# СИСТЕМА ЗАЩИТЫ ПОЛЬЗОВАТЕЛЕЙ ОТ НЕЖЕЛАТЕЛЬНОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

**с.н.с., к.т.н. Чечулин А.А.**

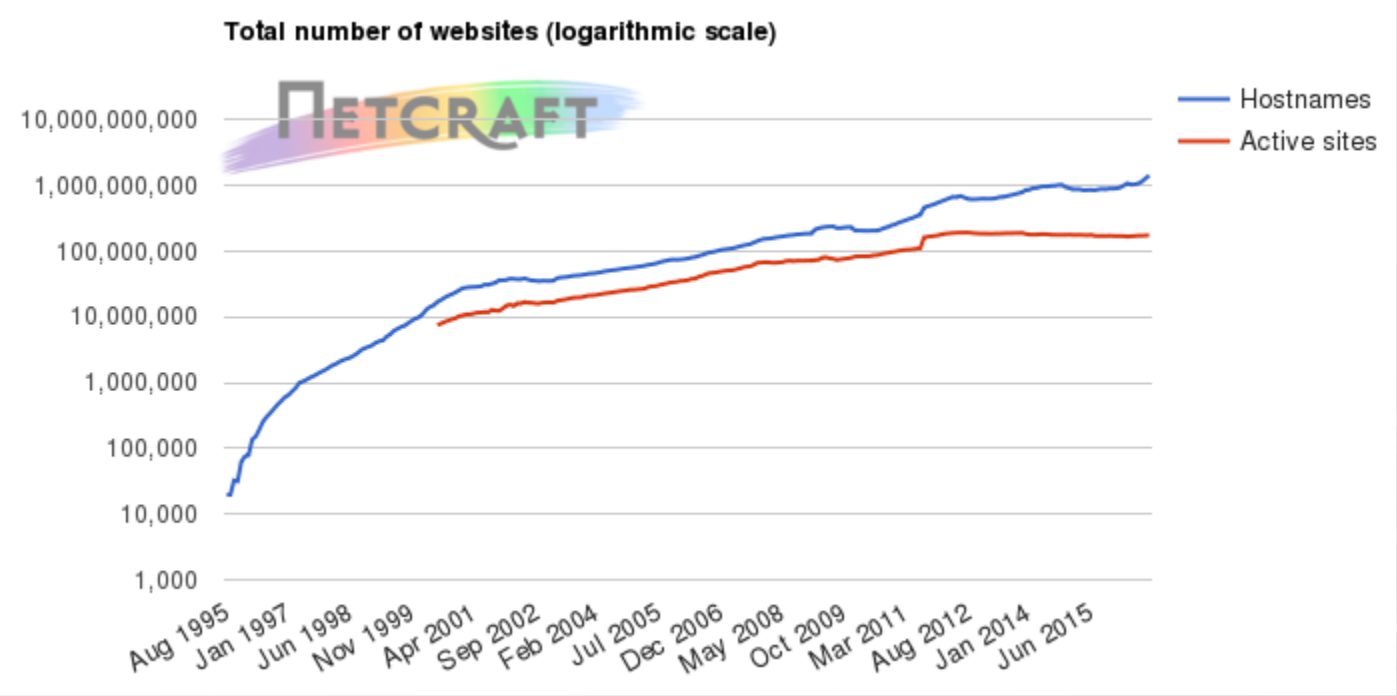
Лаборатория проблем компьютерной безопасности

Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН)

Санкт-Петербург, Россия

# Немного статистики

- На октябрь 2016 года в Рунете (зона .ru, .рф и .su) зарегистрировано 5 381 819 (ru), 906 303 (рф) и 119 106 (su) доменов (http://statdom.ru).
- В то же время по данным Netcraft (http://news.netcraft.com/) в октябре 2016 г. в Интернете функционирует 1 429 331 486 сайтов.



# Неприемлемые сайты

**Федеральный закон № 139-ФЗ от 28 июля 2012 года** описывает необходимость блокировать сайты, содержащие:

- материалы с порнографическими изображениями несовершеннолетних и (или) объявлений о привлечении несовершеннолетних в качестве исполнителей для участия в зрелищных мероприятиях порнографического характера;
- пропаганду употребления наркотиков и психотропных веществ, информацию о способах их производства и местах приобретения;
- информацию о способах совершения самоубийства, а также призывов к совершению самоубийства;
- **любую иную информацию, запрещённую к распространению в России решениями судов**.

# Неприемлемые сайты для детей

- Некоторые категории веб-сайтов могут негативно повлиять на развитие и психику ребенка, например:
  - содержащие материалы порнографического и эротического характера;
  - рекламирующие алкогольные напитки;
  - связанные с пропагандой сектантства;
  - сайты знакомств;
  - посвященные азартным играм;
  - призывающие к расовой, религиозной и т.п. дискриминации;
  - реклама табакокурения;
  - демонстрирующее кровь и насилие;
  - связанные с пропагандой оружия.

# Целевая аудитория и проблемы

- **Кому это надо?**
  - Разработчики в области информационной безопасности
  - Частные лица (родители?)
  - Организации
  - Правительство
- **Проблемы**
  - Объем данных
  - Скорость обработки
  - Сложная структура информации
  - Динамическое содержимое
  - Разные типы данных
  - Разные языки
  - …

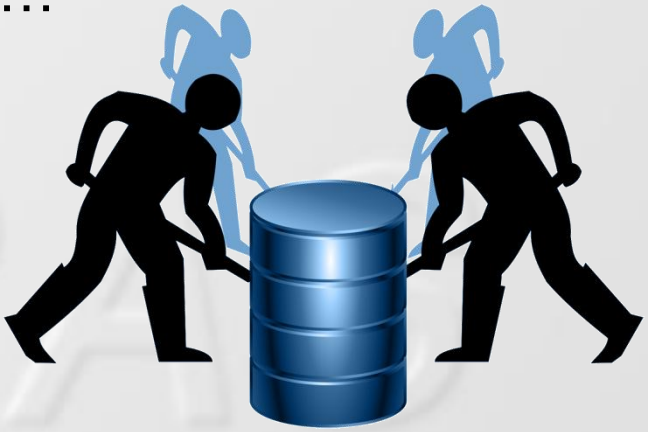
# Постановка задачи

- Цель работы:
  - Разработка системы категоризации веб-сайтов для блокировки сайтов с неприемлемым содержимым, в т.ч на иностранных языках
- Задачи:
  - Анализ способов определения категории веб-сайтов
  - Определение исходных данных
  - Разработка архитектуры системы
  - Реализация архитектуры в программном прототипе
  - Проведение экспериментов для проверки качества работы предложенного подхода
  - Оценка результатов

# Способы классификации

- Способы классификации
  - Ручной анализ
  - Анализ списков классифицированных сайтов
  - Интеллектуальный анализ данных (Data Mining)

- Методы Data Mining
  - Decision Trees (DT)
  - K-nearest neighbors (kNN)
  - Naive Bayes (NB)
  - Support Vector Machine (SVM)
  - Нейросети
  - …

