Website Classification

Mgr. Juraj Hreško`s thesis

7.2.2013 presented by Jaromír Navrátil

Synopsis

- Task
- Possible solutions
- Solution
- Rare classes
- Possible improvements
- Rewriting to C++

The Task

- create application to classify czech websites
- 61 classes
- multi-labeling (1-3 classes for each document)
- real-time classification
- be able to adjust the algorithm to maximize precision or recall

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1\ measure = 2 * \frac{precision\ *\ recall}{precision\ +\ recall}$$

Classes

Occurrences	Class
3159	Advertisement
1281	Alcohol / Tobacco
2442	Arts
9756	Cars / Vehicles
1590	Banking
450	Brokers
27066	Building / Home
15045	Business
16998	Chats / Blogs / Forums
1068	Communications
72	Crime
11805	Education
2613	Entertainment
5553	Environment
1575	Erotic / Adult / Nudity
459	Extreme / Hate / Violence
13302	Fashion / Beauty
12708	Food / Restaurants
2298	Foundations / Charity / Social Services
135	Gambling
3090	Games
6108	Government
18	Hacking / Phishing / Fraud
9225	Health / Medicine
13794	Hobbies
2376	Humour / Cool
13995	IT / Hardware / Software
5163	IT Services / Internet
195	Illegal Drugs
90	Instant Messaging

Occurrences	Class
678	Insurance
1170	Job / Career
6003	Kids / Toys / Family
1059	Military / Guns
1974	Mobile Phones / Operators
11826	Music / Radio / Cinema / TV
3477	News / Magazines
54	Peer-to-peer
10002	Personal / Dating / Lifestyle
2049	Politics / Law
4077	Pornography
4227	Portals / Search Engines
90	Proxies
2475	Real Estate
6966	Regional
1803	Religious / Spirituality
6405	Sale / Auctions
6	Sects
48	Sex Education
42240	Shopping
288	Social Networks
14913	Sports
120	Streaming / Broadcasting
951	Swimwear / Intimate
384	Translation Services
24537	Travelling / Vacation
1788	Uploading / Downloading
816	Warez / Piracy
135	Web Based Mail
888	Web Hosting
1110	Money / Financial

Possible Approaches

- Web structure mining - links
- Web content mining - text, html, multimedia
- Web usage mining - access logs
- combining first two approaches would be ideal, but mining from structure is computationally difficult

