Social Network Analysis

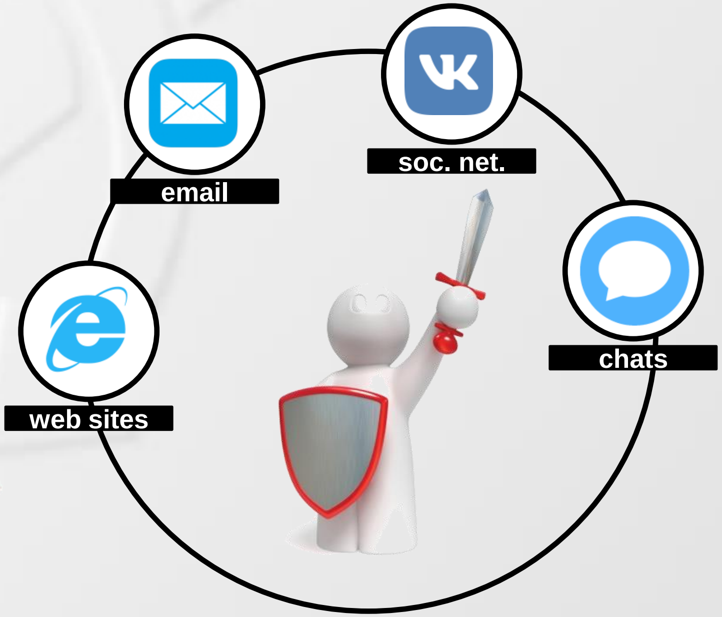
Andrey Chechulin

Laboratory of Computer Security Problems
St. Petersburg Institute for Informatics and Automation
Saint-Petersburg, Russia

Table of Content

- Introduction
- Social network as an information source
- General architecture
- Graph analysis by the visualization means
- Text analysis by the machine learning means
- Images analysis by the machine learning means
- Conclusions

Introduction (1/2)
Problem statement



Challenge:

- growing distribution of **inappropriate information in Internet**
- **terroristic communities** in social networks
- **the absence of control** of information space

Goal:

- to develop a system for information space analysis for **detection and counteraction** against inappropriate information

Proposed approach

- **automatic gathering and analysis** of information objects in information space
- social network **communities analysis**
- **visual analysis** of social networks

Introduction (2/2)
Inappropriate information

Federal law of Russian Federation no. 139-FZ of 2012-07-28 describes the need to block web sites that contain:

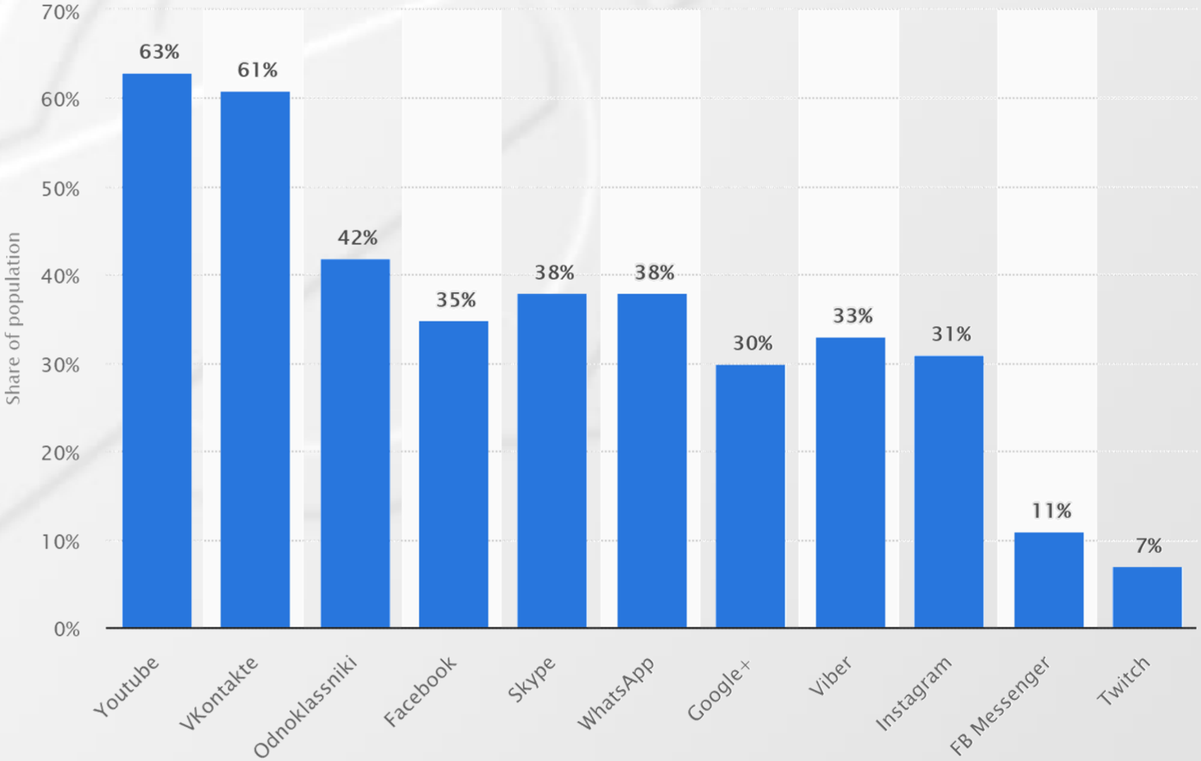
- child pornography or solicitation to participate in such
- information about methods of making, using, getting or locating narcotic drugs and psychotropic substances or their precursors (acetone, potassium permanganate, sulfuric acid, hydrochloric acid, acetic acid) or growing plants containing narcotic drugs
- information about methods of suicide and calls for suicide
- **any Internet-distributed information which has a court decision describing it as a prohibited to be spread in Russia**

Parental control systems should be able to block web sites that contain:

- pornographic and erotic materials
- information associated with the propaganda of sectarianism
- information related to the racial, religious, etc. discrimination
- etc

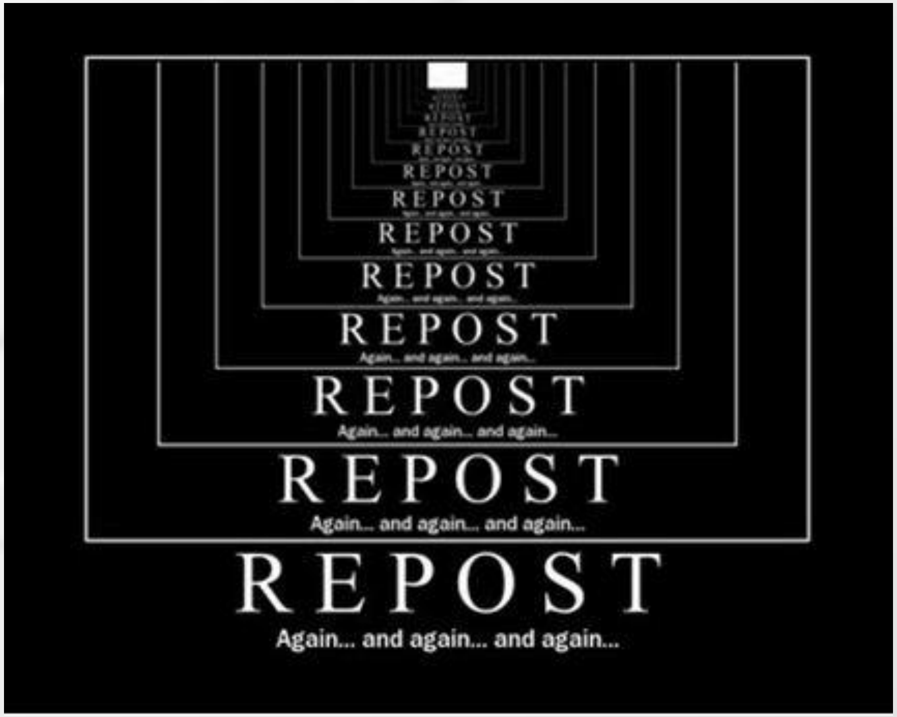
Information gathering (1/4)
Source for information gathering

- **Statistics (2019)**
  - 47 percent of the population in Russia have an active account with any social network
  - VKontakte has over 46.6 million monthly users in Russia and abroad
- **Auditory**
  - The most visited in Russia
  - 2nd most visited in Belarus
  - 3rd most visited in Kazakhstan
  - 4th most visited in Estonia, Kyrgyzstan and Moldova
  - 5th most visited in Latvia



Information gathering (2/4)
VK information dissemination channels

- **Basic objects**
  - **User**
  - **Community**
  - **Wall Post**
  - **Wall Comment**
  - Private Message
  - Chat
  - Note
  - Wiki Page
  - Market Item
  - Market Collection
  - Topic
  - Topic Comment
  - Application
  - Poll



- **Repost**
  - the **direct copy** of an information object from parent object to a new one
  - one of the **most effective** ways of **information dissemination** channels
  - the sources and receivers of information objects are **users and groups**

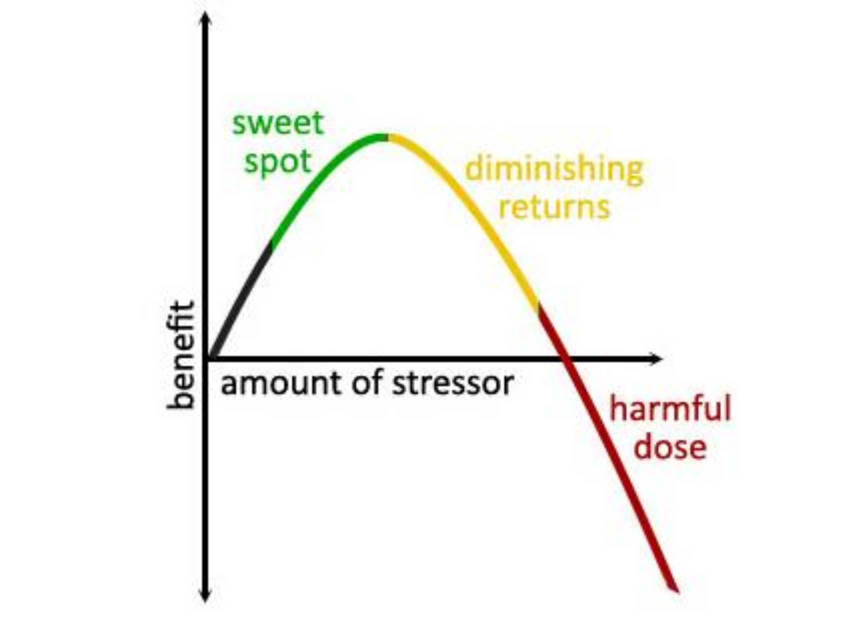
Information gathering (3/4) Data gathering algorithms

- **“Repost” data gathering**
  1. Detection of all information objects, that are in **"repost" relationship** with other objects for the specified time period
  2. Gathering information about all the **sources of information objects** (looking for information objects in "repost" chain "down")
  3. Iterative information gathering about all **receivers of information objects**: (looking for information objects in "repost" chain "up")

Information gathering (4/4)
Data gathering algorithms

- **“Attenuation“ or "distortion" data gathering**
  1. A set of keywords is calculated for text data from all **initial information objects** for the specified time period
  2. For each receiver a set of keywords is calculated for all text data, found in **information objects from information space** of the receiver for the specified time period
  3. For each receiver a **number of keywords** that exists in both sets is calculated.