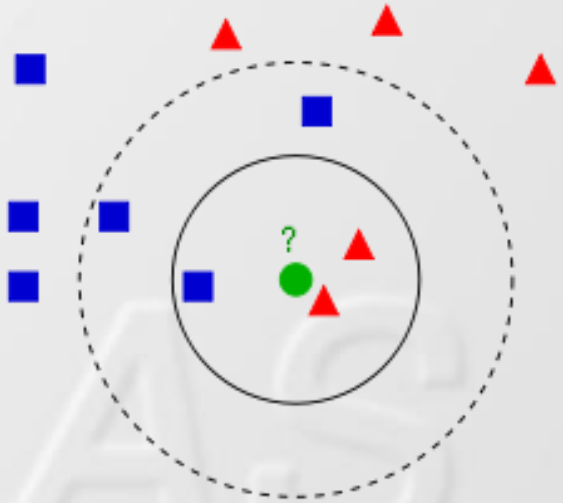


# Методы Data Mining

**Цель** состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

**Дерево принятия решений** (Decision Tree, дерево классификации или регрессионное деревом)

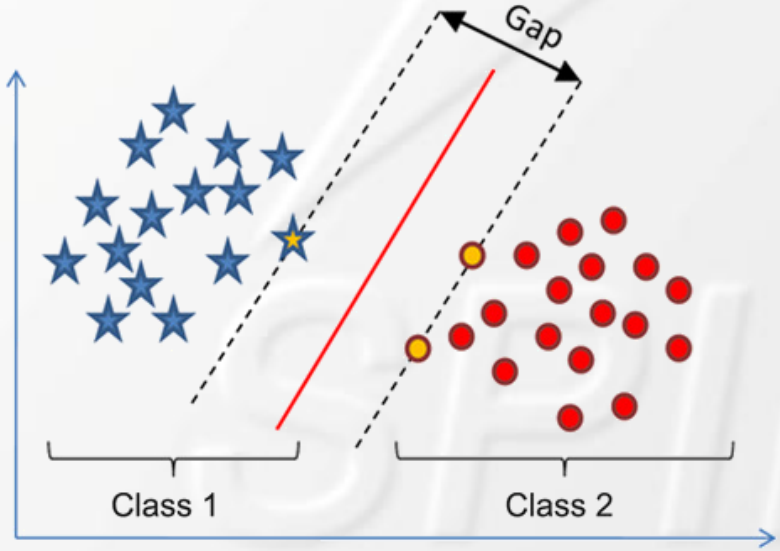

**Метод k ближайших соседей** (k-nearest neighbors algorithm, k-NN)



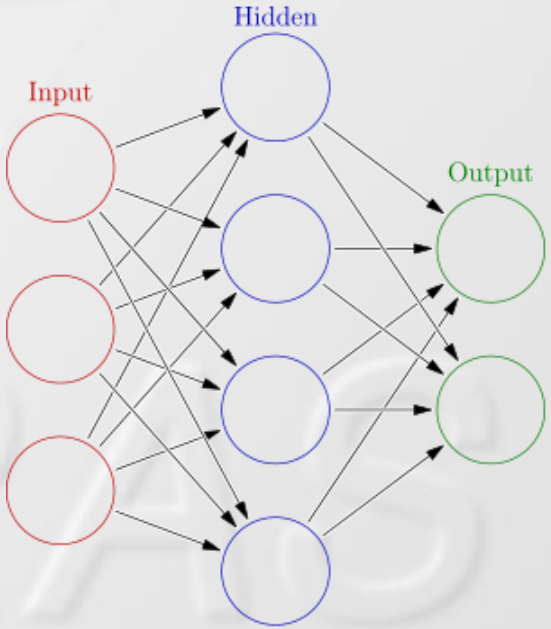
# Методы Data Mining

**Цель** состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

**Метод опорных векторов** (SVM, support vector machine)




**Искусственная нейронная сеть** (ИНС, Neural Networks)



# Исходные данные

- Источники признаков:
  - Текст сайта
  - Текст из тегов сайта
  - Теги сайта
  - Изображения
  - URL
  - Ссылки между сайтами
  - Каталоги сайтов
  - Ответы от WhoIS
  - История изменения
  - Репутация
  - Мультимедия
  - Скрипты
  - …
- Общие характеристики:
  - Количество признаков
  - Типы признаков (binomial, integer, float, enumeration, и т.д.)
  - Источники признаков (списки сайтов для тренировок и оценки)

# Список сайтов

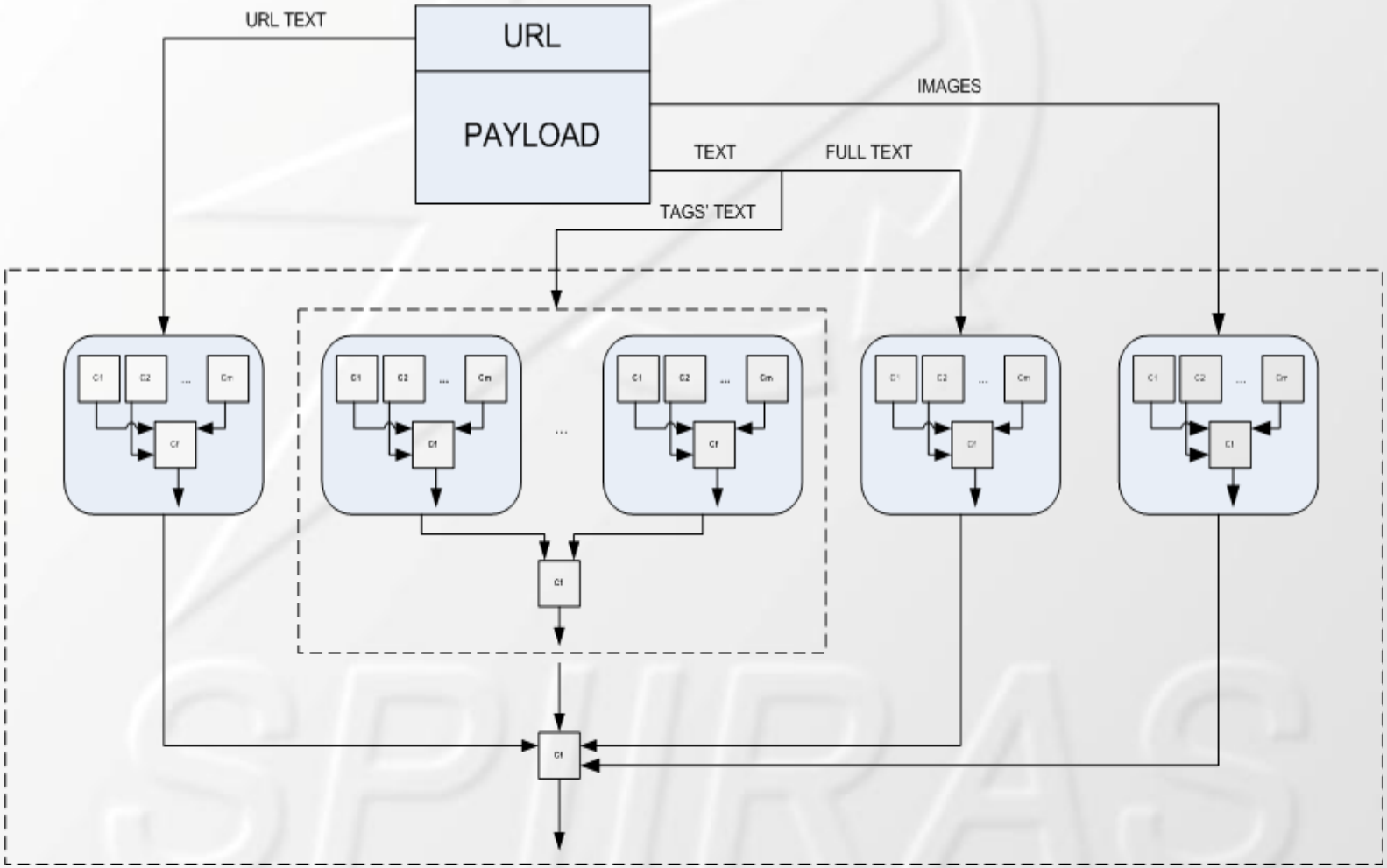
- Подготовка обученной модели производилась на основе выборки состоящей из 40000 сайтов.
- Тестовая выборка состояла из 4000 сайтов.
- Характеристики
  - 19+1 категорий: Категории: Adult, Alcohol, Banking, Blogs, Cults, Dating, Drugs, Forum, Gambling, Games, Hate, Health, Job_Search, News, Sport, Tobacco, Travel, Violence, Weapons + Unknown
  - Источники данных:
    - DMOZ (the Open Directory Project, http://rdf.dmoz.org/rdf/ )
    - URLBlackList (http://urlblacklist.com/)
  - Сайты для обучающей выборки не всегда имели высокую степень релевантности категориям которыми они были обозначены.