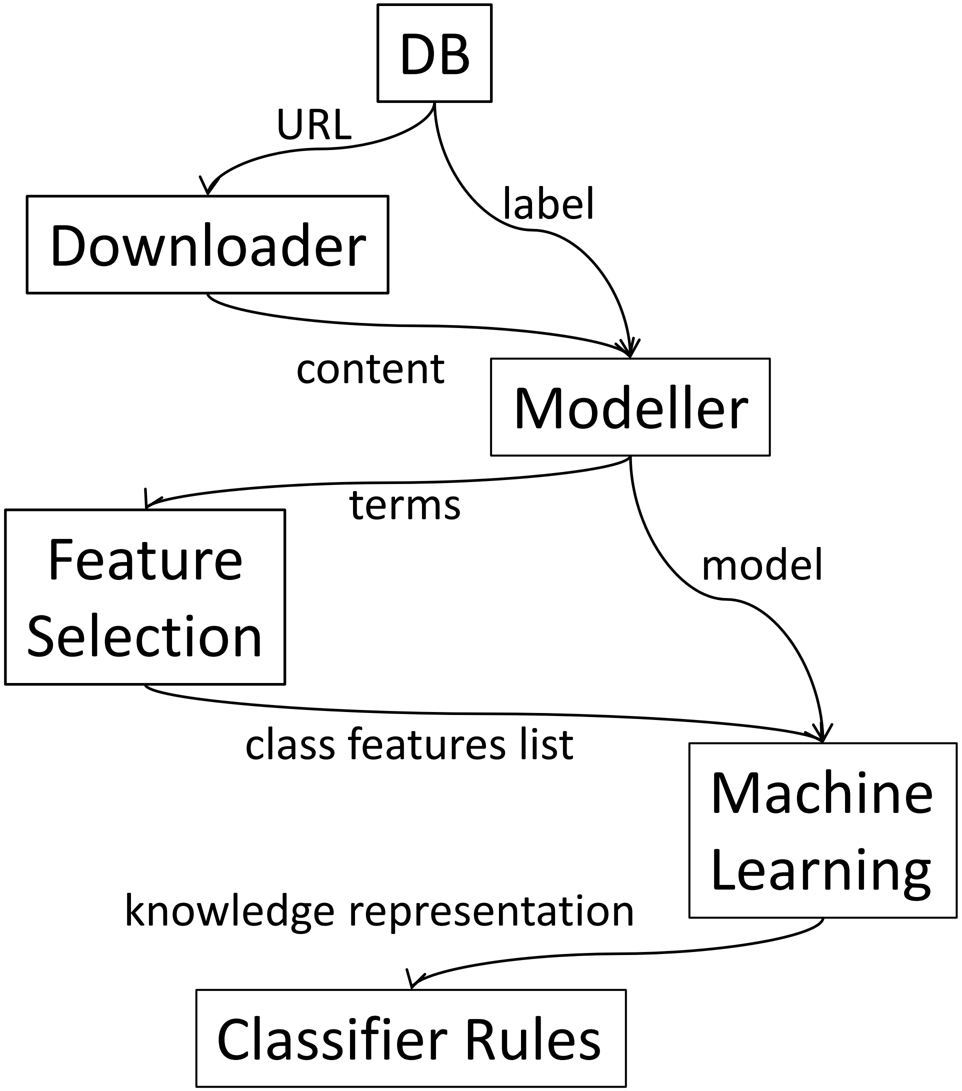
Multi-label Classification

- Algorithms from WEKA are not able to process multi-label data, thus we have to transform the problem or adapt the algorithm
- Transforming the problem:
  - choose one class for each example, forgetting others
  - delete all examples with more than one class
  - change every combination of classes into one new class
  - use one classificator for each class

Categories	Occurence
0	0.41%
1	64.45%
2	31.75%
3	3.38%
4-6	0.01%

Components of Classifier

- downloader
- modeller
- feature selection
- machine learning



Downloader

- download website using wget
- get language coding (mostly Windows-1250, ISO 8859-2 or UTF-8)
- transfer to UTF-8 using Enca

Modeller - source to vertical

- transfer text to so-called vertical
- delete HTML tags, scripts, parts of CSS, interpunction
- divide words by spaces and convert them to lower-case

page source

```
<html>
    <head>
        <title>Interesting article.<title>
    </head>
    <body>
        <h1>The article</h1>
        This is the main part of the article.
    </body>
</html>
```

vertical

word	Tag
interesting	title
article	title
the	h1
article	h1
this	none
is	none
the	none
main	none
part	none
of	none
article	none

Modeller - vertical to vector

- trasfer vertical into vector model using Structure-oriented Weighting Technique
- delete words with high frequency across classes – not used
- stemming (lemmatization) – not used

vertical

word	Tag
interesting	title
article	title
the	h1
article	h1
this	none
is	none
the	none
main	none
part	none
of	none
article	none

$$SWT_w(t_i, d_j) = \sum_{e_k} (w(e_k) * TF(t_i, e_k, d_j))$$

e_k	title	h1	h2	h3	none
$w(e_k)$	10	5	3	2	1

vector model

word	weight
article	16
interesting	10
the	6
this	1
is	1
main	1
part	1
of	1
article	1

Feature Selection

- eliminate attributes with fewer than 50 ocurrences, lessening number of words in dictionary from 1 263 296 to 63 121
- compute information gain for each term
- choose 2000 best terms

Choosing Classifier

- choose 5 categories with average number of positive and negative examples

SVM - sigmoid

Category	Precision	Recall	F_1 Measure
Arts	0.831	0.814	0.812
Entertainment	0.793	0.768	0.763
Foundations	0.811	0.792	0.789
Games	0.787	0.768	0.765
HW-SW	0.767	0.764	0.763
Mean	0.798	0.781	0.778

SVM - linear

Category	Precision	Recall	F_1 Measure
Arts	0.812	0.810	0.810
Entertainment	0.767	0.766	0.766
Foundations	0.766	0.764	0.764
Games	0.814	0.811	0.811
HW-SW	0.782	0.782	0.781
Mean	0.788	0.787	0.786