  Calculated value indicates the **degree of similarity** between information spaces of sources and receivers.



General system architecture

Monitoring

Tracking

Links

Content

Detection of dangerous influence

Attack sources

Target audience

Information distribution channels

Disseminated information

Counteraction

Developing a list of countermeasures

Target of the countermeasure

Type of countermeasure

Decision support

Modeling

Resources evaluation

General system architecture
Graph analysis by the visualization means

Monitoring

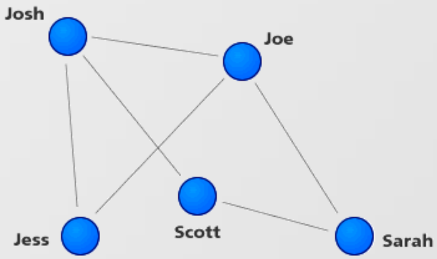
- Tracking
- Detection of dangerous influence
- Links
- Attack sources
- Target audience
- Information distribution channels
- Content
- Disseminated information

Counteraction

- Developing a list of countermeasures
- Decision support
- Target of the countermeasure
- Modeling
- Type of countermeasure
- Resources evaluation

Graph analysis by the visualization means
Graph representation

- **Objects involved in information interaction**
  - **the information source** – the entity that is a starting point for information content and has high level of unique content
  - **the information repeater** – the entity that is repeated by others involved in information interaction with low level of unique content
  - **the information aggregator** – the entity with low level of unique content but big audience
  - **the information consumer** – the entity with low level of unique content and small audience with no further distribution of the content



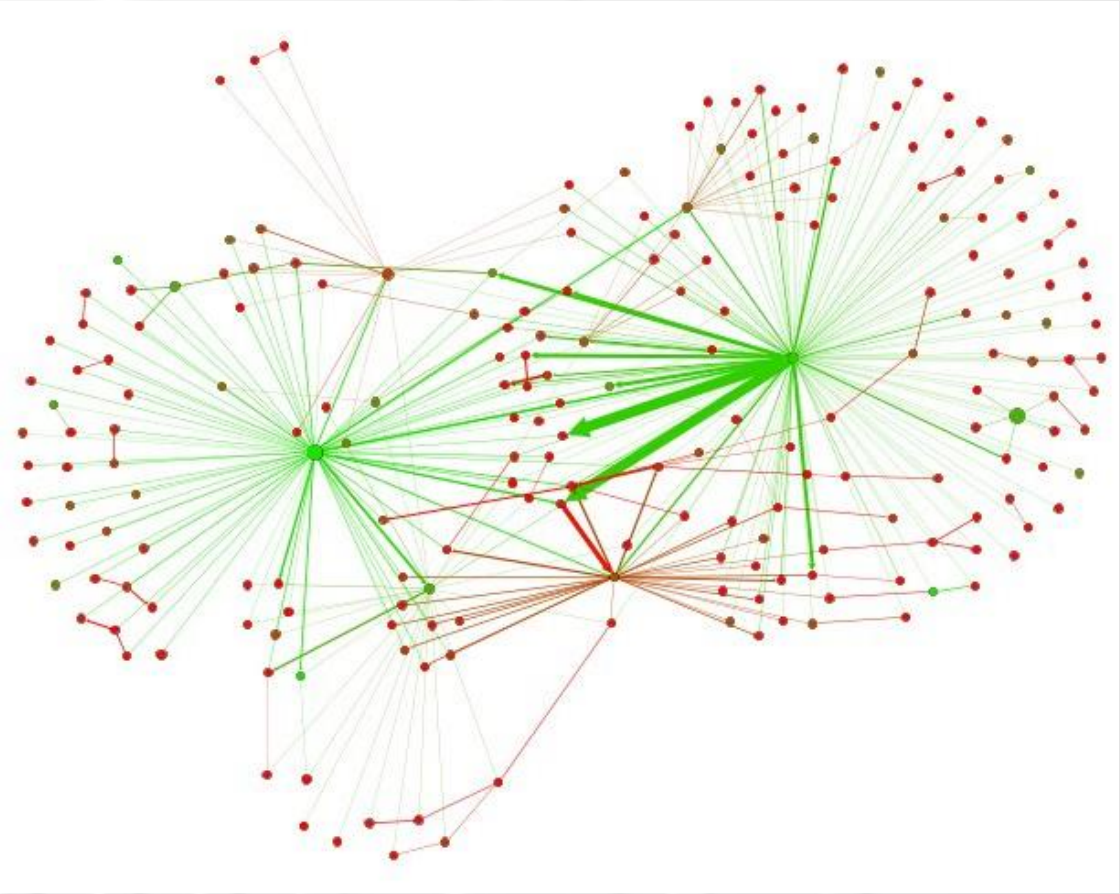
Characteristic	Index
Vertex (size)	Average number of views
Vertex (form)	Type of the social network object
Vertex (color)	Uniqueness of the generated content

Characteristic	Index
Edge (thickness)	The flow saturation
Edge (direction)	Direction of the information flow
Edge (color)	Uniqueness of the generated content

Graph analysis by the visualization means
Source data and general graph

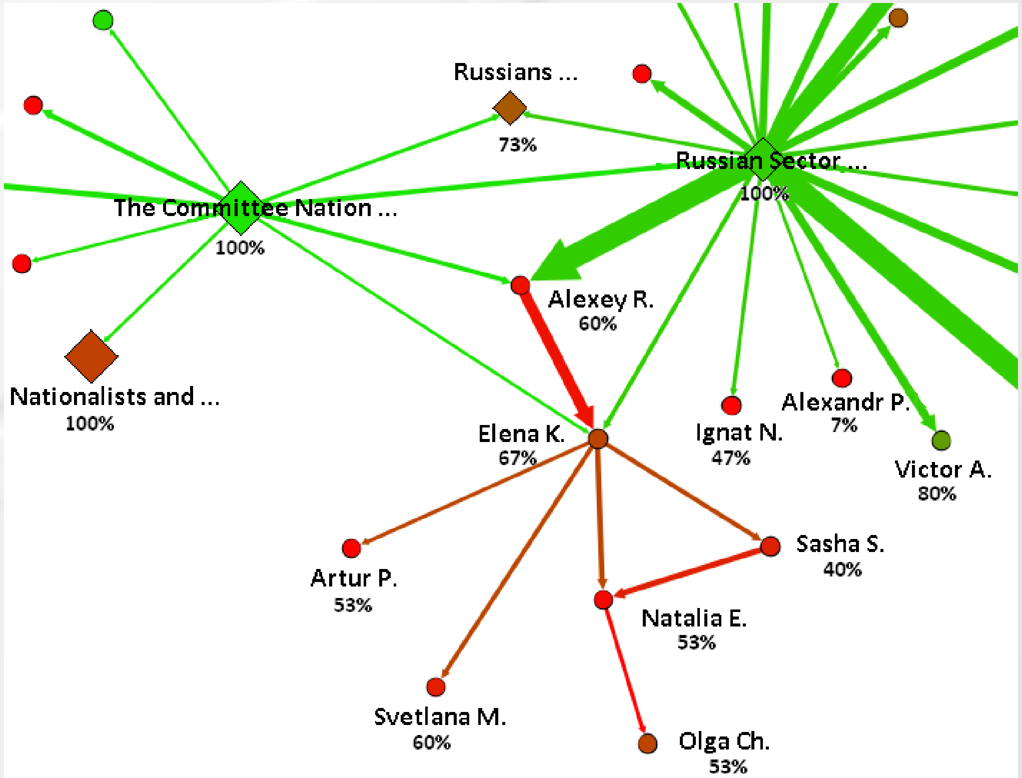
Input data

- the information objects from VK group "Nationalists and …" with 10 additional groups were selected
- a time period was selected as 7 days
- the set of keywords was formed based on the information space of the input data (475 058 signs)
- all data was collected "as is" by the proposed algorithm, with no impact on groups, users or information flows



Graph analysis by the visualization means Visual analysis of the sub graph (1/2)