# Метрики оценки качества классификации

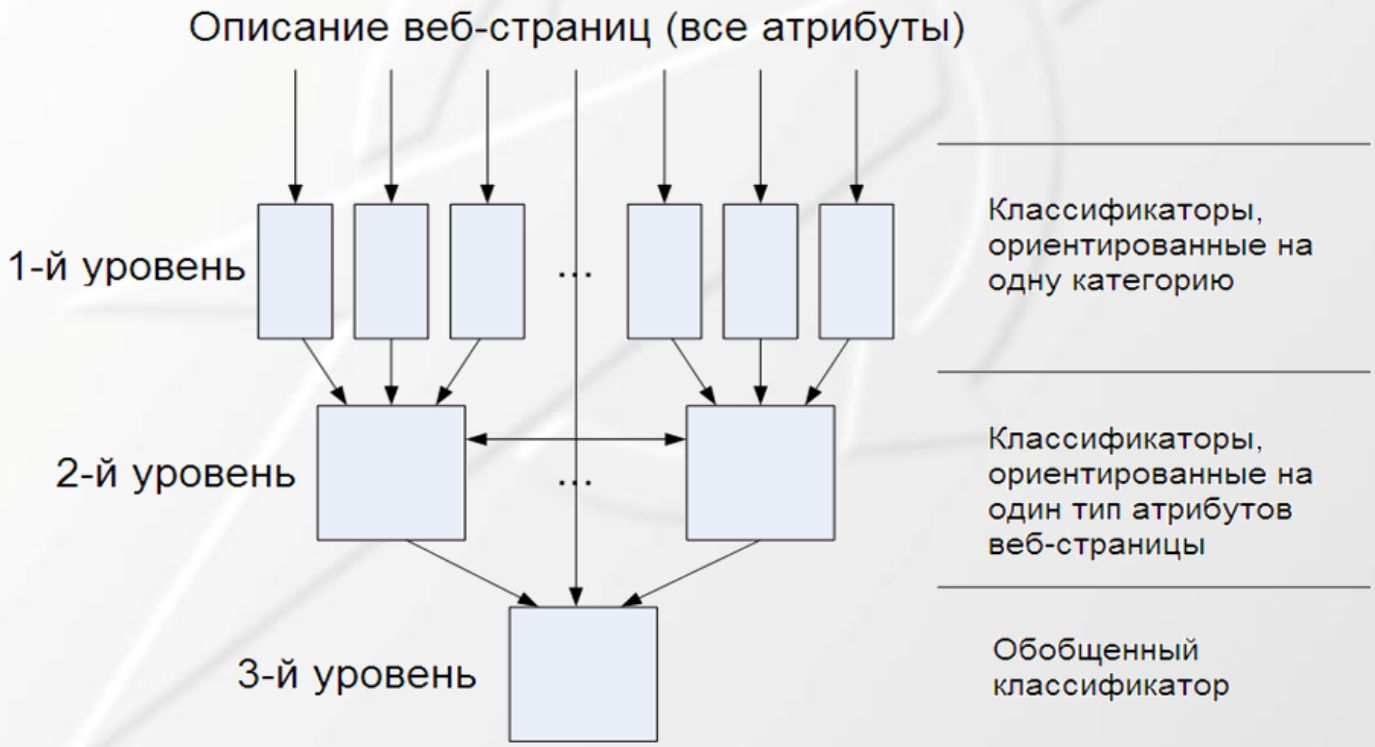
| Относится ли сайт к категории | | Эксперт | |
|---|---|---|---|
| | | TRUE | FALSE |
| Классификатор | TRUE | *TP (true positive)* | *FP (false positive)* |
| | FALSE | *FN (false negative)* | *TN (true negative)* |

- Для каждой категории вычисляется:
  - $a$ (true positive) – количество сайтов правильно распознанных как принадлежащие категории;
  - $b$ (false positive) – количество сайтов неправильно распознанных как принадлежащие категории;
  - $c$ (false negative) – количество сайтов неправильно распознанных как не принадлежащие категории;
  - $d$ (true negative) – количество сайтов правильно распознанных как не принадлежащие категории;
  - *Полнота* $(p) = \frac{a}{a+c}$; *Точность* $(r) = \frac{a}{a+b}$;
  - *F-мера (Fmeasure)* $= \frac{2pr}{p+r}$

# Архитектура
# Извлечение и обработка веб-сайтов



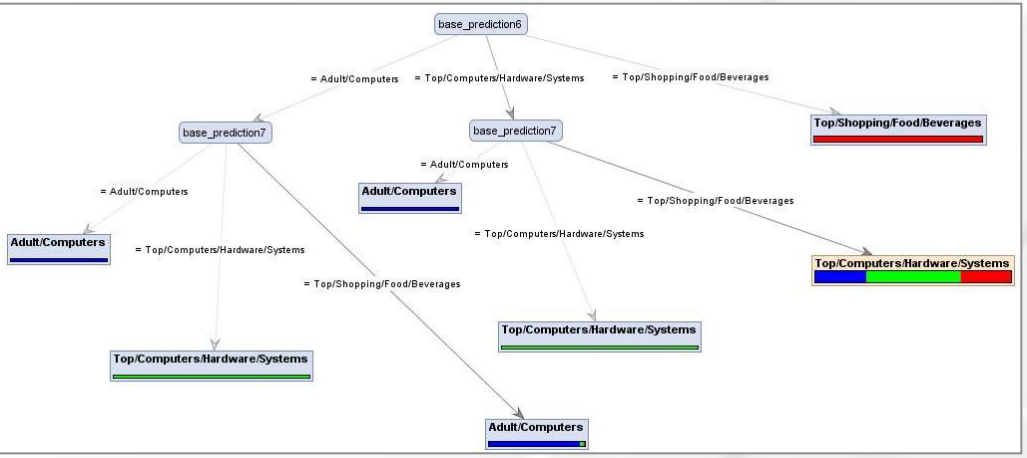
# Архитектура
# Общий подход к классификации



