

# 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

	Recall	Precision
Bankruptcy	66.14	87.86
Constitutional Law & Theory	46.19	55.61
Criminal Justice	78.56	79.50
Labor & Employment	61.87	72.46
Legal Ethics & Professional Responsibility	17.06	69.29
Native Americans Law	82.05	77.84
Taxation	82.92	82.71
Transportation	41.52	55.58
-----		
Averages	59.54	72.61
Averages on all 40 Categories	42.36	67.49

# Comparison: C4.5, Ripper, and Topical View Queries

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

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 (6<sup>th</sup> state)
  - 35 Categories

# Results for C4.5

Category	Precision	Recall	F
			$b = 1/2$
Counties	48.09	12.29	30.39
Sales	73.68	13.00	38.11
Worker's Compensation	78.71	91.73	81.01
Insurance	83.19	84.56	83.46
Warehouse Receipts	92.59	86.21	91.24
Overall	67.62	41.61	56.35

# Categorization of *Urban Renewal* Automatic and Manual

# Automatic Categorization of “Urban Renewal” – C4.5

C 4 . 5 [ r e l e a s e 8 ] r u l e g e n e r a t o r      T u e S e p  
2 4 1 2 : 2 0 : 4 4 1 9 9 6

---

O p t i o n s :

File stem <sim 10 1 7 >

Rulesets evaluated on unseen cases

R e a d 1 2 5 1 8 0 c a s e s ( 7 1 a t t r i b u t e s ) f r o m  
s i m 1 0 1 7

---

P r o c e s s i n g t r e e 0

F i n a l r u l e s f r o m t r e e 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):

Rule Size Error Used Wrong Advantage

-----	-----	-----	-----	-----	-----	-----
19	6	13.3%	19	1 (5.3%)	3 (4 1)	383
25	4	22.9%	5	2 (40.0%)	1 (3 2)	383
34	3	26.9%	7	1 (14.3%)	5 (6 1)	383
17	6	29.3%	4	0 (0.0%)	3 (3 0)	383
29	3	37.0%	3	0 (0.0%)	3 (3 0)	383
23	5	43.4%	7	2 (28.6%)	3 (5 2)	383
14	3	50.0%	2	0 (0.0%)	2 (2 0)	383
16	3	50.0%	2	0 (0.0%)	2 (2 0)	383
27	3	50.0%	2	0 (0.0%)	2 (2 0)	383
31	3	50.0%	2	0 (0.0%)	2 (2 0)	383
8	3	64.8%	4	2 (50.0%)	0 (2 2)	383

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):

Rule	Size	Error	Used	Wrong	Advantage
----	----	-----	----	-----	-----
19	6	13.3%	1	1 (100.0%)	-1 (0 1) 383
29	3	37.0%	1	0 (0.0%)	1 (1 0) 383
23	5	43.4%	6	1 (16.7%)	4 (5 1) 383
8	3	64.8%	2	2 (100.0%)	-2 (0 2) 383
10	2	0.0%	24360	39(0.2%)	0 (0 0) non383
12	6	0.0%	78	1 (1.3%)	0 (0 0) non383
11	2	0.0%	19	1 (5.3%)	0 (0 0) non383

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

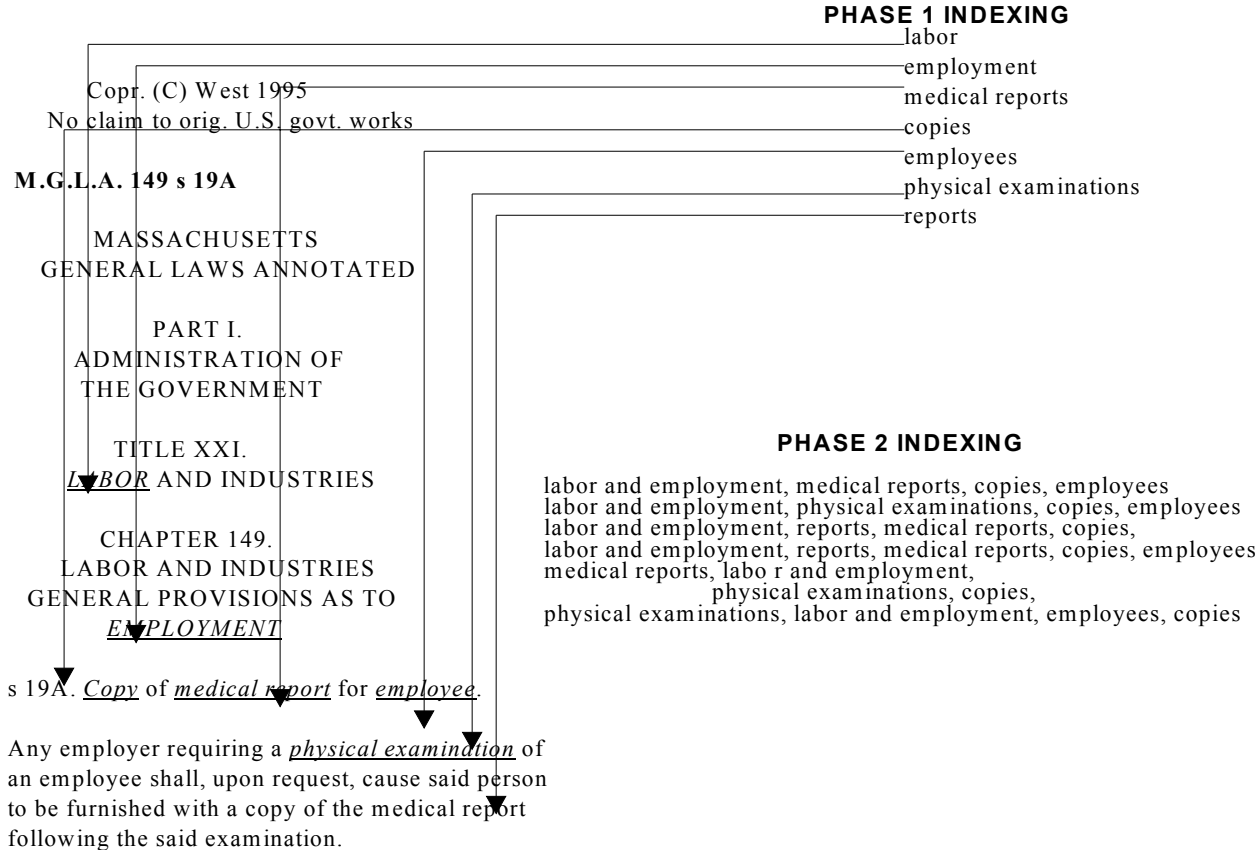
error rate  
0.97 precision  
0.68 recall

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



**Figure 4.** The 3 Phases of Statute Indexing

# 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