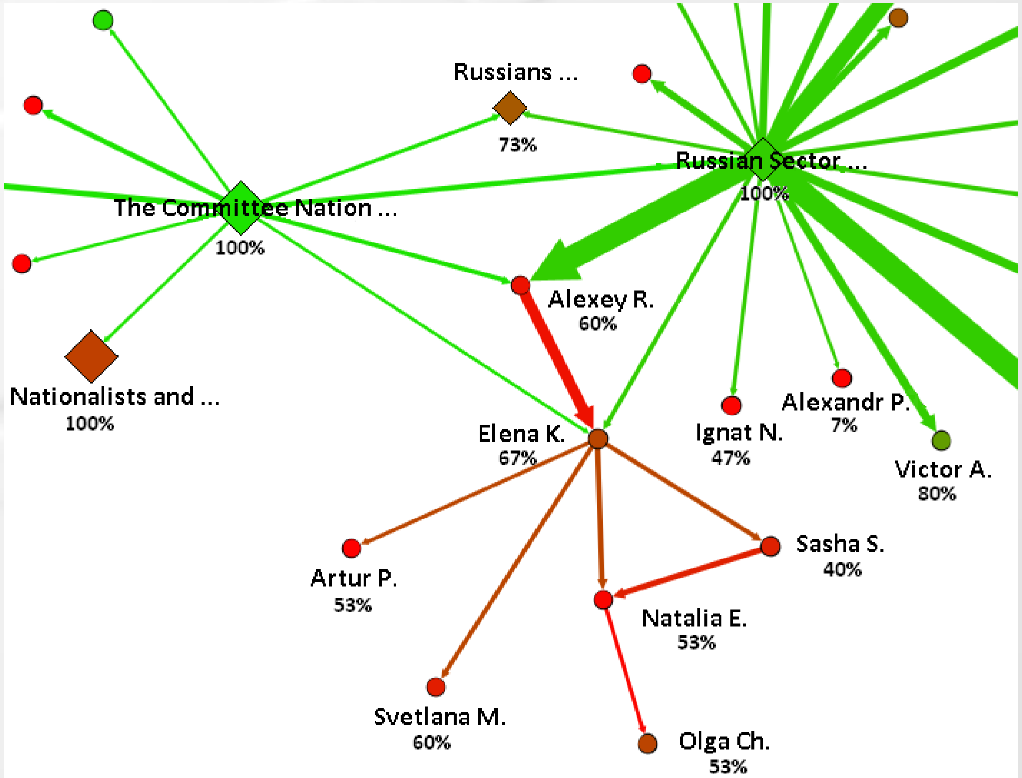
- **Visual analysis**
  - The groups "Nationalists and …", "The Committee Nation …", "Russian Sector …", "Russian …" and the user "Elena K." have the biggest coverage of the audience
  - The content of the groups "The Committee Nation …" and "Russian Sector …" is the most unique and the group "Russians…" disseminate both the unique and reposed content



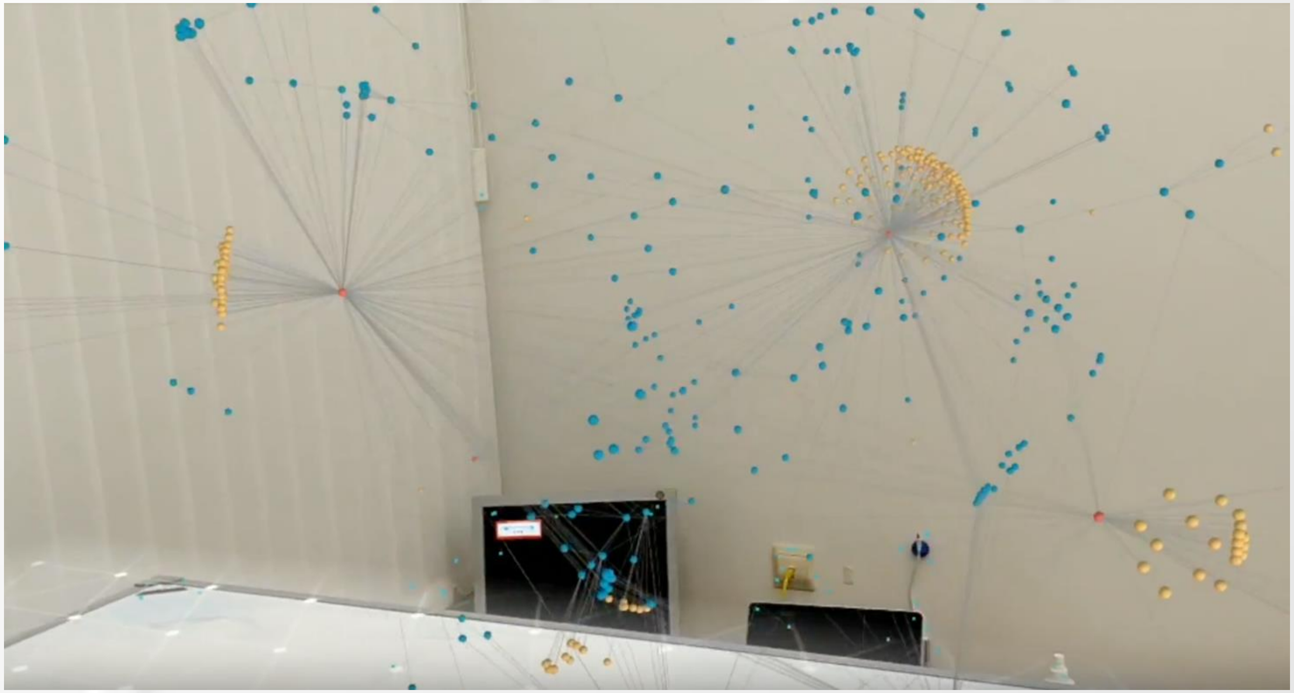
  - The groups "The Committee Nation …" and "Russian Sector …" are the main sources of information (in this case, the first one influences on the latter)
  - The group "Nationalists and …" is the aggregator with a greatest readers amount and it has a great information impact on the audience

Graph analysis by the visualization means Visual analysis of the sub graph (2/2)

- **Visual analysis**
  - The user "Elena K." is an information repeater which takes the information from the groups "The Committee Nation …", "Russian Sector …" and the user "Alexey R." and transmits it to users "Artur P." and others. Despite of a large number of reposts this user also generates unique content
  - The information "attenuation" is also can be seen here. E.g. the keywords for the group "Russian Sector …" are "elections, child, Kemerovo, Putin, Kremlin, Moscow, Russia, nation, Ukraine, Vic, media", meanwhile the keywords for user "Olga Ch." are "elections, child, Kemerovo, Putin, Kremlin, Moscow, education, essence, power, region, administration". So the context of reposted information changes in the user’s information space and the original one "attenuates".



Graph analysis by the visualization means
Augmented reality



View:

- 3 motor and 3 rotational degrees of freedom
- The possibility of sharing in augmented reality mode

Control:

- Decision support
- Ability to receive additional information
- Combining models into a common interactive object

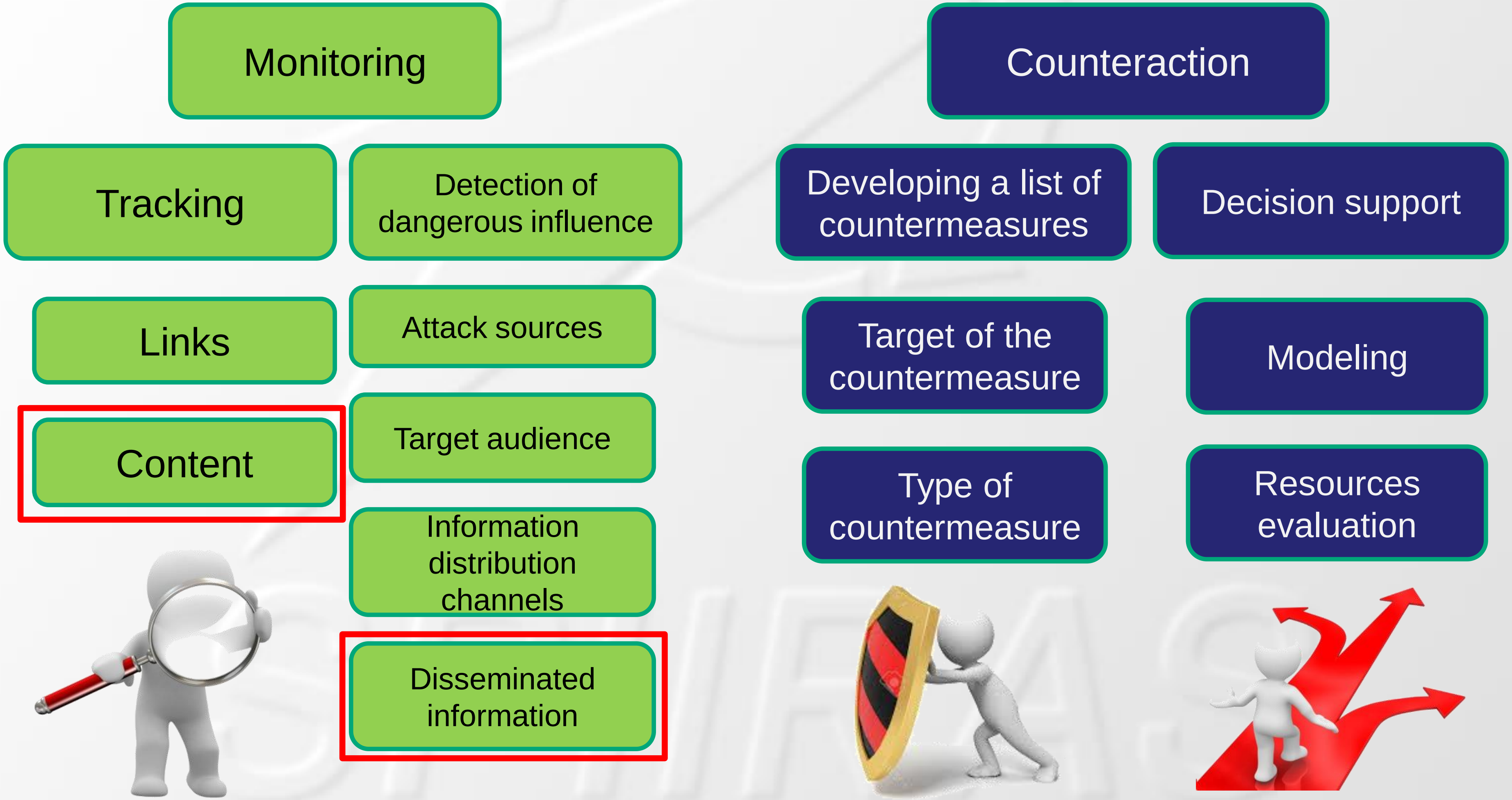
Advantages:

- Modeling with natural movements
- The use of human cognitive characteristics to facilitate the perception of information