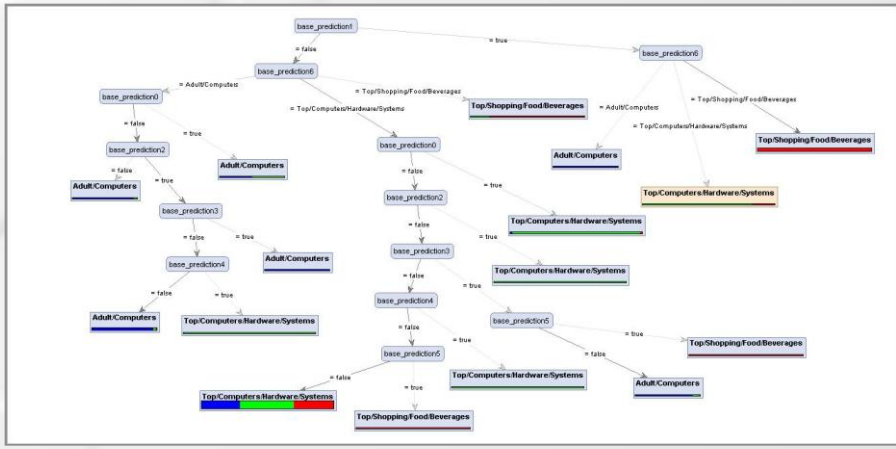

- Три подхода к фильтрации признаков на 3-м уровне
  - ***Main Units*** – классификатор использует только 2-й уровень
  - ***Mixed Units*** – классификатор использует и 1-й и 2-й уровни
  - ***Mixed Units with extended feature set*** – классификатор использует 1-й и 2-й уровни вместе с основными признаками