Random forest

Category	Precision	Recall	F_1 Measure
Arts	0.851	0.848	0.847
Entertainment	0.817	0.815	0.815
Foundations	0.821	0.821	0.821
Games	0.851	0.851	0.851
HW-SW	0.843	0.842	0.841
Mean	0.837	0.835	0.835

J48

Category	Precision	Recall	F_1 Measure
Arts	0.792	0.790	0.790
Entertainment	0.762	0.759	0.758
Foundations	0.761	0.758	0.758
Games	0.798	0.797	0.796
HW-SW	0.809	0.807	0.807
Mean	0.784	0.782	0.782

Naive Bayes

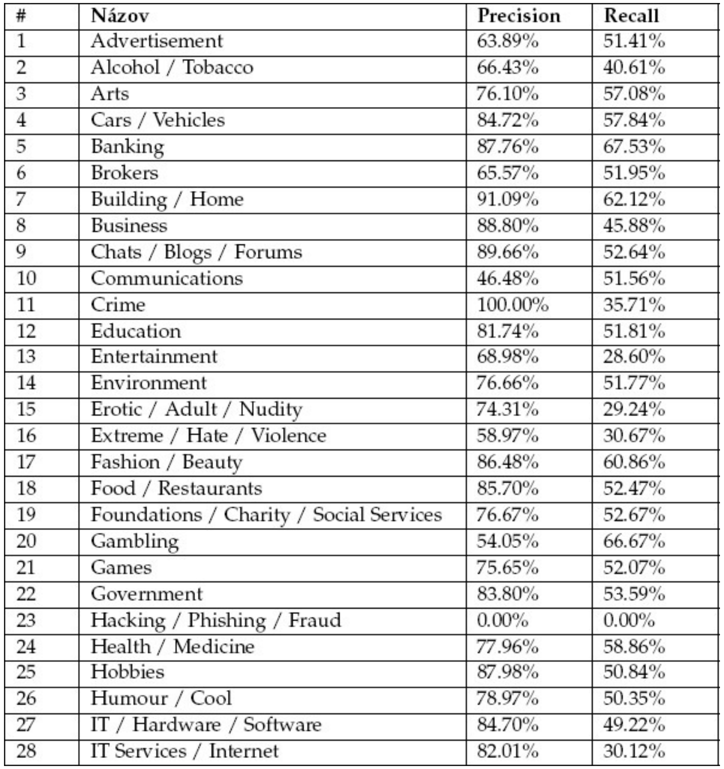
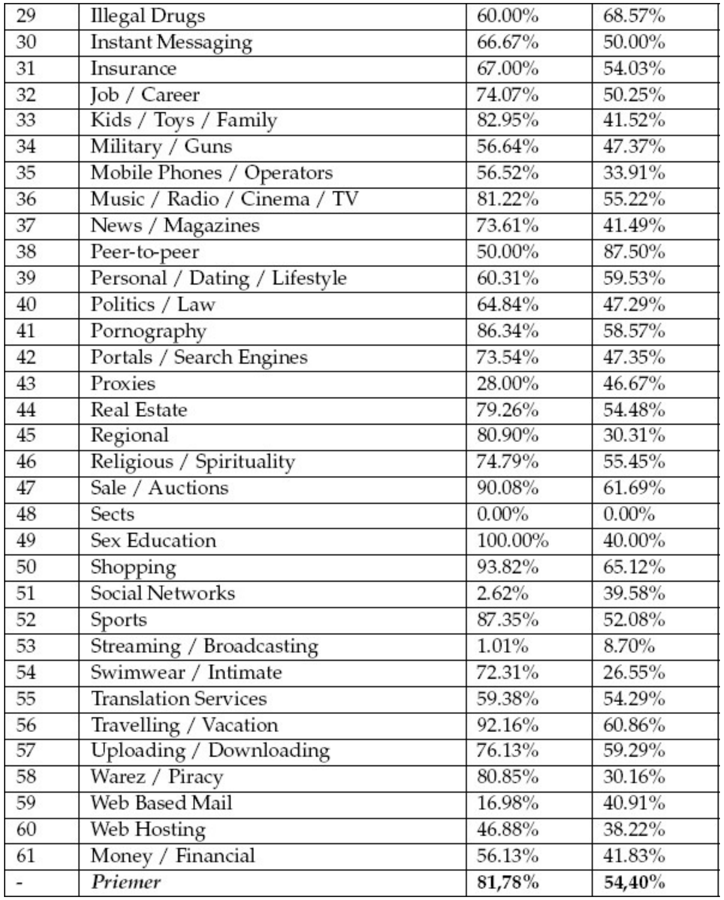
Category	Precision	Recall	F_1 Measure
Arts	0.762	0.762	0.759
Entertainment	0.765	0.761	0.760
Foundations	0.742	0.742	0.741
Games	0.763	0.744	0.740
HW-SW	0.792	0.782	0.780
Mean	0.765	0.758	0.756