Graph analysis by the visualization means Summarizing, future works and examples

- **Summarizing**
  - Social graphs are very complex
  - In some cases visualization can help to solve these challenges
  - Graph analysis can show some information even without content analysis
- **Future works**
  - The development of new visual model (including ones for VR and AR)
  - The development of new techniques for graphs storage, filtration, segmentation, transformation, verification, analysis, processing and visualization
  - The development of information gathering modules for Facebook, Twitter, etc
- **Links**
  - Implementation example: 2D graph: http://comsec.spb.ru/files/forceWithout.html
  - Implementation example: 2D graph: http://comsec.spb.ru/files/forceWith.html
  - Implementation example: 3D graph: http://comsec.spb.ru/files/force3D.html
  - Implementation example: 3D graph in augmented reality (one needs a mobile phone with Google AR Core support) : http://comsec.spb.ru/files/forceAPK.html

General system architecture
Content analysis



Monitoring

Tracking

Detection of dangerous influence

Links

Attack sources

Content

Target audience

Information distribution channels

Disseminated information

Counteraction

Developing a list of countermeasures

Decision support

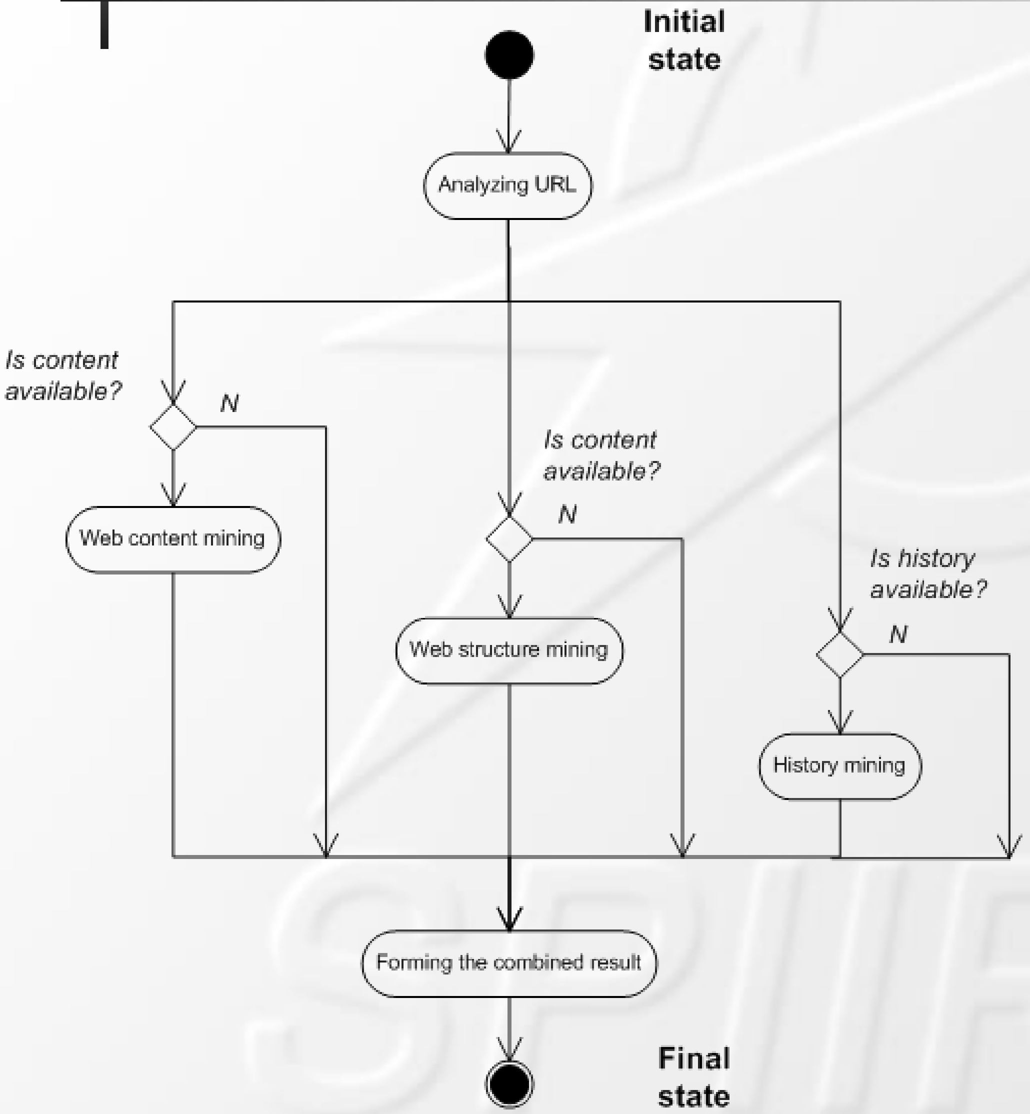
Target of the countermeasure

Modeling

Type of countermeasure

Resources evaluation

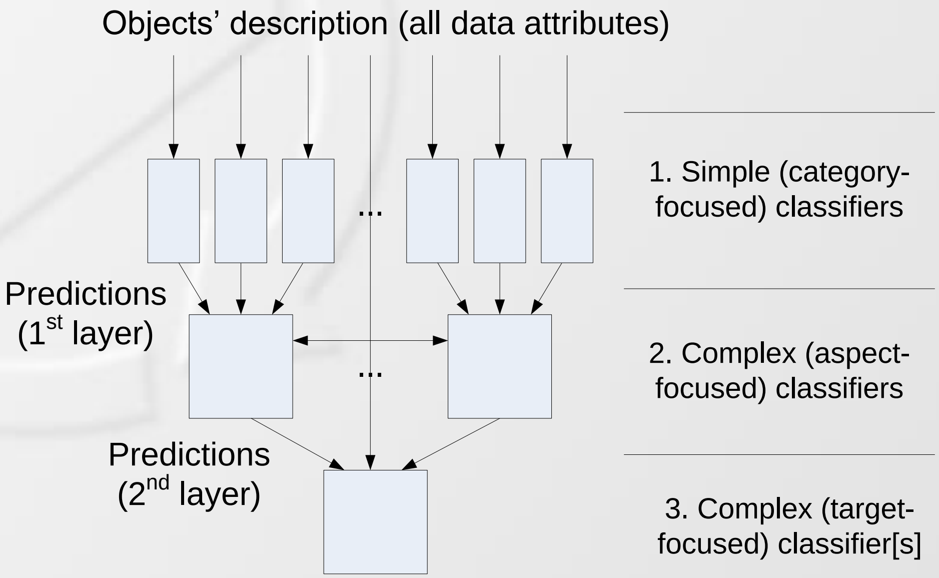
Classification architecture (1/5)
Common scheme of using existing methods



- *Analyzing URL* provides the classification of the object without the analysis of its content, links with neighbor objects and history.
- *Analyzing content*, the object is considered as a set of data which exists independently, apart from environment and its evolution in that environment.
- *Structure analysis* takes object’s environment into account and enables making a decision about the object considering its place in the net of objects linked to each other.
- At *history analysis*, a dynamical component is introduced to object’s description showing main trends of its evolution.

Classification architecture (2/5) General approach to web classification

- Three level decision making
  - **1st level classifiers** focused on making a decision on whether a given feature vector belongs to a particular category
  - ***2nd level classifiers*** recognize a category basis on general data and on first level predictions
  - ***3rd level classifier*** *makes* a final decision relying on both second level predictions and raw data

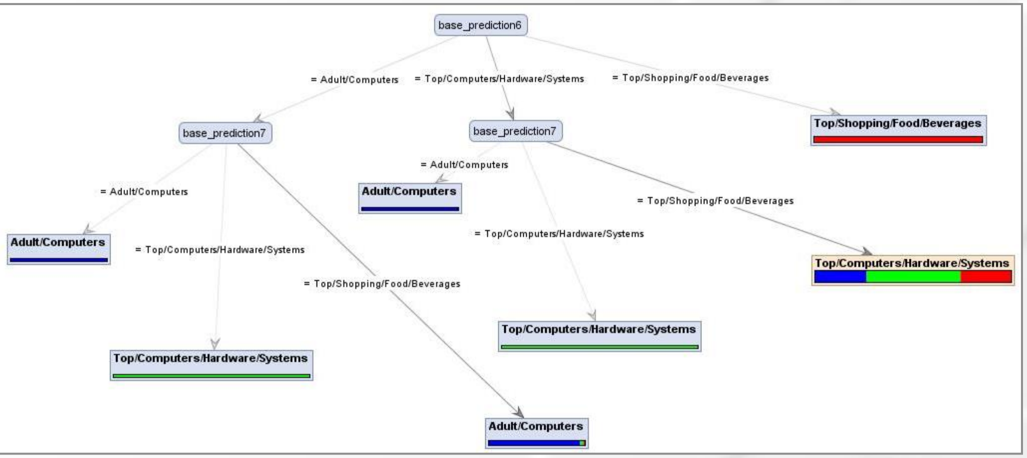


Objects' description (all data attributes)
...
Predictions (1st layer)
...
Predictions (2nd layer)
1. Simple (category-focused) classifiers
2. Complex (aspect-focused) classifiers
3. Complex (target-focused) classifier[s]

- Three basic options of features usage on the 3rd level
  - ***Main Units*** – the classifier uses only on second level predictions
  - ***Mixed Units*** – first and second level predictions are used
  - ***Mixed Units with extended feature set*** – together with data of first two levels the most significant features are used

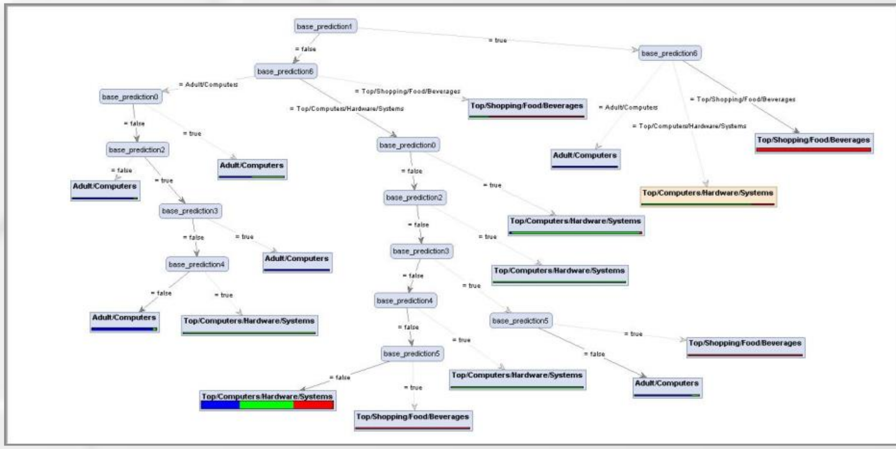
Classification architecture (3/5)
Selection of classification scheme

