Категорирование Web-сайтов для систем блокирования Web-страниц с неприемлемым содержимым

Комашинский Д.В., Котенко И.В.,
Чечулин А.А., Шоров А.В.
(СПИИРАН)

Содержание

- Введение
- Архитектура
- Исходные данные
- Результаты экспериментов
- Заключение

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;
- пропаганду употребления прекурсоров 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.

Проблема языка

- На март 2013 года в Рунете (зона .ru и .рф) зарегистрировано примерно 5 224 000 доменов (http://statdom.ru).
- В то же время по данным Netcraft (http://news.netcraft.com/) в Интернете функционирует 631 521 198 сайт.
- Процент сайтов на русском языке (относительно их общего количества) в лучшем случае составляет менее 0,83%.
- Существует необходимость категоризировать веб-сайты независимо от языка, используемого для их наполнения.

Общая характеристика работы

- Цель работы:
  - Разработка системы категоризации веб-сайтов для блокировки сайтов с неприемлемым содержимым, в т.ч на иностранных языках;
- Задачи:
  - Анализ существующих моделей и методов определения категории веб-страниц;
  - Разработка архитектуры системы;
  - Реализация архитектуры в программном прототипе;
  - Проведение экспериментов для проверки качества работы предложенного подхода;

Релевантные работы (1/2)

- Описание типов текстовых исходных данных
  - Кузнецов Р.Ф. Классификатор веб-страниц на базе SVM-Multiclass // Труды РОМИП'2006.
- Общее описание методик анализа данных
  - Han J., Kamber M. Data Mining: Concepts and Techniques // Elsevier, Morgan Kaufman publishers, 2006.
- Подходы к классификации веб-страниц
  - Qi X., Davison B.D . Web Page Classification: Features and algorithms, ACM Computing Surveys (CSUR). 2009.
  - Calado P. et al. Combining link-based and content-based methods for Web document classification // In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM). 2003.

Релевантные работы (2/2)

Данная работа является продолжением исследований, выполняемых лабораторией с 2009 года.

- Анализ существующих моделей и методов определения категории веб-страниц:

  Зозуля Ю.В., Котенко И.В. Блокирование Web-сайтов с неприемлемым содержимым на основании выявления их категорий // РусКрипто'2010

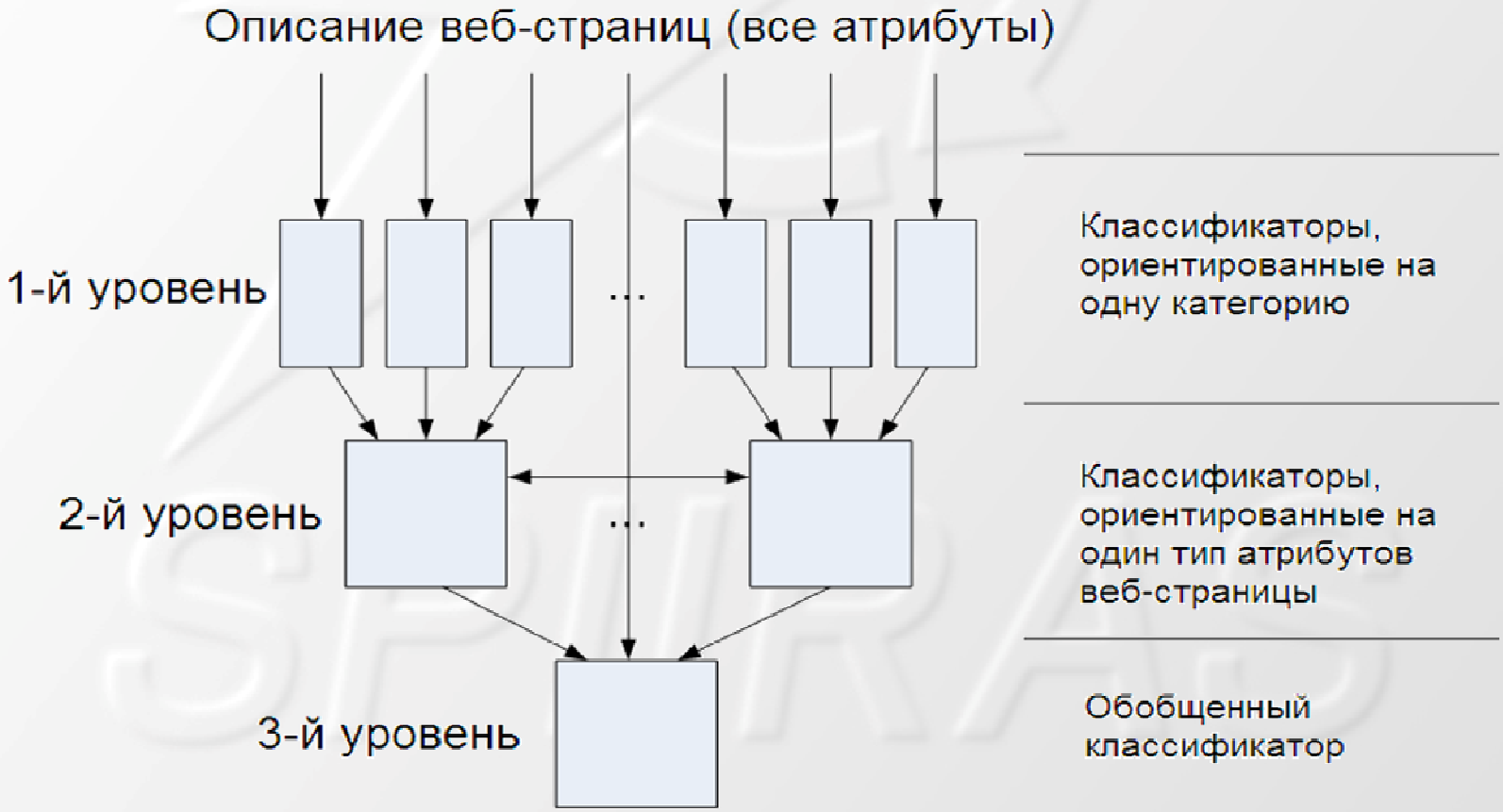
- Подход к категоризации веб-страниц с использованием технологий Machine Learning и Data Mining, архитектура системы категоризации :