Random Forest

- **number of randomly selected attributes (constant k) was set to 50, as well as number of trees**
- $1 < k \leq \log_2(|A| + 1)$, A is set of attributes
- **rate Positive : Negative was set to 1:5 using meta classificator**

System Evaluation

- cross-validation
  - training : testing data set to 1:4
  - precision 59.68%
- second approach took each class as one problem

#	Názov	Precision	Recall
1	Advertisement	63.89%	51.41%
2	Alcohol / Tobacco	66.43%	40.61%
3	Arts	76.10%	57.08%
4	Cars / Vehicles	84.72%	57.84%
5	Banking	87.76%	67.53%
6	Brokers	65.57%	51.95%
7	Building / Home	91.09%	62.12%
8	Business	88.80%	45.88%
9	Chats / Blogs / Forums	89.66%	52.64%
10	Communications	46.48%	51.56%
11	Crime	100.00%	35.71%
12	Education	81.74%	51.81%
13	Entertainment	68.98%	28.60%
14	Environment	76.66%	51.77%
15	Erotic / Adult / Nudity	74.31%	29.24%
16	Extreme / Hate / Violence	58.97%	30.67%
17	Fashion / Beauty	86.48%	60.86%
18	Food / Restaurants	85.70%	52.47%
19	Foundations / Charity / Social Services	76.67%	52.67%
20	Gambling	54.05%	66.67%
21	Games	75.65%	52.07%
22	Government	83.80%	53.59%
23	Hacking / Phishing / Fraud	0.00%	0.00%
24	Health / Medicine	77.96%	58.86%
25	Hobbies	87.98%	50.84%
26	Humour / Cool	78.97%	50.35%
27	IT / Hardware / Software	84.70%	49.22%
28	IT Services / Internet	82.01%	30.12%
29	Illegal Drugs	60.00%	68.57%
30	Instant Messaging	66.67%	50.00%
31	Insurance	67.00%	54.03%
32	Job / Career	74.07%	50.25%
33	Kids / Toys / Family	82.95%	41.52%
34	Military / Guns	56.64%	47.37%
35	Mobile Phones / Operators	56.52%	33.91%
36	Music / Radio / Cinema / TV	81.22%	55.22%
37	News / Magazines	73.61%	41.49%
38	Peer-to-peer	50.00%	87.50%
39	Personal / Dating / Lifestyle	60.31%	59.53%
40	Politics / Law	64.84%	47.29%
41	Pornography	86.34%	58.57%
42	Portals / Search Engines	73.54%	47.35%
43	Proxies	28.00%	46.67%
44	Real Estate	79.26%	54.48%
45	Regional	80.90%	30.31%
46	Religious / Spirituality	74.79%	55.45%
47	Sale / Auctions	90.08%	61.69%
48	Sects	0.00%	0.00%
49	Sex Education	100.00%	40.00%
50	Shopping	93.82%	65.12%
51	Social Networks	2.62%	39.58%
52	Sports	87.35%	52.08%
53	Streaming / Broadcasting	1.01%	8.70%
54	Swimwear / Intimate	72.31%	26.55%
55	Translation Services	59.38%	54.29%
56	Travelling / Vacation	92.16%	60.86%
57	Uploading / Downloading	76.13%	59.29%
58	Warez / Piracy	80.85%	30.16%
59	Web Based Mail	16.98%	40.91%
60	Web Hosting	46.88%	38.22%
61	Money / Financial	56.13%	41.83%
-	*Priemer*	**81,78%**	**54,40%**

Complications

- classes with very low number of positive examples
- some pages stopped existing
- system cannot handle HTTPS protocol, nor redirection
- existing solution was very slow when it came to classifying multiple webpages

