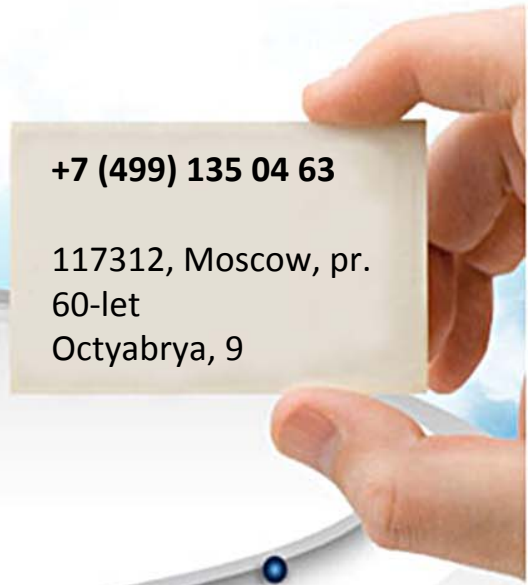
Institute for Systems Analysis
of Russian Academy of Sciences

Training Datasets Collection and Evaluation of Feature Selection Methods for Web Content Filtering

Roman Suvorov
Ilya Sochenkov
Ilya Tikhomirov



AIMSA 2014, Varna, Bulgaria

Introduction: Content Filtering

WebSpy[1]:

- **40% of Internet** use is not related to business;
- **64% of employees** use the Internet for personal interest at work;
- average employee spends **1-2 hours per day** for unproductive browsing;
- wasting 1 hour per day employee yields approximately **$ 7500 of losses per year**.

TopTenReviews by TechMediaNetworks[2]:

- **12% of total web sites** contain pornography;
- **32% of users** complain on receiving unwanted pornographic exposure.

Solution: dynamic content filtering software (automatic classification)

Content Filtering Problem

- Methods
  - Automatic classification using machine learning
- Nature of categories varies
  - Thematic: pornography, purchase of tobacco, web proxies etc.
  - Functional: file sharing sites, social networks, chats and forums etc.
- Differs from text classification
  - Processing time is crucial
  - Desired ratio of false positives and false negatives may vary (strictness)
  - Target data constantly changes
- Consequences
  - Cannot use complex feature selection techniques
  - Need to simplify classifier retraining procedure

Work Parts

1. Upgrade previous work[1] by taking into account URLs found on a page
2. Evaluate classifiers in near-real conditions

1. Roman Suvorov, Ilya Sochenkov, Ilya Tikhomirov. Method for Pornography Filtering in the WEB Based on Automatic Classification and Natural Language Processing // in Proceedings of 15th International Conference, SPECOM 2013. Ed. Miloš Železný, Ivan Habernal, Andrey Ronzhin. Pilsen, Czech Republic, 2013, pp 233-240. ISBN 978-3-319-01930-7

Work Parts

1. Upgrade previous work[1] by taking into account URLs found on a page
2. Evaluate classifiers in near-real conditions

1. Roman Suvorov, Ilya Sochenkov, Ilya Tikhomirov. Method for Pornography Filtering in the WEB Based on Automatic Classification and Natural Language Processing // in Proceedings of 15th International Conference, SPECOM 2013. Ed. Miloš Železný, Ivan Habernal, Andrey Ronzhin. Pilsen, Czech Republic, 2013, pp 233-240. ISBN 978-3-319-01930-7

Our Previous Work: nTIC

- Thematic Importance Characteristic (nTIC)
- Calculate a measure similar to information gain, normalize it and compare with a threshold
- Use stems of words as features

$$I(d,c) = \sum_{t \in L(d)} lTF(t,d) IDF(t,c)$$

$$I(d_{bad}, c_{bad}) << I(d_{bad}, c_{good}) \quad I(d_{good}, c_{bad}) >> I(d_{good}, c_{good})$$

$$nTIC(d, c_{bad}, c_{good}) = \frac{I(d, c_{good}) - I(d, c_{bad})}{I(d, c_{good})}$$

$$nTIC(d, c_{bad}, c_{good}) > Threshold(c_{bad}, c_{good})$$

Proposed Modifications (1)

- Take into account interlinked nature of the Web
  - Use categories of the neighbor pages
  - Utilize peculiarities of URLs found on a page

Proposed Modifications (1)

- Take into account interlinked nature of the Web
  - Use categories of the neighbor pages
  - Utilize peculiarities of URLs found on a page

Proposed Modifications (2)

- Use categories of the neighbor pages (thematic isolation)
  1. Extract URLs from the body of a page
  2. Extract server domain names from these URLs
  3. Map domains to category labels using a dictionary
  4. Calculate weights of these labels as if they were usual lexis
  5. Treat these weights as features

- Initial dictionary is built from all open sources we have found: DMOZ, various black lists etc.
- The dictionary then expanded with resources that were classified with high confidence

Proposed Modifications (3)

- Take into account interlinked nature of the Web
  - Use categories of the neighbor pages
  - Utilize peculiarities of URLs found on a page

Proposed Modifications (4)

- Utilize peculiarities of URLs found on a page
  - Human-friendly URLs are getting more popular

- Algorithm
  1. Extract URLs from the body of a page
  2. Split each URL in a set of tokens delimited by special characters (/, ?, &, # etc)
  3. Normalize each token
  4. Calculate weights of these tokens as if they were usual lexis
  5. Treat these weights as features