# Архитектура
# Выбор принципов фильтрации признаков

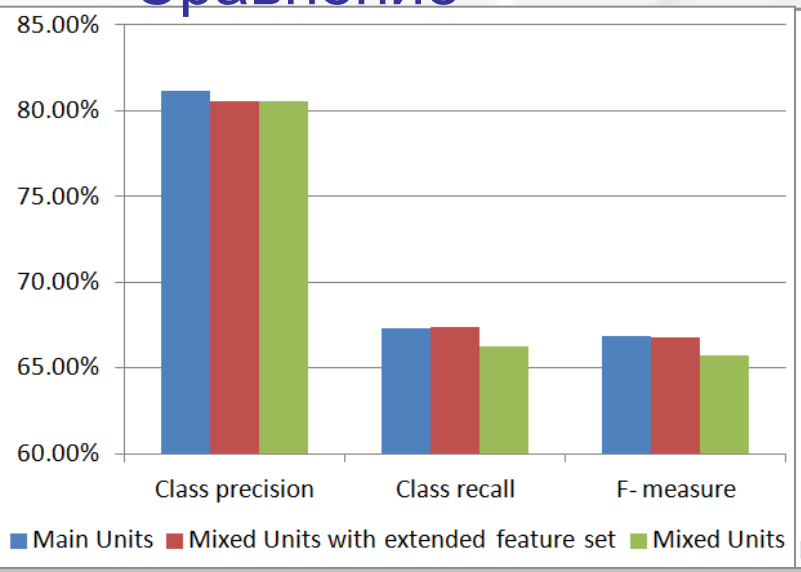
## Main Units



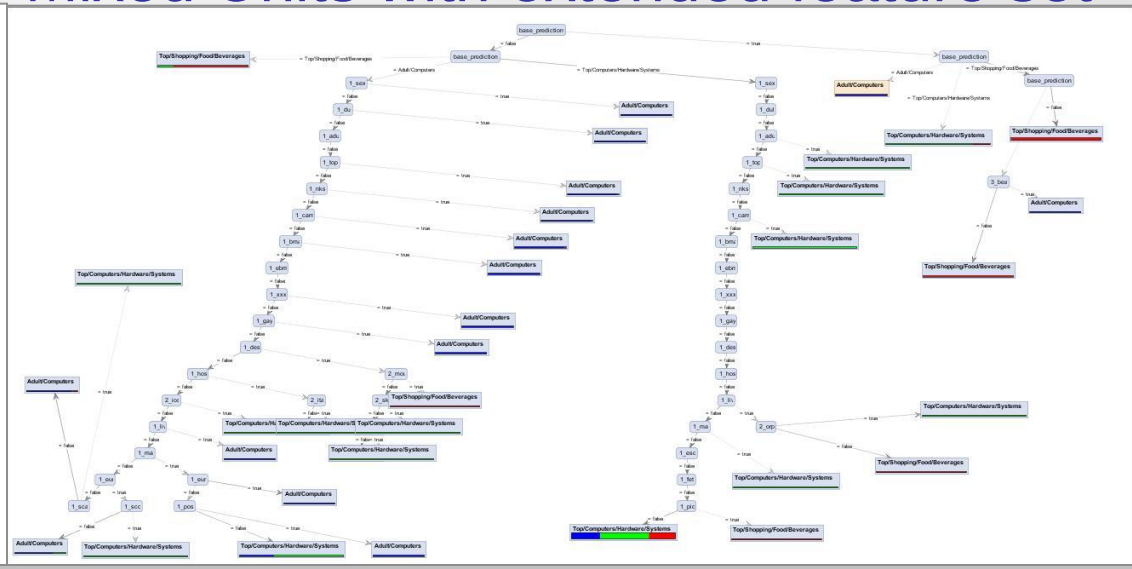

## Mixed Units



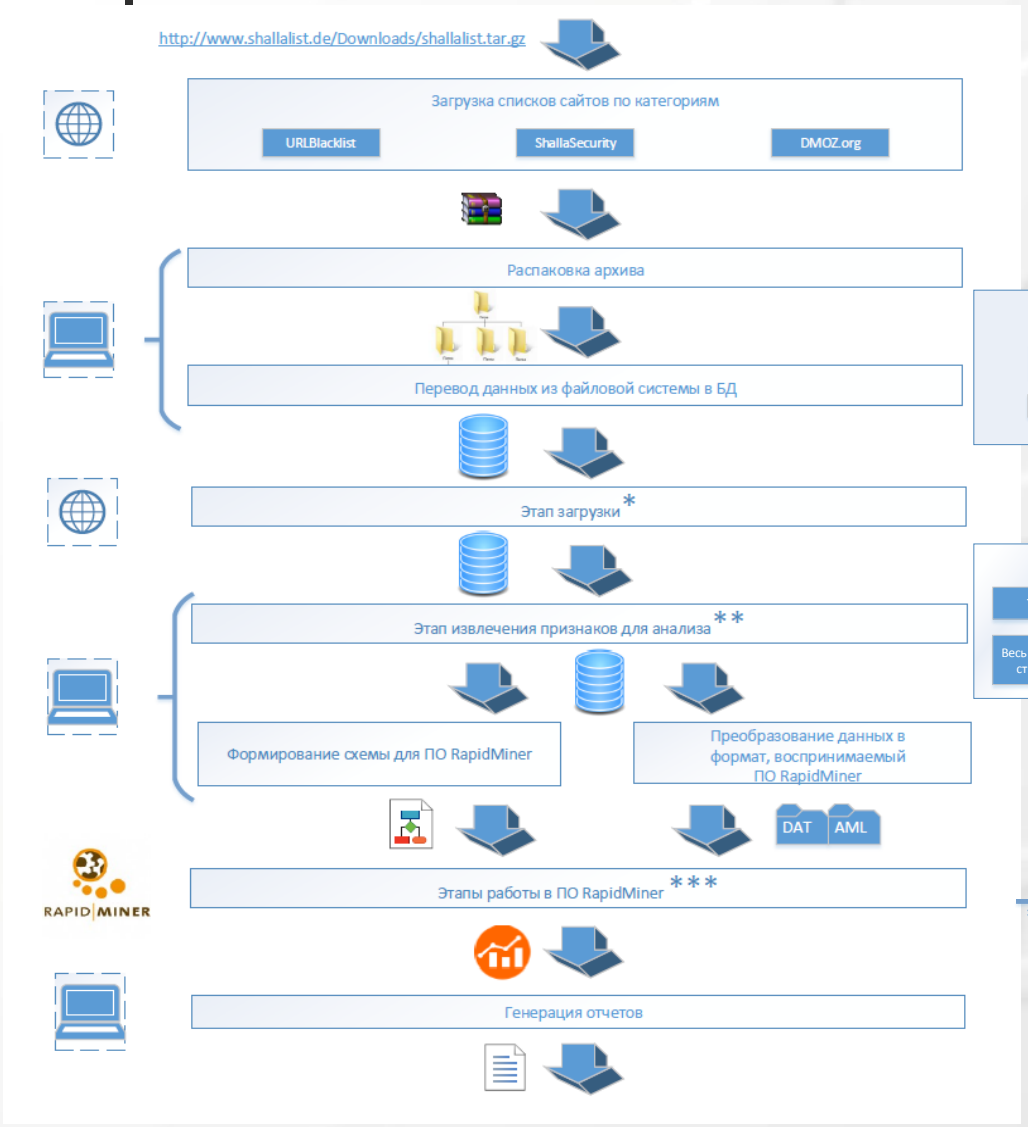
## Сравнение

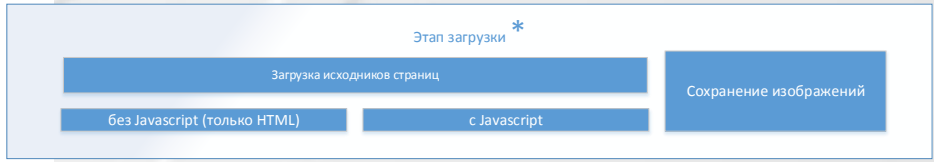


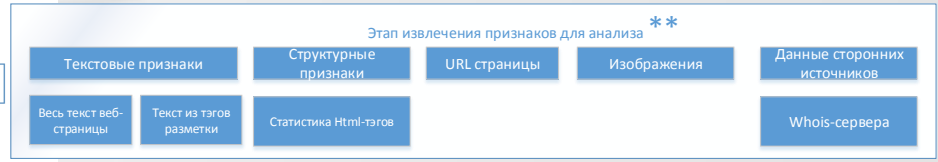

## Mixed Units with extended feature set

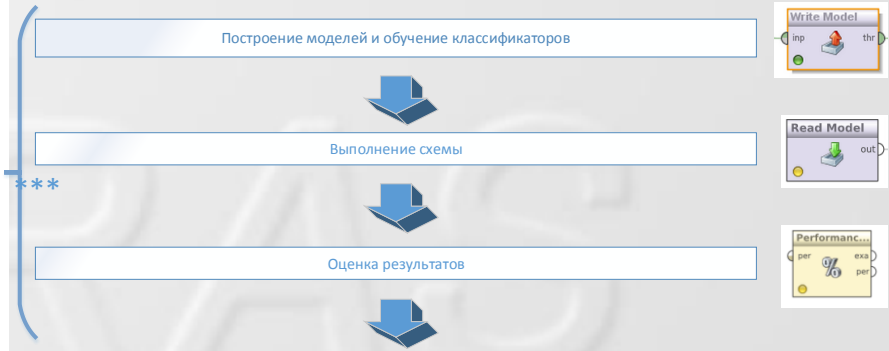


# Реализация
# Тестовый стенд

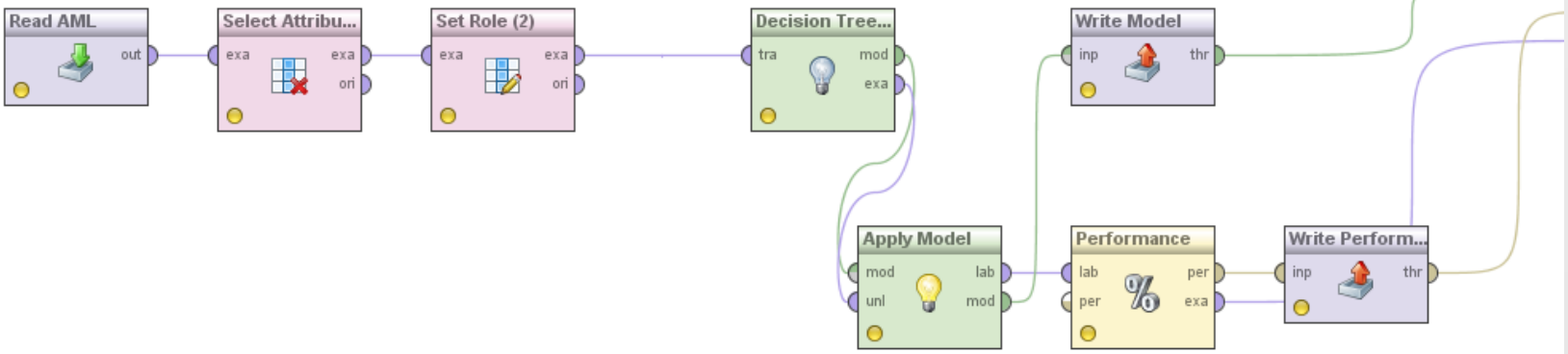


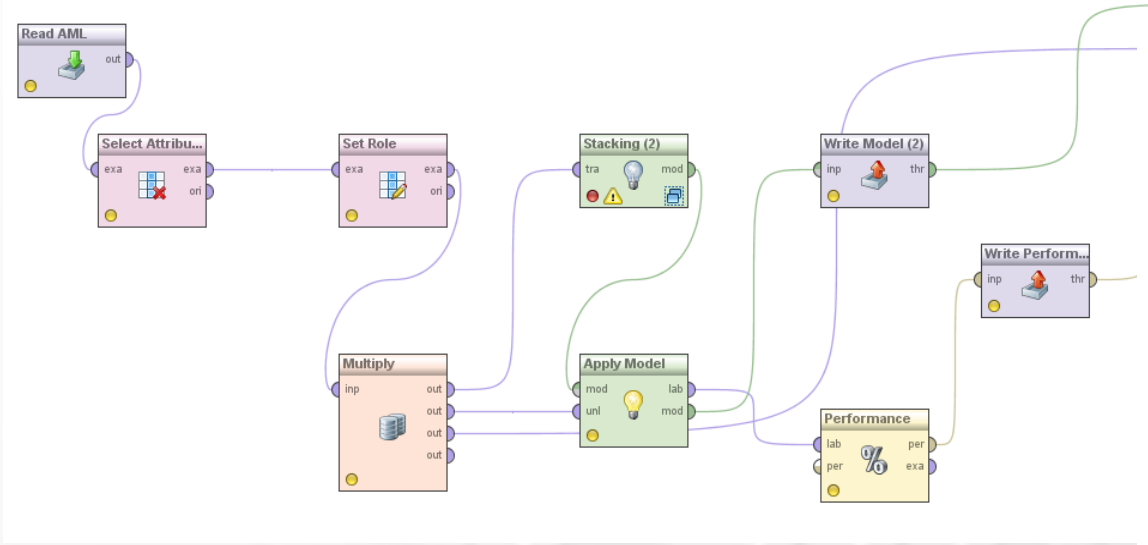



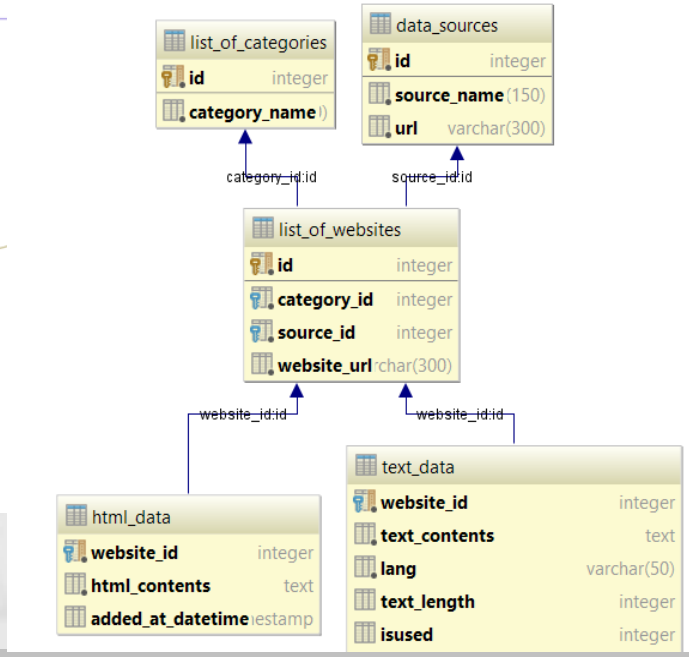



# Реализация
# Тестовый стенд



# Содержимое сайтов

Для экспериментов использовались следующие данные:

- URL web-страницы;
- HTML файл web-страницы;
  - Текст;
  - Текст из 10 наиболее встречаемых тегов

  (title, p, a, div, meta:content, li, link:title, span, h1, h2);

В перспективе предполагается использовать изображения, находящиеся на веб-странице, для повышения точности классификации.

```
47          </td></tr></table></td>
48      <td width="1" bgcolor="#000000" rowspan="101"><img src="/img/spacer.gif" width="1" height="1"></td>
49    </tr>
50  <tr><td height='34' valign='top' background='/img/bottom-header.jpg'><img src='/img/space.gif' width='1' height='3'>
51    being organized by Researchers of Laboratory of Computer Security Problems</h3>
52      <ul>
53      <li>Session on “Advanced research in cyber security" in the International Conference “RusCrypto’2013” (Solnechno