Rare Classes

- task is to examine two classes with low occurence
- Illegal Drugs (418 URLs)
  - some pages do not exist anymore, are redirected or requires confirmation
  - only 96 pages (23%) were classified correctly
- Alcohol / Tobacco (5631 URLs)
  - some websites caused utility wget to enter infinite loop
  - 2289 pages (41%) classified correctly
  - category Shopping assigned many times, along with Social Networks

Rare Classes - data

- classification of six thousand pages runned for about 18 hours (it would be much longer if SSD was not used)

Illegal Drugs (418 examples)

Category	Times Assigned
Illegal Drugs	96
Shopping	56
Health / Medicine	43
Social Networks	39
Chats / Blogs / Forums	19
Alcohol / Tobacco	11
News / Magazines	10
Streaming / Broadcasting	10
other (classified < 10 times)	122
empty pages	59

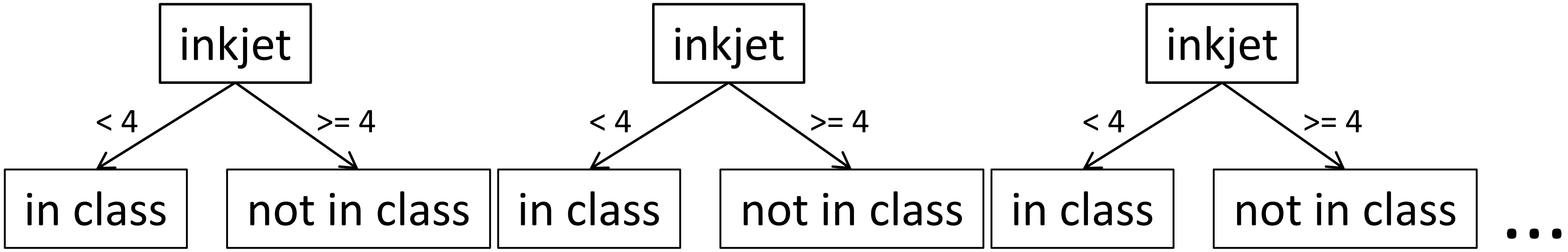
Alcohol / Tobacco (5631 examples)

Category	Times Assigned
Alcohol / Tobacco	2289
Shopping	648
Social Networks	461
Food / Restaurants	316
Health / Medicine	203
Travelling / Vacacion	167
Chats / Blogs / Forums	105
Streaming / Broadcasting	51
other (classified < 50 times)	375
empty pages	402

Possible Improvements

- remove obstacles preventing downloading some pages, such as use of HTTPS, redirection, age prompt
- relearn forests using verified data
- use faster classifier or parallelize Random Forest
- rewrite system from Python and Bash to C++
- improve feature selection

Forest classifying rare class Sects



Rewriting to C++

- rapid increase of speed (now 42 examples per min., was 5.5)
- somehow different results using same URLs

former solution (~1h 15min)

Category	Times Assigned
Illegal Drugs	96
Shopping	56
Health / Medicine	43
Social Networks	39
Chats / Blogs / Forums	19
Alcohol / Tobacco	11
News / Magazines	10
Streaming / Broadcasting	10
other (classified < 10 times)	122

C++ version (9min 50s)

Category	Times Assigned
Social Networks	128
Illegal Drugs	81
Health / Medicine	33
Shopping	31
Chats / Blogs / Forums	17
Alcohol / Tobacco	9
Web Based Mail	7
Streaming / Broadcasting	7
other (classified < 7 times)	58

Conclusion

- rewriting system to C++ made it viable for real-time application
- the main problem is preprocessing now
  - downloading webpage takes most time
  - using more pages from same domain could improve accuracy
  - utility wget enters infinite loop on some sites
- classifier itself could be improved as well
  - independent list of attributes for each class
  - another algorithm can be tried (e.g. Bayesian classifier)
- dividing program into parts operating independently would slightly improve speed

Sources

- Thesis of Mgr. Juraj Hreško, Masaryk University, Faculty of Informatics, Brno, 2012
- http://weka.wikispaces.com/
- http://www.ide.bth.se/~hgr/Papers/cuda-rf_mcc10_v1.0_crc.pdf
- http://www.cplusplus.com/