Main Units



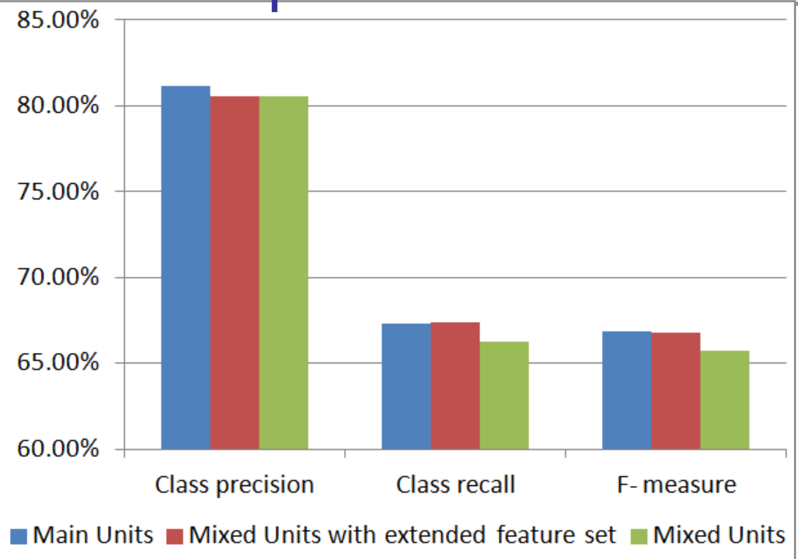
base_prediction6
= Adult/Computers
= Top/Computers/Hardware/Systems
= Top/Shopping/Food/Beverages
base_prediction7
base_prediction7
Top/Shopping/Food/Beverages
= Adult/Computers
Adult/Computers
= Adult/Computers
= Top/Shopping/Food/Beverages
Adult/Computers
= Top/Computers/Hardware/Systems
= Top/Computers/Hardware/Systems
Top/Computers/Hardware/Systems
= Top/Shopping/Food/Beverages
Top/Computers/Hardware/Systems
Top/Computers/Hardware/Systems
Adult/Computers

Mixed Units

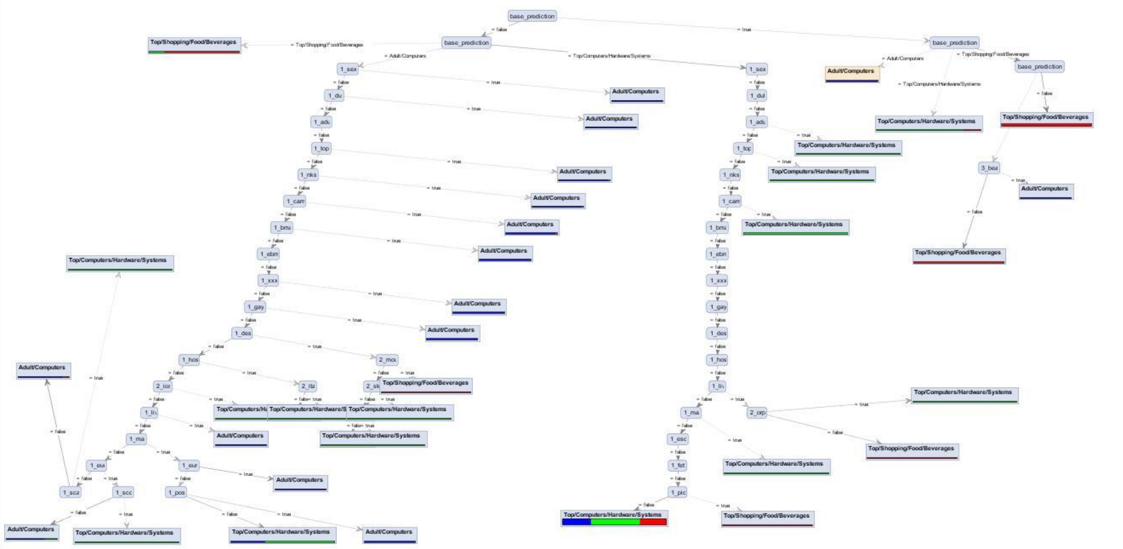


base_prediction1
= false
= true
base_prediction6
base_prediction6
base_prediction0
= Adult/Computers
= Top/Shopping/Food/Beverages
= Top/Computers/Hardware/Systems
Top/Shopping/Food/Beverages
= Adult/Computers
= Top/Shopping/Food/Beverages
= Top/Computers/Hardware/Systems
Top/Shopping/Food/Beverages
base_prediction2
base_prediction0
Adult/Computers
Adult/Computers
Adult/Computers
Top/Computers/Hardware/Systems
base_prediction2
base_prediction3
Top/Computers/Hardware/Systems
base_prediction4
Adult/Computers
base_prediction3
Top/Computers/Hardware/Systems
Adult/Computers
Top/Computers/Hardware/Systems
base_prediction4
base_prediction5
Top/Shopping/Food/Beverages
base_prediction5
Top/Computers/Hardware/Systems
Adult/Computers
Top/Computers/Hardware/Systems
Top/Shopping/Food/Beverages

Comparison

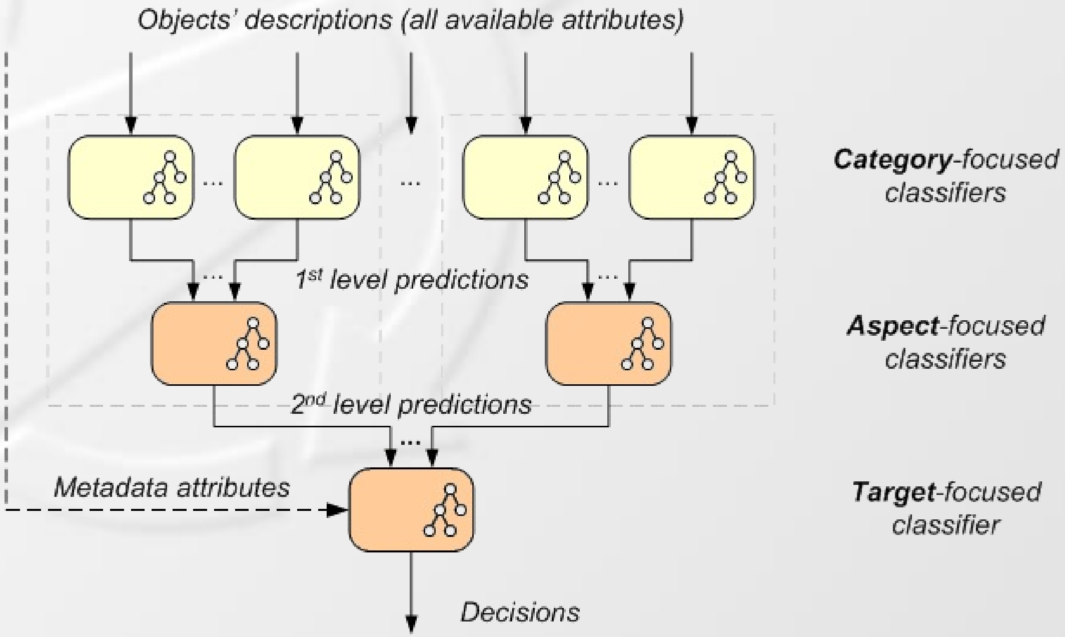


Mixed Units with extended feature set



base_prediction
Top/Shopping/Food/Beverages
base_prediction
base_prediction
base_prediction
Adult/Computers
Top/Computers/Hardware/Systems
Top/Shopping/Food/Beverages
Top/Computers/Hardware/Systems
Adult/Computers
Top/Shopping/Food/Beverages
Top/Computers/Hardware/Systems

Classification architecture (4/5) General classification scheme



Objects' descriptions (all available attributes)
...
...
...
1st level predictions
...
...
2nd level predictions
...
Metadata attributes
Decisions
Category-focused classifiers
Aspect-focused classifiers
Target-focused classifier

- Classification features
  - multilevel classifier
  - category-focused attributes are used only on 1st level
  - 1st level classifiers are binomial
  - 2nd and 3rd level classifiers use only result of previous levels
  - there is no separate classifier for “unknown” category
  - aspect-focused classifiers can be hierarchical
  - metadata attributes (i.e. general attributes that cannot be used for separate classifier , e.g. text volume, country of IP, server’s features, etc.) can be used on the 3rd level for classification improvement

