
<th>Categorization Method</th>
<th>Recall</th>
<th>Percent Change</th>
<th>Precision</th>
<th>Percent Change</th>
<th>Error</th>
<th>Percent Change</th>
<th>F-measure b = 2</th>
<th>Percent Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5 release 8 (baseline)</td>
<td>47.39</td>
<td></td>
<td>67.34</td>
<td></td>
<td>3.80</td>
<td></td>
<td>0.4954</td>
<td></td>
</tr>
<tr>
<td>C4.5 release 8 exp. frequency</td>
<td>51.39</td>
<td>8.44</td>
<td>67.98</td>
<td>0.95</td>
<td>3.68</td>
<td>(3.16)</td>
<td>0.5348</td>
<td>7.95</td>
</tr>
<tr>
<td>Ripper exp. frequency L1</td>
<td>52.83</td>
<td>11.48</td>
<td>69.26</td>
<td>2.85</td>
<td>3.54</td>
<td>(6.84)</td>
<td>0.5446</td>
<td>9.93</td>
</tr>
<tr>
<td>Ripper exp. frequency L0.25</td>
<td>67.32</td>
<td>42.06</td>
<td>51.88</td>
<td>(22.96)</td>
<td>4.50</td>
<td>18.42</td>
<td>0.6376</td>
<td>28.70</td>
</tr>
<tr>
<td>Official MTV: Unaltered</td>
<td>60.23</td>
<td>27.09</td>
<td>54.68</td>
<td>(18.80)</td>
<td>4.79</td>
<td>26.05</td>
<td>0.5784</td>
<td>16.75</td>
</tr>
<tr>
<td>Official MTV: Altered</td>
<td>57.05</td>
<td>20.38</td>
<td>55.00</td>
<td>(18.32)</td>
<td>4.78</td>
<td>25.79</td>
<td>0.5540</td>
<td>11.83</td>
</tr>
</tbody>
</table>

Average Performance for 18 Topics
Categorization of Statutes (SIG/CR 1997)

- Machine Learning Approaches: C4.5 and Ripper
  - Training on 125,180 labeled example statute sections (5 states)
  - Testing on 24,475 labeled statute sections (6th state)
  - 35 Categories
<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counties</td>
<td>48.09</td>
<td>12.29</td>
<td>30.39</td>
</tr>
<tr>
<td>Sales</td>
<td>73.68</td>
<td>13.00</td>
<td>38.11</td>
</tr>
<tr>
<td>Worker’s Compensation</td>
<td>78.71</td>
<td>91.73</td>
<td>81.01</td>
</tr>
<tr>
<td>Insurance</td>
<td>83.19</td>
<td>84.56</td>
<td>83.46</td>
</tr>
<tr>
<td>Warehouse Receipts</td>
<td>92.59</td>
<td>86.21</td>
<td>91.24</td>
</tr>
<tr>
<td>Overall</td>
<td>67.62</td>
<td>41.61</td>
<td>56.35</td>
</tr>
</tbody>
</table>
Categorization of *Urban Renewal*
Automatic and Manual
Automatic Categorization of “Urban Renewal” – C4.5

C 4.5 [release 8] rule generator Tue Sep 24 12:20:44 1996

Options:
  File stem <siml017>
  Rulesets evaluated on unseen cases

Read 125180 cases (71 attributes) from siml017

Processing tree 0

Final rules from tree 0:
Rule 19:
agenc > 1.5018
blight <= 1.1588
development <= 0.7393
redevelop <= 1.0815
renew > 0.3004
urban > 0.4519
-> class 383 [86.7%]

Rule 25:
decad <= 0.9587
oper > 1.8955
renew > 0.3004
urban > 0.4519
-> class 383 [77.1%]
Rule 34:
  decad > 0.9587
  renew > 0.6823
  time <= 0.7423
  -> class 383 [73.1%]

Rule 17:
  agenc <= 1.5018
  decad <= 0.9587
  grant > 4.1484
  land <= 0.081
  renew > 0.3004
  urban > 0.4519
  -> class 383 [70.7%]
Rule 29:
  blight > 1.1588
  house > 2.3675
  house <= 2.8787
  ->  class 383  [63.0%]

Rule 23:
  blight <= 1.1588
  oper <= 1.8955
  redevelop > 1.0815
  renew > 0.3004
  resolv <= 0.307
  ->  class 383  [56.6%]
Rule 14:
  authority > 1.5306
  fami > 2.1165
  urban > 0.4519
  -> class 383 [50.0%]

Rule 16:
  reloc > 2.8552
  resident > 0.4041
  urban > 0.4519
  -> class 383 [50.0%]
Rule 27:
  blight > 1.1588
  zone > 0.2438
  zone <= 0.2956
  -> class 383 [50.0%]

Rule 31:
  blight > 1.1588
  tax > 1.7082
  zone > 0.2438
  -> class 383 [50.0%]
Rule 8:
  neighborhood > 3.4954
  neighborhood <= 4.9301
  plan > 1.6727
  -> class 383  [35.2%]