  Чечулин А.А., Котенко И.В., Комашинский Д.В. Категорирование веб-страниц с неприемлемым содержимым // РусКрипто'2011

Общая архитектура системы (1/2)

Система классификации работает в двух режимах:

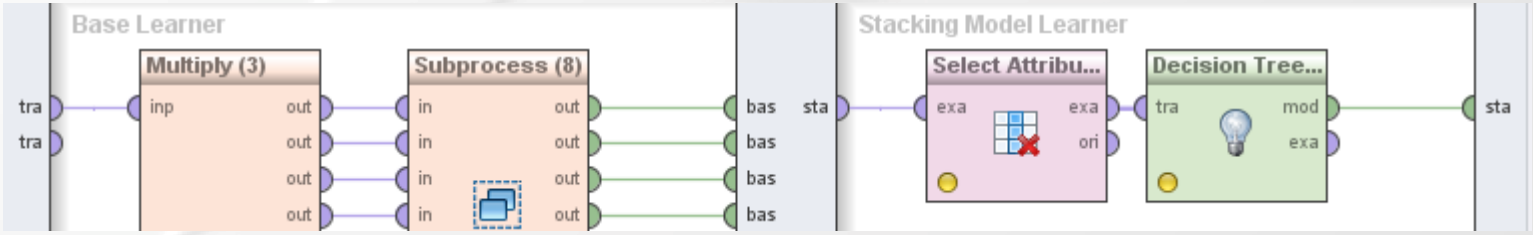
1. Подготовка общей обученной модели и ее тестирование.
2. Использование обученной модели для классификации неизвестных веб-сайтов.



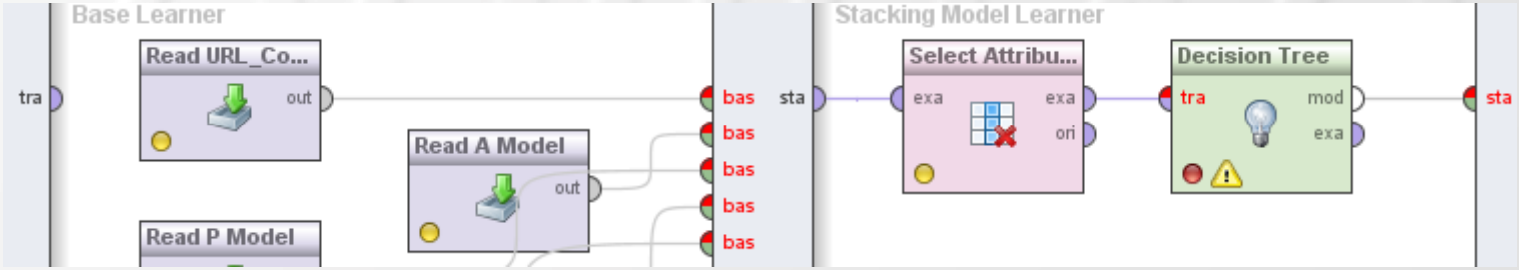
Общая архитектура системы (2/2)