Classification architecture (5/5)
Classification quality metrics

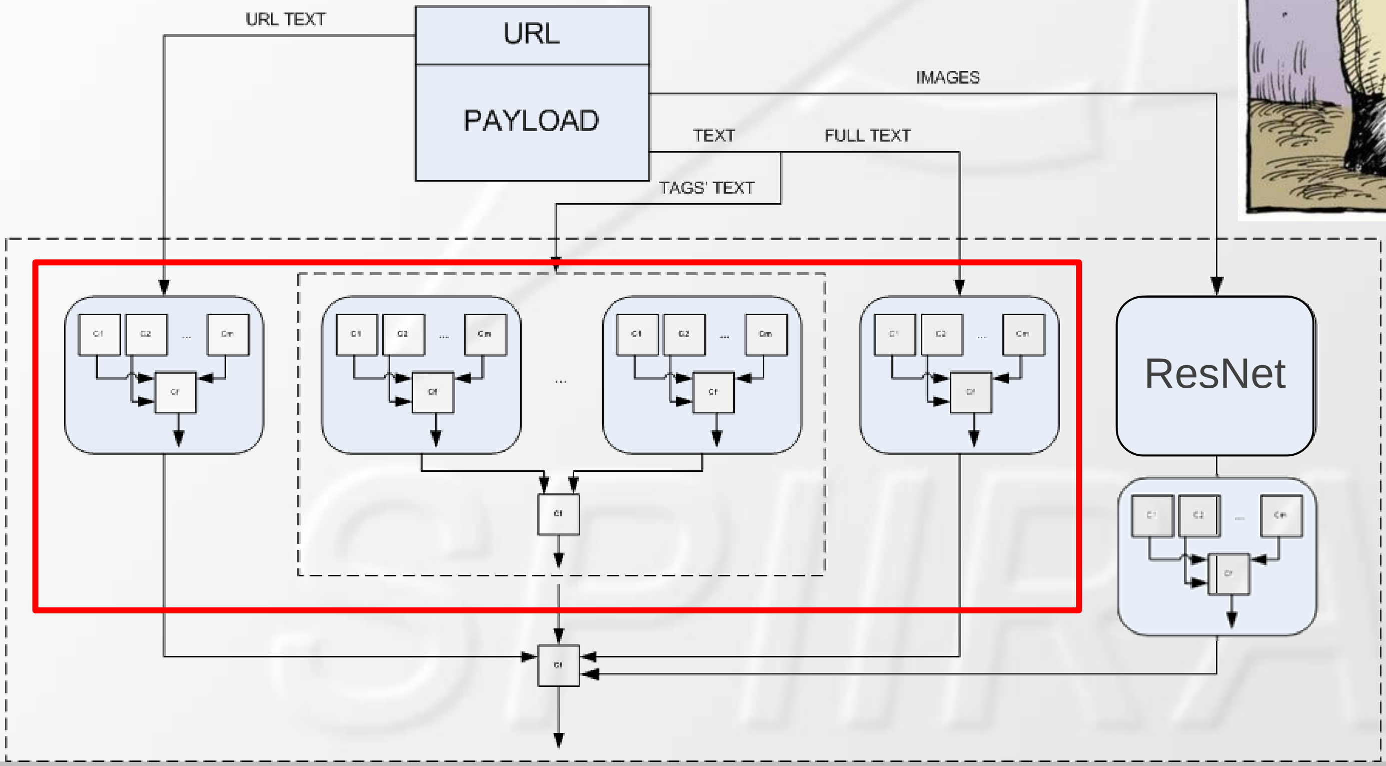
Belonging to category		Expert	
		TRUE	FALSE
Classifier	TRUE	*TP (true positive)*	*FP (false positive)*
	FALSE	*FN (false negative)*	*TN (true negative)*

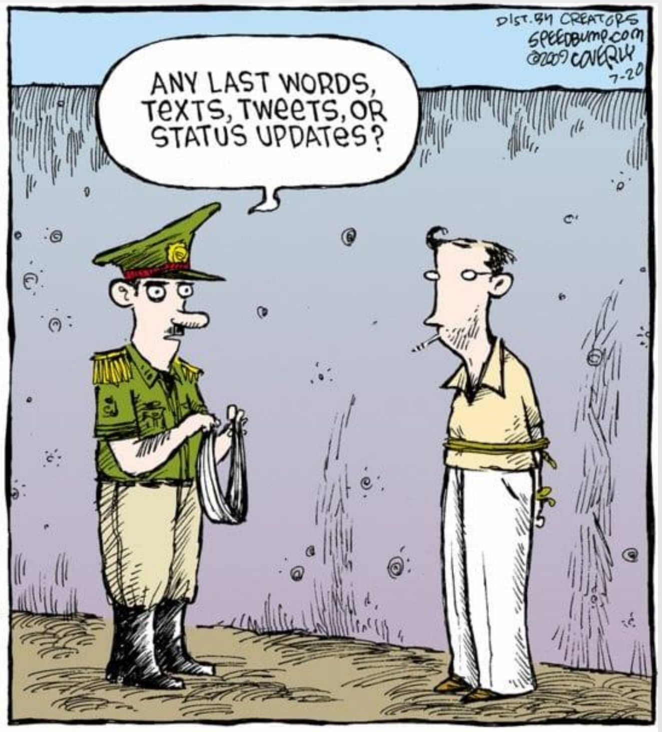
- ***Recall*** is calculated as the ratio of quantity of correctly classified Web-pages to the total of the Web-pages concerning to the chosen category: r=TP/(TP+FN)
- ***Precision*** is calculated as the ratio of correctly classified Web-pages to the total of the Web-pages classified on the chosen category: *p=TP/(TP+FP)*
- ***Accuracy*** is calculated as the ratio of decisions, correctly determined by the system, to the total number of decisions: *p=(TP+TN)/(TP+FP+FN+TN)*
- ***F-measure*** is a harmonic mean of precision and recall:

$$\textit{F-measure} = \frac{2pr}{p+r}$$

Text analysis
General scheme

- Data sources
  - **Textual content**
  - Images
  - Links
- General page processing scheme:





Text analysis
Features for text classification

The dictionary was formed by using two approaches to TF (term frequency) calculation:

- **standard (sTF)** - the sequence of actions is as follows: calculate numbers of keyword's appearances per each site belonging to a category, sum them and divide the obtained value with the total amount of all words presenting in all categories' sites
- **Modified (mTF)** - the keyword's appearance is accounted only once

The similar way was used for approaching to IDF (inverse term frequency) :

- **Standard (sIDF)** - it is decided that a keyword belongs to a category if it appears even once in any site of the category)
- **Modified (mIDF)** - a keyword belongs to a category if its appearances' number exceeds some predefined threshold value

Text analysis Dictionaries

- Keywords dictionary
  - TF/IDF approach and its modification
  - Amount of keywords for each category
- Text processing
  - Stemming
  - Tokens
  - Hyponyms
  - Hyperonyms
- Input data attributes
  - Site's unique identifier
  - Whether the keyword presents in the text
  - The category the page belongs to

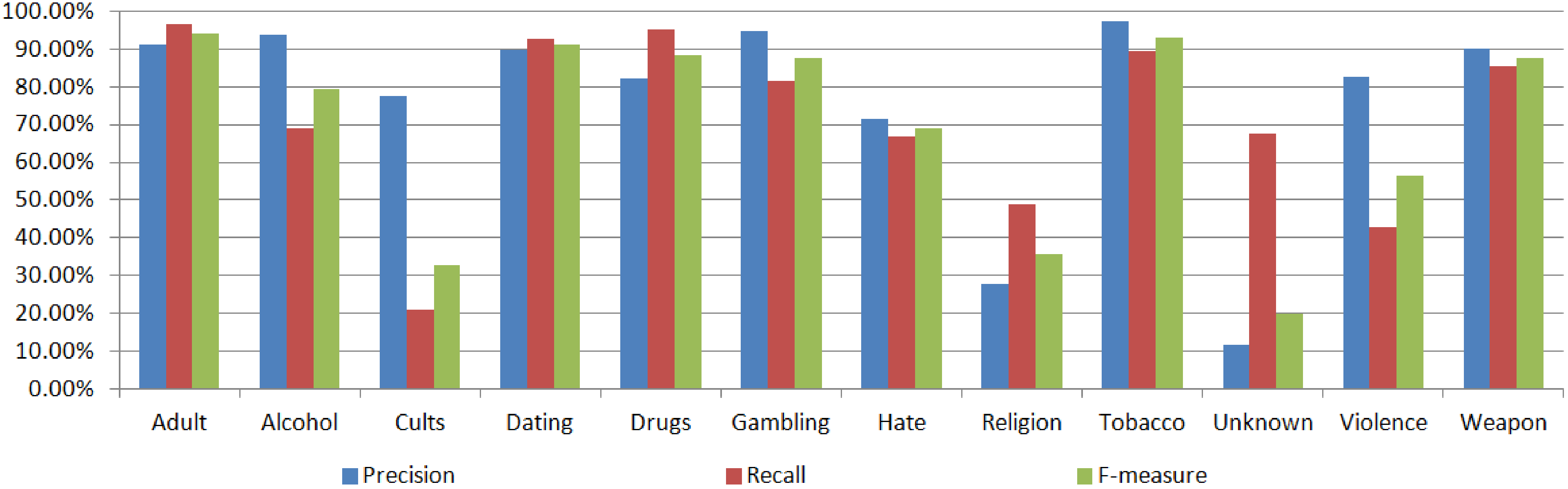
Category	Keywords
Adult	porn, sex, pic, xxx, hardcor
Alcohol	wine, tast, wineri, vineyard, beer
Banking	bank, loan, credit, union, financi
Blogs	septemb, juli, novemb, august, wordpress
Cults	church, bibl, christ, god, ministri
Dating	rencontr, singl, est, profil, vou
Drugs	whoi, eng, traffic, verifi, legitim
Forum	gmt, vbulletin, phpbb, guest, moder
Gambling	casino, poker, gambl, bet, bonu
Games	xbox, wii, psp, game, charact
Hate	hate, jew, jewish, truth, god
Health	clinic, treatment, patient, health, therapi
Job_Search	recruit, employ, resum, execut, candid
News	radio, opinion, classifi, newspap, digit
Sport	leagu, athlet, golf, season, basketbal
Tobacco	tobacco, smoke, cigarett, cigar, smoker
Travel	trip, cruis, charter, island, destin
Violence	violenc, abus, domest, victim, sexual
Weapons	gun, shoot, rifl, firearm, pistol

Text analysis
Source data preparation

N	Stage	Challenges	Decision
1	**Creation of the list of categories of pages**	Similar categories (“hate” and “violence”, “medicine” and “drugs”, “religion” and “cults”, etc)	Many iteration during categories selection
2	**Preparation of the input lists of pages**	Selection of “good” list; Combining of different lists sources	Groups, hashtags,…
3	**Loading the pages content to the internal storage**	Dynamic context	History collection
4	**Data pre-processing and extraction of features, which are used to train classifier models**	Feature selection (e.g. textual features cannot be used for all categories)	Collection of several different types of source data