Work Parts

1. Upgrade previous work[1] by taking into account URLs found on a page
2. Evaluate classifiers in near-real conditions

1. Roman Suvorov, Ilya Sochenkov, Ilya Tikhomirov. Method for Pornography Filtering in the WEB Based on Automatic Classification and Natural Language Processing // in Proceedings of 15th International Conference, SPECOM 2013. Ed. Miloš Železný, Ivan Habernal, Andrey Ronzhin. Pilsen, Czech Republic, 2013, pp 233-240. ISBN 978-3-319-01930-7

Data Collection Problem

- No etalon datasets for content filtering
- Standard datasets for text categorization misfit content filtering:
  - Categories are less thematic and more associative
  - They lack linking information (no chance to evaluate URL-based features)
- Publicly available access lists are outdated

- Possible solutions
  - Use unsupervised machine learning
  - Use methods that require less data to learn
  - Introduce a technique for datasets collecting that require small manual labor

Data Collection Problem

- No etalon datasets for content filtering
- Standard datasets for text categorization misfit content filtering:
  - Categories are less thematic and more associative
  - They lack linking information (no chance to evaluate URL-based features)
- Publicly available access lists are outdated

- Possible solutions
  - Use unsupervised machine learning
  - Use methods that require less data to learn
  - Introduce a technique for datasets collecting that require small manual labor

Thematic Crawler (1)

- Collect only pages on the topic of interest
- Subtask of focused crawling

- Create a simple system that needs almost no tuning or training

- Similar to Babouk
  - differs in walk order and page selection rule

Thematic Crawler (2)

- Main algorithm
  1. Collect seed URLs using metasearch with queries related to the subject of interest
  2. Recursively crawl pages starting from the seed URLs
     - Breadth-first order
     - If a seed URL is not a root page (e.g. http://www.isa.ru/index.php?...), download it and proceed with next seed URL
     - If a seed URL is a root page (e.g. http://www.isa.ru/), proceed recursively with **topic filtering** enabled
- Topic filtering principles
  - Maintain a list of keywords (stemming, TF-IDF) of the current walk graph (global keywords)
  - Compare current page keywords with global keywords (size of intersection)
  - Merge current page keywords into the global list and reduce it if necessary
- Parameters
  - Sizes of keywords lists to maintain
  - How often and how aggressively to reduce the global keyword list

Evaluation (1): Data and Problem

- Languages: English, Russian
- 170000 pages total from 17 “bad” categories
  - 10000 pages per category
  - 5000 pages per language within category
- 20000 “good” pages from Wikipedia and informational sites

- Multi-class and multi-label classification problem
  - If no labels are assigned to a document, it is “good” and “bad” otherwise
- Reduce to a set of binary problems using “one-vs-all” technique

- Use classic measures: accuracy, precision, recall and F1

Evaluation (2): Classifiers and Features

- Test two classifiers (nTIC and linear SVM) with two sets of featuers
  - Base - only plain lexis
  - Cat&Tok – Base + modifications to nTIC for URLs

Classifier	Feature set	Precision			Recall			F1		
		Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
nTIC	Base	0.739	0.972	0.895	0.918	0.994	**0.968**	0.819	0.983	0.929
nTIC	Cat&Tok	0.812	0.986	**0.934**	0.909	0.988	0.963	0.878	0.986	**0.948**
SVM	Base	0.98	0.999	**0.996**	0.962	0.996	**0.988**	0.971	0.997	**0.992**
SVM	Cat&Tok	0.98	0.999	0.996	0.953	0.996	0.985	0.973	0.997	0.991

- Results
  - +7% to precision
  - +2% to F1
  - SVM is not improved (most probably due to optimization)

Evaluation (3): Small Feature Sets

- Content filtering software must run fast on restricted hardware as well
  - Choose smaller feature sets (original contained about 600 000 features)
  - Compare three feature selection techniques: IDF, nTIC, Information Gain (IG)

Technique	IDF			nTIC			IG		
N	P	R	F1	P	R	F1	P	R	F1
5 000	0.977	**0.948**	**0.962**	**1**	0.852	0.921	0.99	0.83	0.908
10 000	0.983	**0.962**	**0.972**	0.981	0.955	0.968	**0.99**	0.908	0.951
100 000	0.992	**0.975**	**0.984**	0.995	0.951	0.972	**0.997**	0.958	0.977

- Results
  - IDF is a good choice for most applications
  - nTIC is better for accurate content filtering in very scarce environments (on smartphones, old servers etc)

Conclusion

- Described and compared in near-real conditions
  - two classifiers (nTIC, SVM)
  - two feature extraction techniques (lexis, lexis + URL-based features)
  - three feature selection techniques (IDF, nTIC, IG)
- Thematic crawler is proposed and used for establishing near-real conditions of experiments
- Taking into account URL-based features
  - does not significantly increase quality in general case
  - may be the only features when dealing with e.g. Google Images (almost no text)
- Feature selection
  - nTIC gains better quality on small feature sets
  - Feature selection is not that important in middle- and large-scale applications

Future Work

- Develop a comprehensive classification system that
  - analyses the structure of a page, its functions and graphics
  - takes into account user's behavior.

Acknowledgements

- Russian Foundation for Basic Research grant 12-07-33012
- Exactus project (semantic search and NLP)
  - http://exactus.ru
  - http://expert.exactus.ru
- “TSA Filtratus” dynamic content filtering system

Institute for Systems Analysis Russian Academy of Sciences

117312, Moscow, pr. 60-let Octyabrya, 9

Konstantin Yakovlev, PhD

yakovlev@isa.ru