- 3 уровня классификаторов:
  - 1-й уровень – классификаторы, принимающие решение о принадлежности поданного на анализ вектора атрибутов к конкретной категории;

  - 2-й уровень – принимают решение о принадлежности вектора атрибутов к одной категории из списка (аспектные классификаторы);



  - 3-й уровень – принимает окончательное решение на основе обобщения результатов аспектных классификаторов.



Используемые данные

Для категоризации web-сайтов 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" (Solnechnoc
54      <li>21th Euromicro International Conference on Parallel, Distributed and network-based Processing (PDP 2013). Spe
55      </ul>
56
57     <h3>Former Conferences and Workshops
58     organized by Researchers of Computer Security Research Group</h3>
59      <ul>
```

Подготовка обученной модели (1/2)

- Подготовка обученной модели производилась на основе выборки состоящей из 79063 сайтов.
  - Сайты для обучающей выборки имели высокую степень релевантности категориям которыми они были обозначены.

	id integer	url text
1	1	http://0-2-amateur-xxx-gay-lesbian-adult-videos.com/xxx-videos-adult-movies-home.asp
2	4	http://1-absolute-anal-sex-pics.com/cabanas-anal-sex-pics.htm
3	7	http://1-karas-xxx-adult-playground-kara-s-amateurs-adult-sex.com/karas-adult-pussy-playground.htm
4	9	http://1.stopreiner.org/interracial-anal
5	11	http://104sluts.ws/amateur cunts
6	12	http://123-sexe-amateur.com/enpleinair
7	20	http://123sexe.mylinea.com/amateurs
• • •		