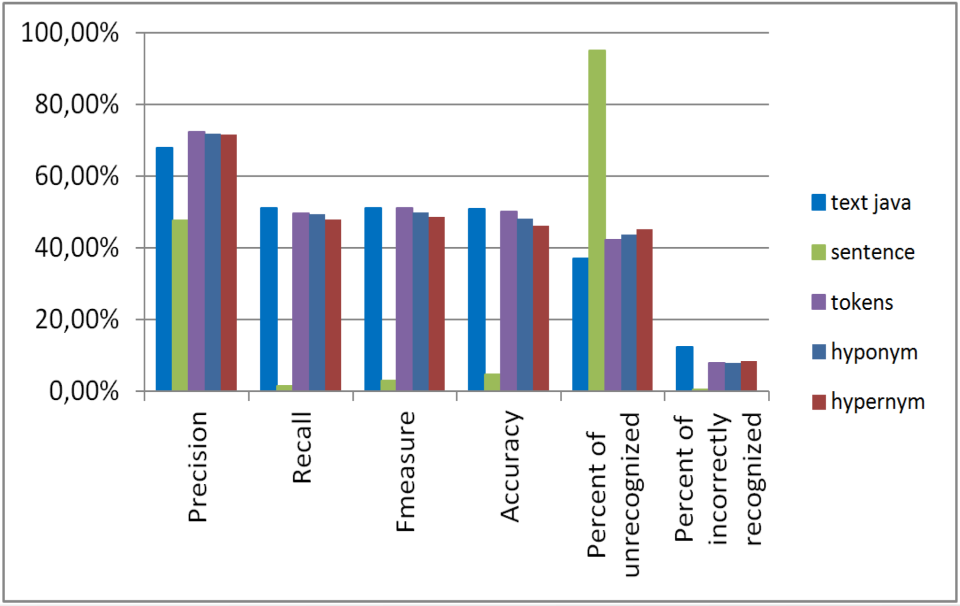
Text analysis
Experiment results (1/2)

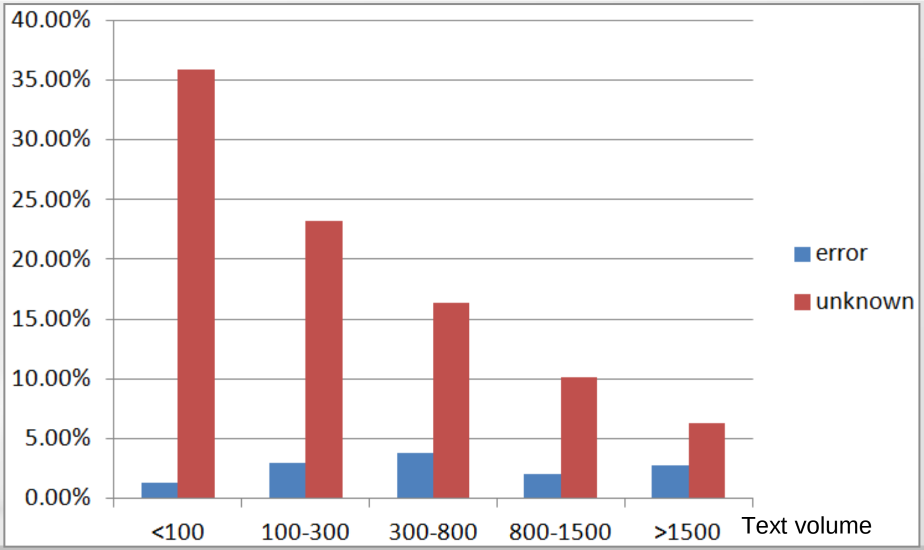
- Experiments results
  - The presence of similar categories can lead to a decrease of general classification quality (e.g., “Hate” and “Violence”, “Cults” and “Religion”, etc)
  - The use of combination of individual classifiers of different types in different categories leads to a significant increase in accuracy
  - The selection of Decision Trees as a basis of classification models leaded to prevalence of the precision



Text analysis
Experiment results (2/2)

- Experiments results
  - The best accuracy was demonstrated by the method of partitioning the text on the basis of tokens
  - Introduction of a new category “unknown” improves the accuracy of the combined classifier
  - Pages classification based on text analysis significantly depends on text volume
  - Text classifiers can be used as standalone classifiers





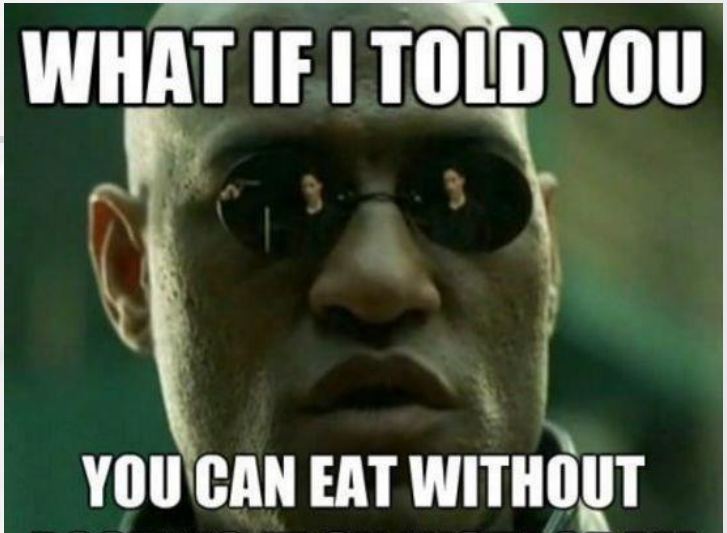
Text analysis
Summarizing, future works and examples

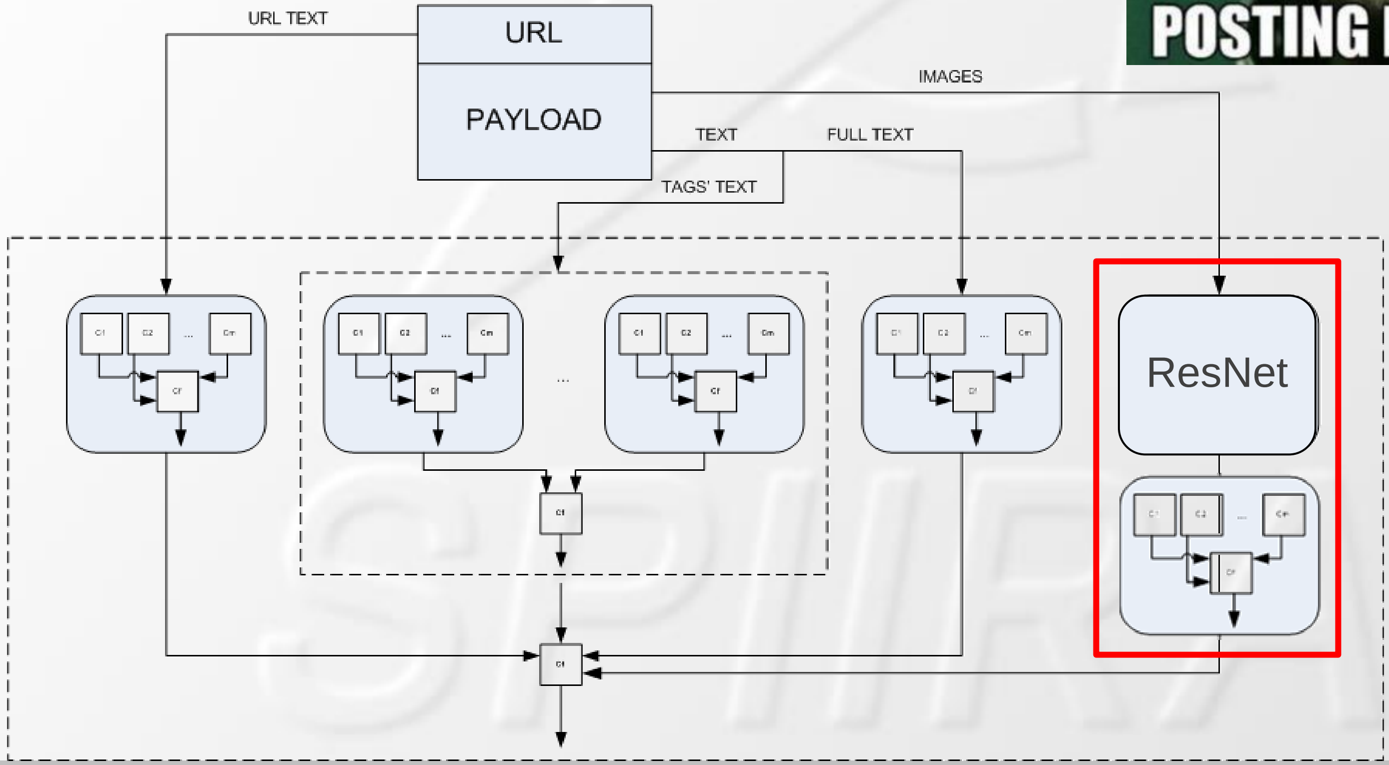
- **Summarizing**
  - Text analysis is the best way for pages analysis, but it is not the only one
  - Challenges here are both small and large text
  - Various language can also create some challenges
- **Future works**
  - The development of new techniques for text analysis
  - The development of new techniques for text processing pararellization
  - The development of information gathering modules for Facebook, Twitter, etc
- **Software**
  - DBMS Postgresql 9.2
  - pgAdmin 1.18.1
  - Python 3.7
  - Rapid Miner 5.3
  - etc.