54      <li>21th Euromicro International Conference on Parallel, Distributed and network-based Processing (PDP 2013). Sp
55      </ul>
56
57    <h3>Former Conferences and Workshops
58    organized by Researchers of Computer Security Research Group</h3>
59      <ul>
```

# Подход к классификации текста

- Словарь
  - Подход TF/IDF и его модификация
  - Количество текста
- Обработка текста
  - Стемминг
  - Токены
  - Гипонимы
  - Гиперонимы
- Входные данные
  - Уникальный ID
  - Наличие слова на сайте
  - Категория сайта

| 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 |

# Внешние источники и иностранные языки

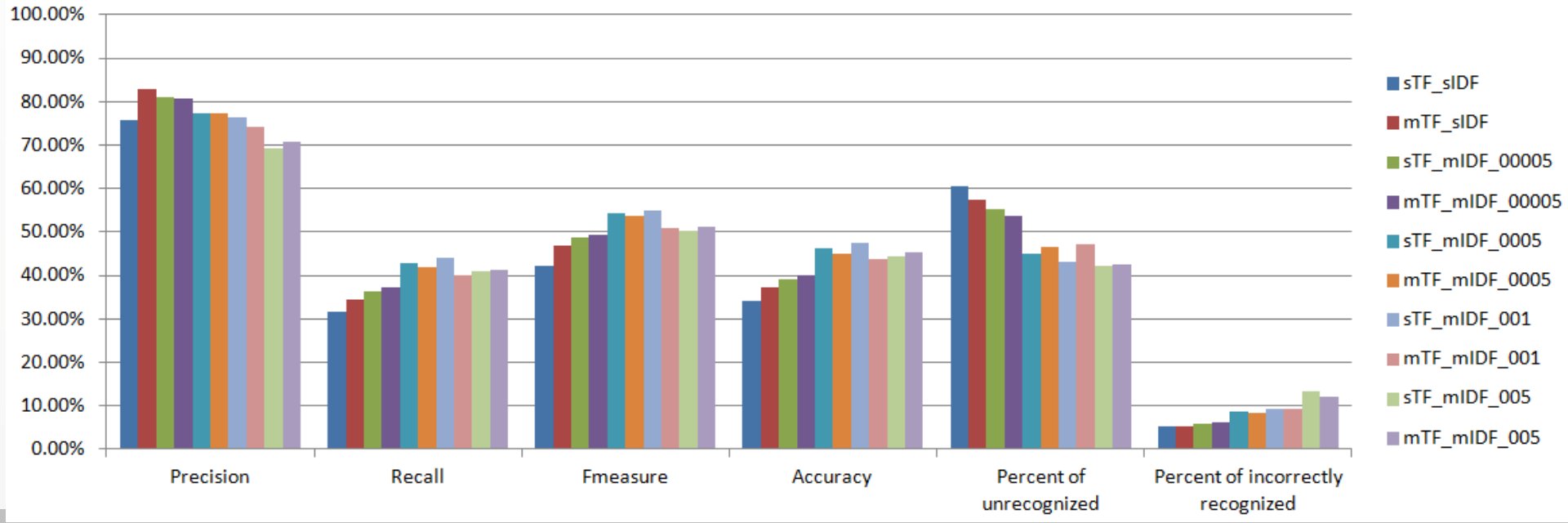
- Ответы от WhoIs
  - Первоначальные результаты показали, что ответы не имеют строгой структуры
  - Некоторые элементы ответов от WhoIs могут быть использованы как метаинформация для классификации
- Иностранные языки
  - Два подхода:
    - Формировать новые признаки
    - Переводить сайт на английский
  - Был использован 2-й подход через Yandex переводчик

```
domain:       SPB.RU
nserver:      ns3-geo.nic.ru.
nserver:      ns4-geo.nic.ru.
nserver:      ns8-geo.nic.ru.
state:        REGISTERED, DELEGATED, VERIFIED
org:          JSC 'RU-CENTER'
registrar:    RU-CENTER-REG-RIPN
admin-contact: https://www.nic.ru/whois
created:      1997.03.11
paid-till:    2015.04.01
free-date:    2015.05.02
source:       TCI

Last updated on 2014.07.15 13:56:34 MSK
```