79057	165464	http://www.youwin.com/en/sports-betting/football-odds&sa=u&ei=82ghumdrdirnmaxl-ecyaw&ved=0cbmqfjaaofe&usg=afq
79058	165465	http://www.youwin.com/en/sports-betting/tennis-odds&sa=u&ei=z2ghuilfmvdymax-1c24aw&ved=0cceqfjagoey&usg=afqjc
79059	165468	http://www.yuhip.org/top-soccer-betting-online&sa=u&ei=-mghup2khqtnmawh7pxaaw&ved=0cceqfjagonqc&usg=afqjcned-
79060	165469	http://www.zimbio.com/.../michael+phelps+betting+odds+2012+london+olympics
79061	165470	http://www.zimbio.com/.../articles/0h0zi4cel-q/important+tips+tennis+betting
79062	165476	http://www2.acttab.com.au/site/livebetting.php
79063	165485	http://zomobo.net/hockey-betting&sa=u&ei=428hupjteutjmawnrncvaw&ved=0ccgqfjafoj4f&usg=afqjcne9nwi0lyatpfoyfyb

- Тестовая выборка состояла из 7697 сайтов.

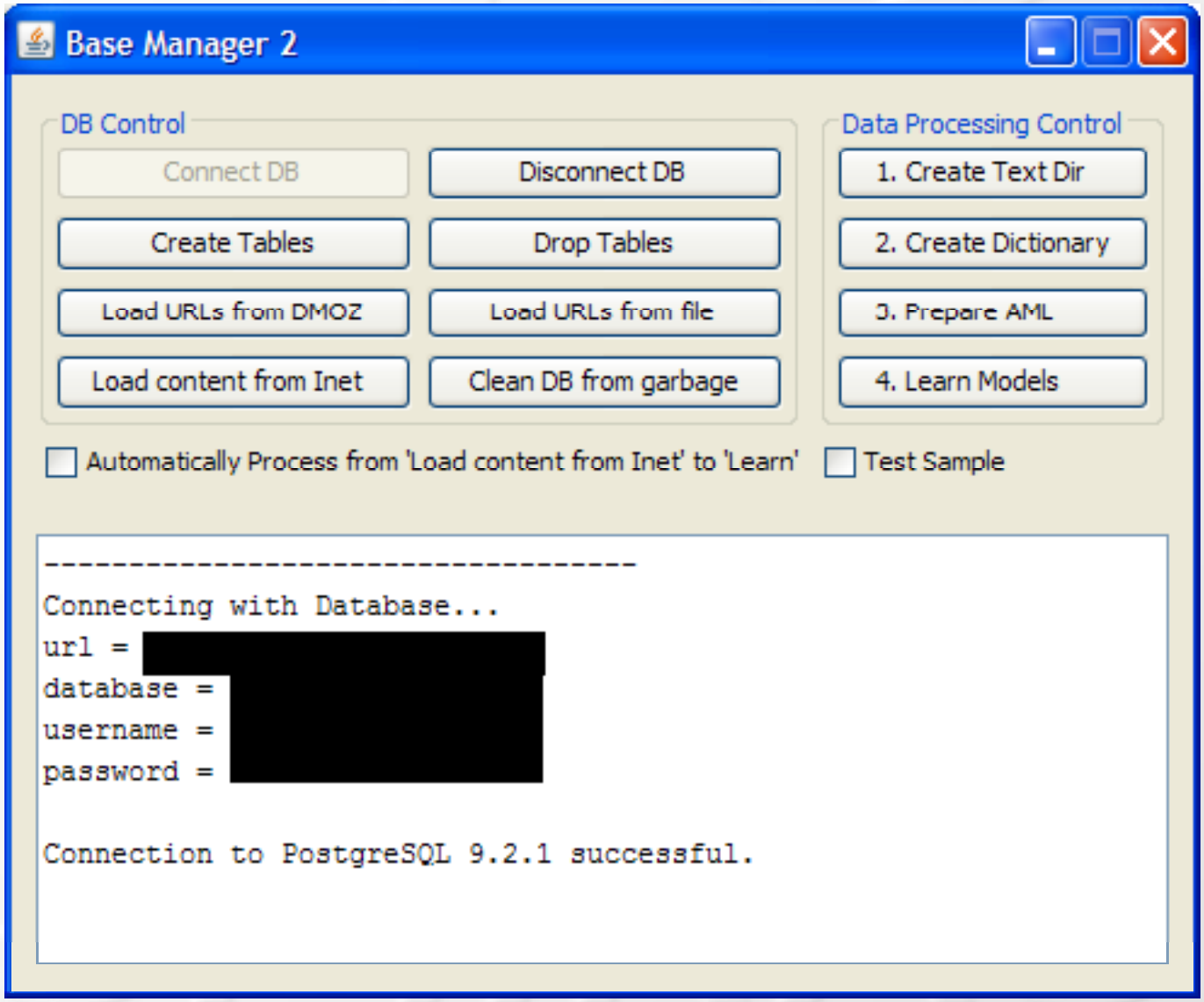
Подготовка обученной модели (2/2)

Для загрузки данных и обучения модели разработан инструмент Base Manager v2.

Данный инструмент работает в двух режимах:

1. Режим обучения. Данный режим предназначен для автоматизированной подготовки обученной модели. Для этого требуется список URL сайтов, принадлежащих к той или иной категории.
2. Режим тестирования. Позволяет оценить качество обученной модели. Также требует список URL сайтов с категориями.

Интерфейс Base Manager 2

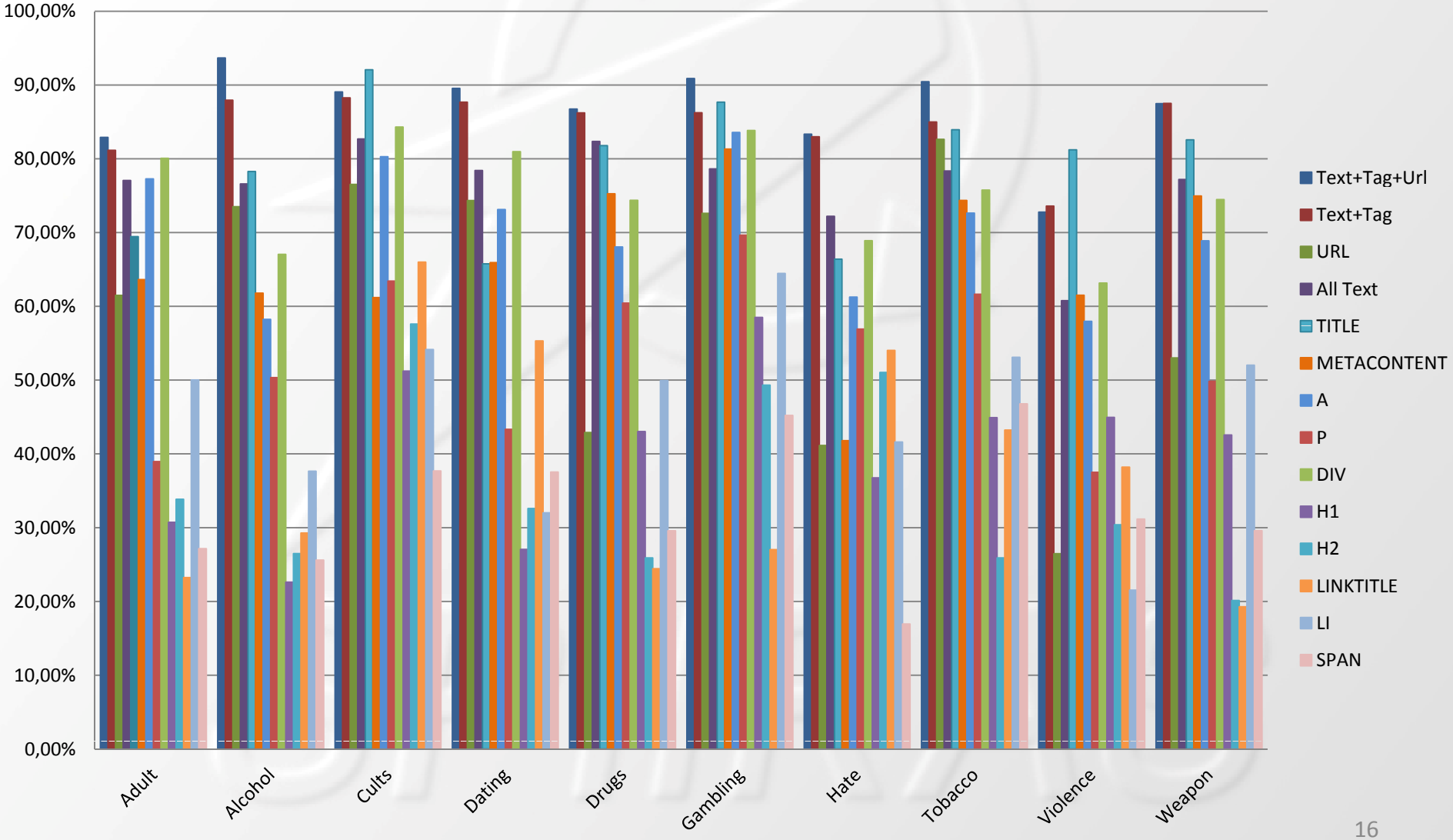


Эксперименты (1/2)

- Для проведения экспериментов были выбраны 10 основных категорий:
  - Adult, Alcohol, Cults, Dating, Drugs, Gambling, Hate, Tobacco, Violence, Weapon.