Rule 10:
  blight <= 1.1588
  urban <= 0.4519
  -> class non383  [100.0%]
Rule 12:
    agenc <= 1.5018
    blight <= 1.1588
    fami <= 2.1165
    grant <= 4.1484
    redevelop <= 1.0815
    resident <= 0.4041
-> class non383 [100.0%]

Rule 11:
    blight <= 1.1588
    renew <= 0.3004
-> class non383 [100.0%]

Default class: non383
Evaluation on training data (125180 items):

<table>
<thead>
<tr>
<th>Rule</th>
<th>Size</th>
<th>Error</th>
<th>Used</th>
<th>Wrong</th>
<th>Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>6</td>
<td>13.3%</td>
<td>19</td>
<td>1 (5.3%)</td>
<td>3 (4</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>22.9%</td>
<td>5</td>
<td>2 (40.0%)</td>
<td>1 (3</td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>26.9%</td>
<td>7</td>
<td>1 (14.3%)</td>
<td>5 (6</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>29.3%</td>
<td>4</td>
<td>0 (0.0%)</td>
<td>3 (3</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>37.0%</td>
<td>3</td>
<td>0 (0.0%)</td>
<td>3 (3</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>43.4%</td>
<td>7</td>
<td>2 (28.6%)</td>
<td>3 (5</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>50.0%</td>
<td>2</td>
<td>0 (0.0%)</td>
<td>2 (2</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>50.0%</td>
<td>2</td>
<td>0 (0.0%)</td>
<td>2 (2</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>50.0%</td>
<td>2</td>
<td>0 (0.0%)</td>
<td>2 (2</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>50.0%</td>
<td>2</td>
<td>0 (0.0%)</td>
<td>2 (2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>64.8%</td>
<td>4</td>
<td>2 (50.0%)</td>
<td>0 (2</td>
</tr>
</tbody>
</table>
10  2  0.0%  124268  25 (0.0%)  0 (0|0) non383
12  6  0.0%   676     3 (0.4%)   0 (0|0) non383
11  2  0.0%   111     2 (1.8%)   0 (0|0) non383

Tested 125180, errors 39 (0.0%)  <<

(a) (b)  <-classified as
        ---- ----
   49   31 (a): class 383
  8125092   (b): class non383
Evaluation on test data (24475 items):

<table>
<thead>
<tr>
<th>Rule</th>
<th>Size</th>
<th>Error</th>
<th>Used</th>
<th>Wrong</th>
<th>Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>6</td>
<td>13.3%</td>
<td>1</td>
<td>1</td>
<td>(100.0%)</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>37.0%</td>
<td>1</td>
<td>0</td>
<td>(0.0%)</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>43.4%</td>
<td>6</td>
<td>1</td>
<td>(16.7%)</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>64.8%</td>
<td>2</td>
<td>2</td>
<td>(100.0%)</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.0%</td>
<td>24360</td>
<td>39</td>
<td>(0.2%)</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>0.0%</td>
<td>78</td>
<td>1</td>
<td>(1.3%)</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.0%</td>
<td>19</td>
<td>1</td>
<td>(5.3%)</td>
</tr>
</tbody>
</table>
Tested 24475, errors 51 (0.2%)  

(a) (b)  <-classified as
---- ----
  6  47    (a): class 383
  4 24418   (b): class non383

Recall: 11.32%  Precision: 60.00%
Manual Rule Set for *Urban Renewal*

Urban (132)
   Renewal (13)
      Plan > class 383 (7 out of 8 correct)
      Project > class 383 (2 out of 2 correct)
      Blight > class 383 (0 cases)

Slum (27)
      Clearance > class 383 (23 out of 23 correct)
      Urban > class 383 (1 out of 1 correct)

Blighted (3)
      Area > class 383 (3 out of 3 correct)

>not class 383
Manual Rule Set (cont.)

The number in parentheses in a nonterminal node indicates how many docs satisfied the node. Searching was adjusted so that a document once classified was removed from consideration.

- 36 17
- 1 24,385

error rate
0.97 precision
0.68 recall

Recall: 67.92%  Precision: 97.30%

Error Rate: 0.07%
Questionable Validity of Training and Testing Data

• Statutes editors: three phases of index assignment
  – Phase 1: list all possible category assignments
  – Phase 2: eliminate categories not suited to print environment
  – Phase 3: concatenate categories hierarchically

• Arguably automatic categorization is phase 1, but this data not retained
Any employer requiring a physical examination of an employee shall, upon request, cause said person to be furnished with a copy of the medical report following the said examination.
Categorization Applied to Misinformation

• Text may be deliberately misleading
  – Web search engine optimization (Lynch JASIS&T vol. 52 no. 1, 2001)
  – Computer virus hoaxes
  – Stock market fraud
  – Information warfare
  – Manipulating users’ perceptions

• Semantic Hacking project – ISTS