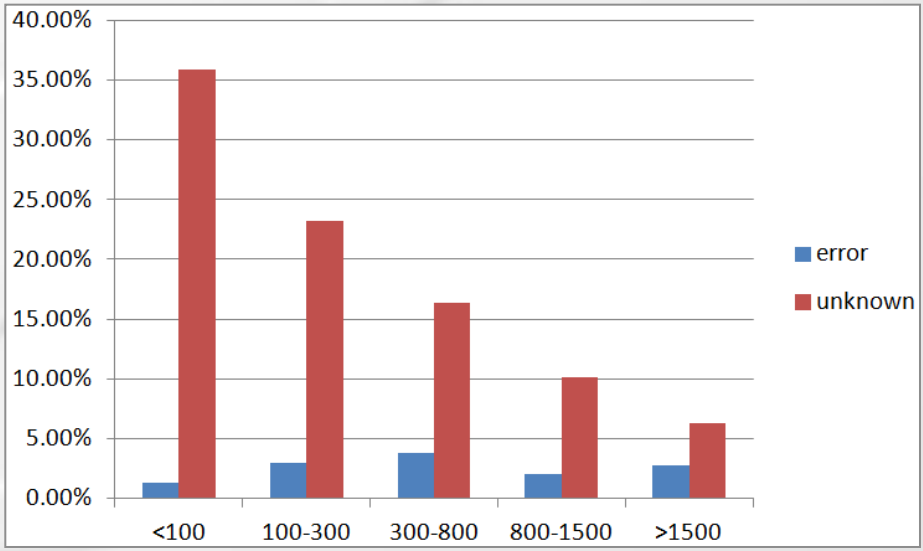
# Результаты текстовой классификации

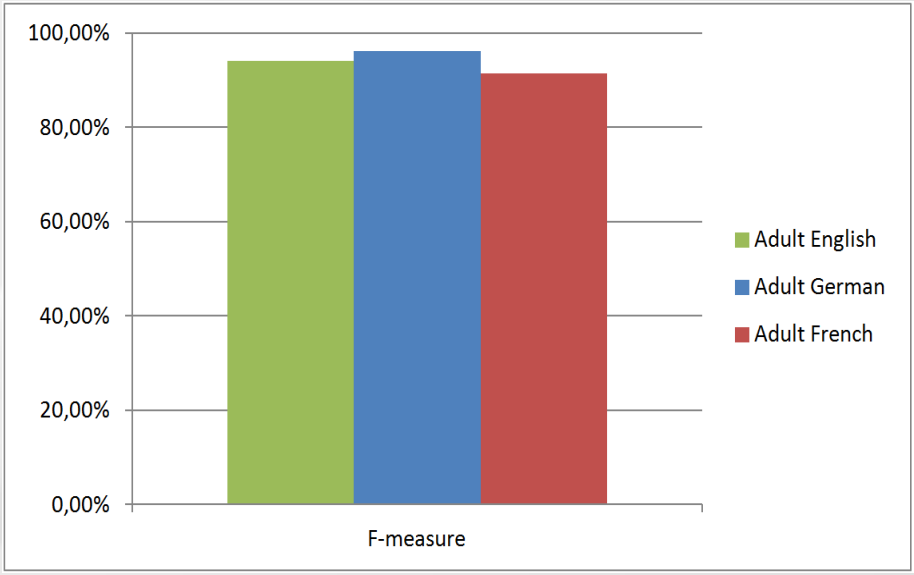
- Результаты экспериментов
  - Наличие схожих категорий приводит к понижению качества классификации (например, “Hate” and “Violence”, “Cults” and “Religion”, и т.д.)
  - Выбранная архитектура классификаторов позволяет получить хорошие показатели качества
  - Выбор деревьев решений в качестве классификаторов привело к превалированию точности над полнотой



# Результаты текстовой классификации

- Дополнительные результаты
  - Категория Unknown повышает качество классификации
  - Классификация основанная на тексте зависит от количество текста
  - Данный классификатор может быть использован в одиночку
- Иностранные языки
  - Система автоматического перевода (Yandex-translate)
  - Основной язык – английский
  - Снижается вариативность языка, что повышает качество классификации