- Для уменьшения количества ошибок второго рода использовался набор данных, содержащий легитимные сайты, и отнесенный к категории Unknown.
- Если классификатор не мог определить сайт к одной из категорий, он относил ее к категории Unknown.
- Тестирование проводилось на выборке состоящей из 7697 сайтов.

Эксперименты (2/2)

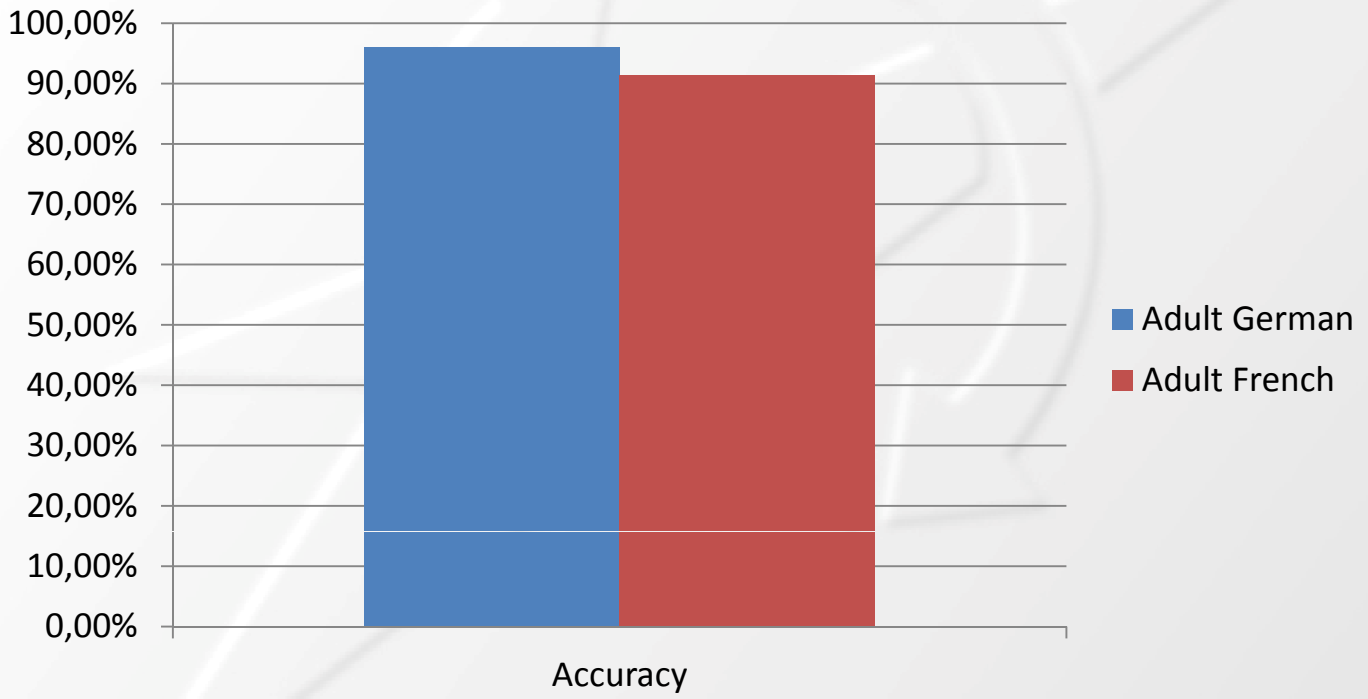
Данная диаграмма демонстрирует показатель F-меры, полученный после категоризации с помощью различных типов обученных моделей (см. легенду).



Решение языковой проблемы (1/2)

- Для классификации сайтов на иностранных языках использовалась трансляция контента web-страницы с языка оригинала на язык, который использовался для подготовки обученной модели.
- После чего использование обученного классификатора позволяло определять принадлежность сайта к той или иной категории.
- Для верификации данного подхода использовалась тестовая выборка состоящая из веб-сайтов содержащие 4191 и 3205, которые принадлежат к категориям Adult на немецком и французском языках.

Решение языковой проблемы (2/2)



- Результаты показали, что классификатор подготовленный на данных, не включавшие эти категории смог отнести 96,11 % и 91,39% к категориям adult на немецком и французском языках соответственно.
- Для классификации сайтов на иностранных языках в текущей версии программы **Classifier** встроена система автоматического перевода с использованием API Яндекс.Перевод.

Заключение