We have analyzed your performance in the company over the past year ...
Conclusion: So far we are fairly confident that your first name is 'Rob' ...
joyreactor.com

Images analysis
General scheme

- Data sources
  - Textual content
  - **Images**
  - Links
- General page processing scheme:





Visual content analysis (1/3)
Source data

- Source data
  - 10 000 pages in total (1000 for each of the ten categories)

Category	Web pages with images	Total amount of images
adult	881	31563
alcohol	844	12489
chat	704	6411
ecommerce	862	16843
gamesonline	728	7061
hunting	902	24684
medical	814	9057
music	751	8197
news	817	9194
religion	830	9346
Total	8133	134845

Category	The most relevant images categories
adult	swimming_trunks, bikini, tub, miniskirt, bathtub, brassiere, maillot, diaper, bathing_cap, dumbbell
alcohol	wine_bottle, red_wine, barrel, beer_bottle, beer_glass, rapeseed, lotion, cocktail_shaker, goblet, valley
chat	daisy, Egyptian_cat, beacon, iPod, pencil_sharpener, gown, oscilloscope, crossword_puzzle, bookshop, bow_tie
ecommerce	pill_bottle, hair_slide, nipple, mailbag, clog, vase, necklace, wallet, loupe, wool
gamesonline	slot, mask, balloon, jackolantern, snowplow, toyshop, tractor, soccer_ball, parachute, drilling_platform
hunting	hartebeest, ibex, impala, gazelle, bighorn, bison, American_black_bear, water_buffalo, ox, Arabian_camel
medical	stethoscope, lab_coat, barber_chair, stole, jellyfish, airliner, notebook, ambulance, swab, paintbrush
music	acoustic_guitar, violin, electric_guitar, stage, cello, grand_piano, banjo, sax, oboe, bassoon
news	football_helmet, flagpole, unicycle, volleyball, kimono, military_uniform, streetcar, missile, dock, harp
religion	church, vestment, cloak, altar, academic_gown, monastery, obelisk, iron, cleaver, restaurant

Visual content analysis (2/3)
Images classification

- Experiments with images classifier
  - some of the categories can be recognized by images (e.g. adult, alcohol, hunting), some of them - not (e.g. news, chat)
  - the classification quality depends on the category, e.g. the category “news” can have various content. That is why the list of terms for this category contains the same words related to sport (football_helmet, volleyball), army (missile, military_uniform), music (harp), transport (streetcar), etc

Category	Precision	Recall	F-measure
adult	0,877551	0,5	0,637037
alcohol	0,9	0,313953	0,465517
chat	0,6	0,034483	0,065217
ecommerce	0,619048	0,149425	0,240741
gamesonline	0,478261	0,127907	0,201835
hunting	0,974651	0,255814	0,407407
medical	0,521739	0,139535	0,220183
music	0,846154	0,127907	0,222222
news	0,54	0,046512	0,085106
religion	0,586207	0,197674	0,295652

Accuracy	Errors	Unknowns	Accuracy w/o unknowns	Errors w/o unknowns
0,1891	0,0696	0,7412	0,7309	0,2691

Visual content analysis (3/3)
Combining of the classifiers

- Experiments with combination of the classifiers
  - not all pages contained images, some of the images were advertising banners, and all sites contained the text for classification, nevertheless the use of the image classifier allowed to improve the quality of classification by more than 6%
  - the image classifier was not retrained

Category	W/o images	Images	Total
adult	0,942857	0,877551	0,961568
alcohol	0,947712	0,9	0,958742
chat	0,64	0,6	0,645161
ecommerce	0,792453	0,619048	0,804954
gamesonline	0,904255	0,478261	0,907548
hunting	0,961039	0,974651	0,980584
medical	0,903226	0,521739	0,910265
music	0,8125	0,846154	0,851684
news	0,75	0,54	0,751218
religion	0,841463	0,586207	0,851613

	Accuracy	Errors	Unknowns	Accuracy w/o unknowns	Errors w/o unknowns
w/o images	0,52	0,0695	0,4105	0,88210	0,117897
Images	0,1891	0,0696	0,7412	0,7309	0,2691
Combined	0,58	0,0715	0,3485	0,89025	0,129747

Images classification
Summarizing, future works and examples

Summarizing

- The proposed approach provides an opportunity to setup more easily (than monolithic classifier) the process of maintaining and extending of web categorization scheme
- The addition of the image classification module made it possible to classify pages that for some reason cannot be classified by text (information in a foreign language, insufficient amount of text on the web page, etc.)

Future works

- To add simultaneously several image classifiers to the common architecture
- To analyze the possibility of using classifiers based on the information about the domain (WhoIs servers' response)
- To perform additional experiments

Software

- DBMS Postgresql 9.2
- pgAdmin 1.18.1
- Python 3.7
- ImageNet + ResNet
- etc.

General system architecture
Content analysis

Monitoring

- Tracking
- Links
- Content

- Detection of dangerous influence
- Attack sources
- Target audience
- Information distribution channels
- Disseminated information

Counteraction

- Developing a list of countermeasures
  - Target of the countermeasure
  - Type of countermeasure
- Decision support
  - Modeling
  - Resources evaluation

Information counteraction

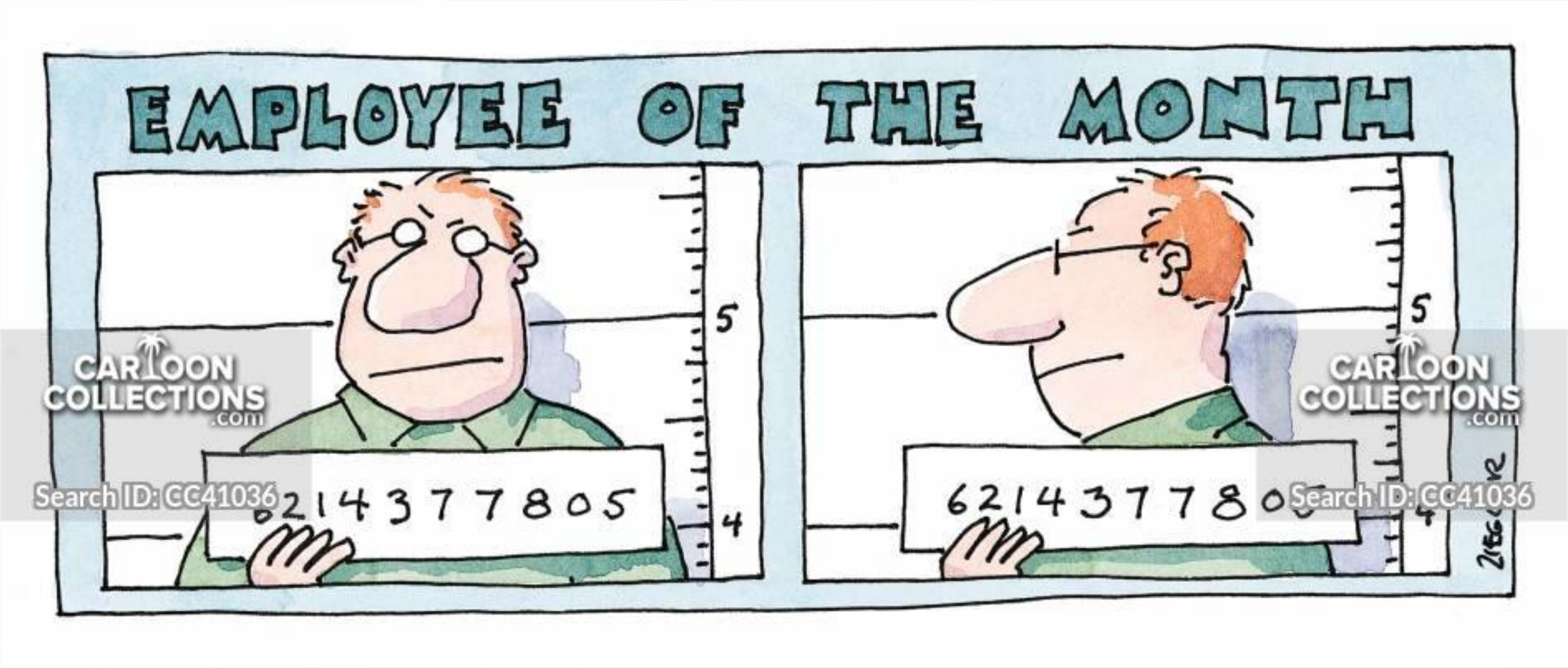
Countermeasures:

- Blocking of information objects that contain inappropriate information
- Blocking of inappropriate information sources
- Disruption of connectivity of inappropriate information distribution networks
- Noising of distribution channels and information sources of the target audience
- Switching attention of the target audience

"Someone's tweeting that there's a fly in his soup."

Acknowledgements

- Dmitry Komashinskiy (F-Secure)
- Dmitry Levshun (SPIIRAS, ITMO, UPS)
- Lidia Vitkova (SPIIRAS, SPbSUT)
- Maxim Kolomeets (SPIIRAS, ITMO, UPS)



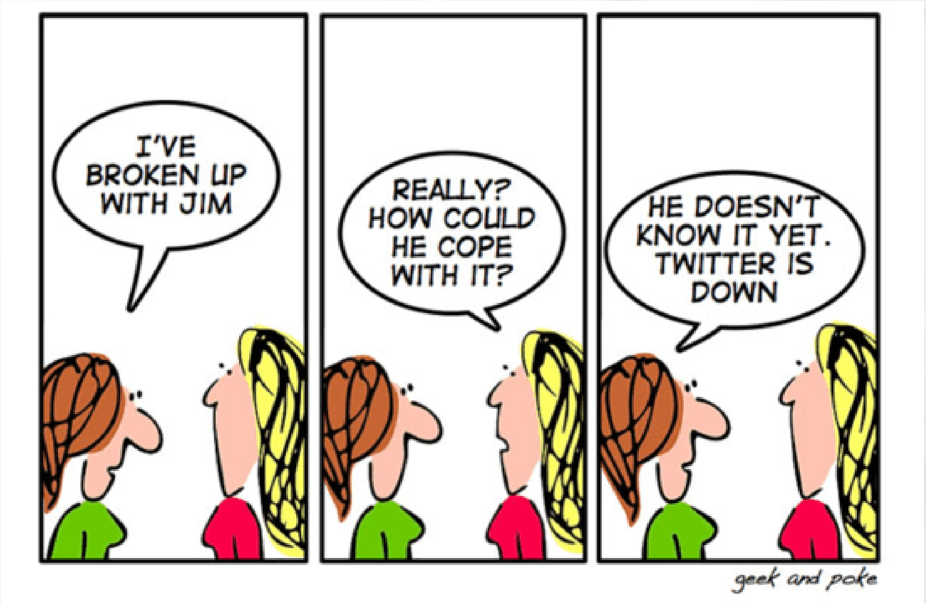
Discussion?

Thank you for your attention!

Questions?

Contact information:

Andrey Chechulin (chechulin@comsec.spb.ru)


I'VE BROKEN UP WITH JIM
REALLY? HOW COULD HE COPE WITH IT?
HE DOESN'T KNOW IT YET. TWITTER IS DOWN
geek and poke

http://comsec.spb.ru/chechulin