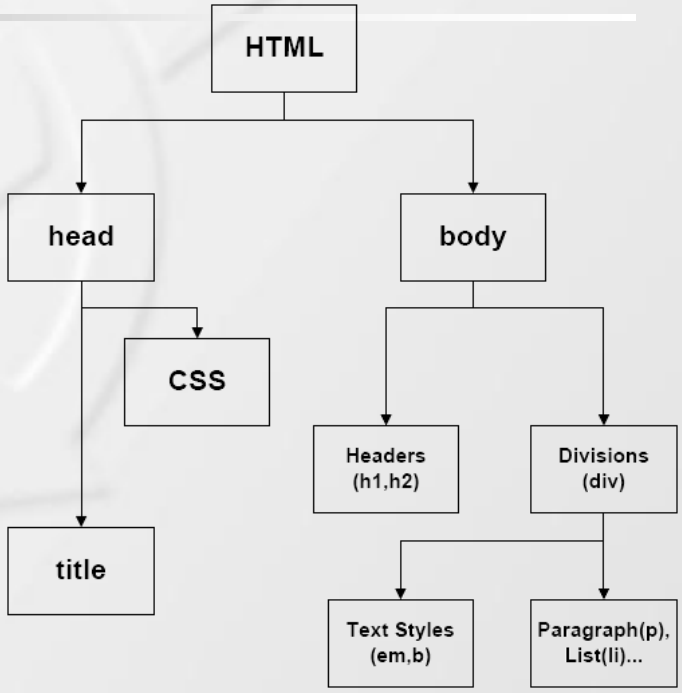
# Классификация по URL

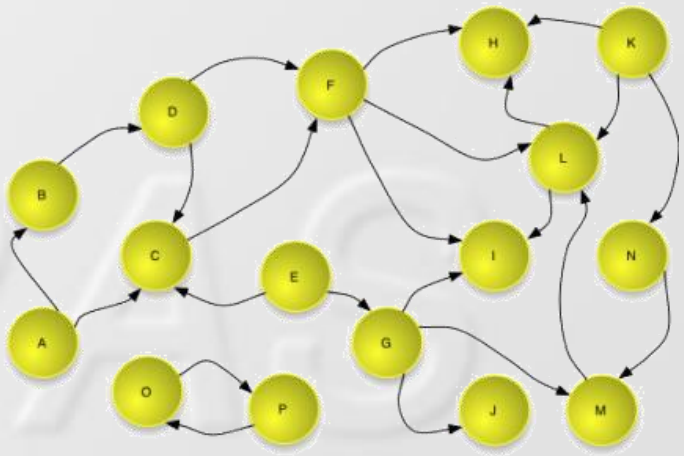
- Словарь (n-граммы)
  - (1) Сбор всех адресов веб-страниц (URL) относящихся к одной категории в файл
  - (2) формирование токенов на основе специальных символов
  - (3) устранение стоп-слов и применение стеммера
  - (4) формирование n-грамм
- Словарь
  - Подход TF/IDF
  - 50 3-грамм для каждой категории
- Входные данные
  - Уникальный ID
  - Наличие 3-граммы в URL
  - Категория сайта

| Category | Keywords |
|---|---|
| Adult | apd, bds, bep, deo, drt, dtu, eos, exv, fap, fuc, ... |
| Cults | aba, aex, aht, aib, aji, aof, asa, avy, bab, baj, ... |
| Dating | afu, all, ate, ati, aur, bsi, cli, cup, dat, dda, ... |
| Drugs | acy, agr, apv, axp, axs, dsf, ecs, gea, grx, hfr, ... |

# Структура HTML и ссылки

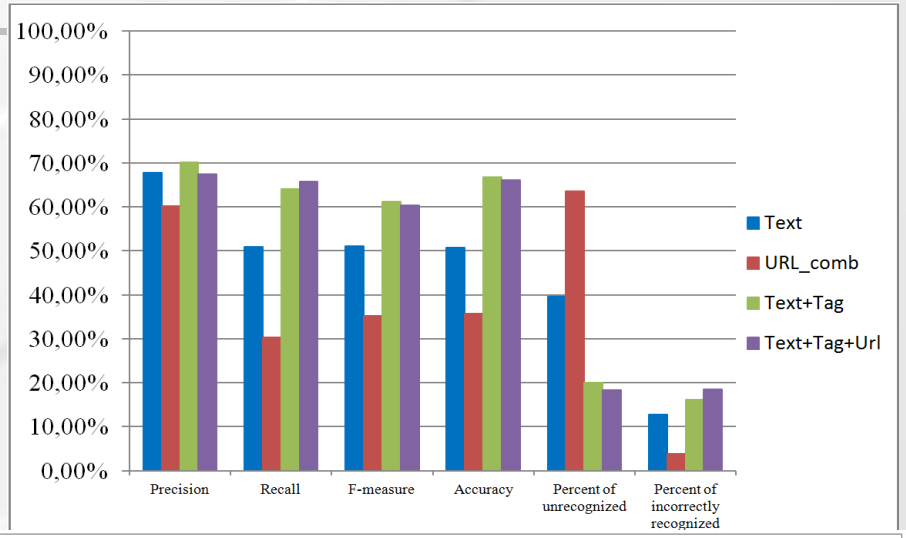
- Структура HTML
  - Эти признаки могут помочь с распознаванием сложных категорий (News, Blogs, Forums, etc)
  - Входные данные
    - Уникальный ID
    - Наличие относительное количество тегов
    - Категория сайта
- Ссылки
  - Эксперименты требуют больших объемов связанных сайтов
  - Результаты показали плохое качество, но причина – качество выборки

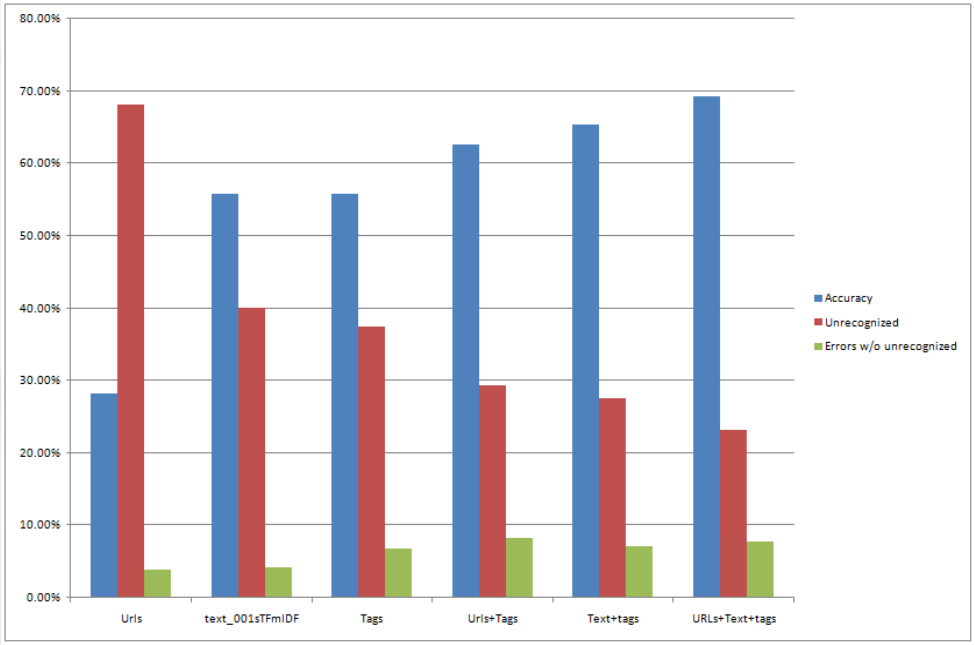


# Комбинирование классификаторов

- Результаты экспериментов
  - Предложенный подход может быть использован для построения систем защиты от нежелательной информации
  - Комбинирование источников позволяет повысить качество классификации
  - Основные источники данных:
    - Общая текстовая информация и текст из тегов
    - URL



# Заключение (1/2)
# Использованное ПО

- Для сбора и первичной обработки данных
  - Jsoup (http://jsoup.org);
  - NetBeans IDE (http://netbeans.org/);
  - Яндекс.Перевод (http://translate.yandex.ru/);
- Для хранения данных
  - Postgresql (http://www.postgresql.org/);
  - pgAdmin (http://www.pgadmin.org/);
- Для проведения экспериментов
  - RapidMiner (http://rapid-i.com/).

# Заключение

- Была представлена технология классификация веб-сайтов с помощью методов Machine Learning и Data Mining
- Представлены технология и инструменты, позволяющие автоматизировать процесс подготовки исходных данных и обученной модели
- Эксперименты показали высокую точность категоризации веб-сайтов, что дает возможность использования разработанной технологии в системах блокирования веб-сайтов с неприемлемым содержимым
- В дальнейшей работе планируется расширить список анализируемых языков и перейти от жесткой классификации к мягкой

# Вопросы

Спасибо за внимание

Вопросы?

Контактная информация:

Чечулин Андрей Алексеевич (chechulin@comsec.spb.ru)

http://comsec.spb.ru/chechulin