- Данный доклад представляет технологию категоризации web-сайтов с помощью методов Machine Learning и Data Mining;
- Представлены технология и инструменты, позволяющие автоматизировать процесс подготовки исходных данных и обученной модели ;
- Для категоризации веб-страниц на иностранных языках использовалась методика перевода веб-сайтов с языка оригинала на язык, использованный для обучения моделей;
- Эксперименты показали высокую точность категоризации веб-сайтов, что дает возможность использования разработанной технологии в системах блокирования веб-сайтов с неприемлемым содержимым.

Используемое ПО

- Для сбора и первичной обработки данных
  - Jsoup 1.7.1 (http://jsoup.org);
  - NetBeans IDE 7.2 (http://netbeans.org/);
  - Яндекс.Перевод (http://translate.yandex.ru/);
- Для хранения данных
  - Postgresql 9.2.1 (http://www.postgresql.org/);
  - pgAdmin 1.16.0 (http://www.pgadmin.org/);
- Для проведения экспериментов
  - RapidMiner 5.2 (http://rapid-i.com/).

Контактная информация

Комашинский Дмитрий Владимирович
komashinskiy@comsec.spb.ru
http://comsec.spb.ru/Komashinskiy
Котенко Игорь Витальевич
ivkote@comsec.spb.ru
http://comsec.spb.ru/Kotenko
Чечулин Андрей Алексеевич
chechulin@comsec.spb.ru
http://comsec.spb.ru/Chechulin
Шоров Андрей Владимирович
ashorov@comsec.spb.ru
http://comsec.spb.ru/shorov

Благодарности

Работа выполняется при финансовой поддержке Министерства образования и науки РФ (государственный контракт 11.519.11.4008), РФФИ, программы фундаментальных исследований ОНИТ РАН, проектов Евросоюза SecFutur и MASSIF