СПИИРАН

КАТЕГОРИРОВАНИЕ ВЕБ-СТРАНИЦ С НЕПРИЕМЛЕМЫМ СОДЕРЖИМЫМ

Комашинский Д.В.,
Чечулин А.А., Котенко И.В.

Учреждение Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН

РусКрипто'2011, 30 марта – 2 апреля 2011 г.

Содержание

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

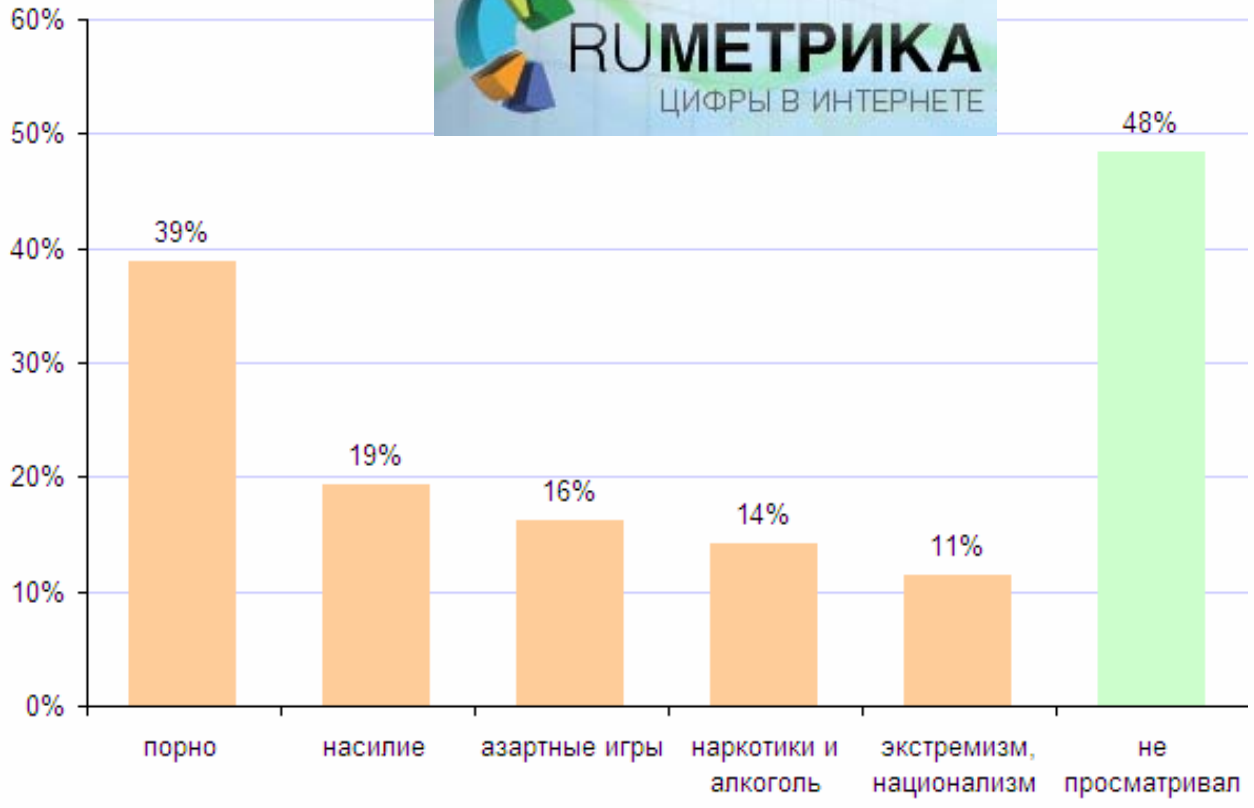
Неприемлемые сайты

- Сайты, запрещенные законодательством РФ
  - принадлежащие тоталитарным и деструктивным религиозным сектам;
  - посвященные изготовлению и применению психотропных препаратов и наркотиков;
  - предлагающие взломанное программное обеспечение;
- Мошеннические сайты
  - Веб-страницы, имитирующие сайты банков, электронных магазинов и т.д.

Неприемлемые сайты для детей

- Сайты для взрослых
  - содержащие материалы порнографического и эротического характера;
  - посвященные азартным играм;
  - сайты знакомств;
- Динамические сайты
  - социальные сети;
  - блоги;
  - чаты;

Ключевые показатели



- В России около 9 млн. Интернет-пользователей в возрасте до 14 лет;
- 75% юных интернет-пользователей выходят в сеть самостоятельно;
- 88% четырёхлетних выходят в сеть вместе с родителями, к 14 годам совместное пользование сетью сохраняется лишь для 7% подростков

Общая характеристика работы

- Цель работы:
  - Разработка архитектуры системы определения категории веб-страниц для блокировки сайтов с неприемлемым содержимым;
- Задачи:
  - Анализ существующих моделей и методов определения категории веб-страниц;
  - Разработка архитектуры системы;
  - Проведение экспериментов для проверки разработанной архитектуры;

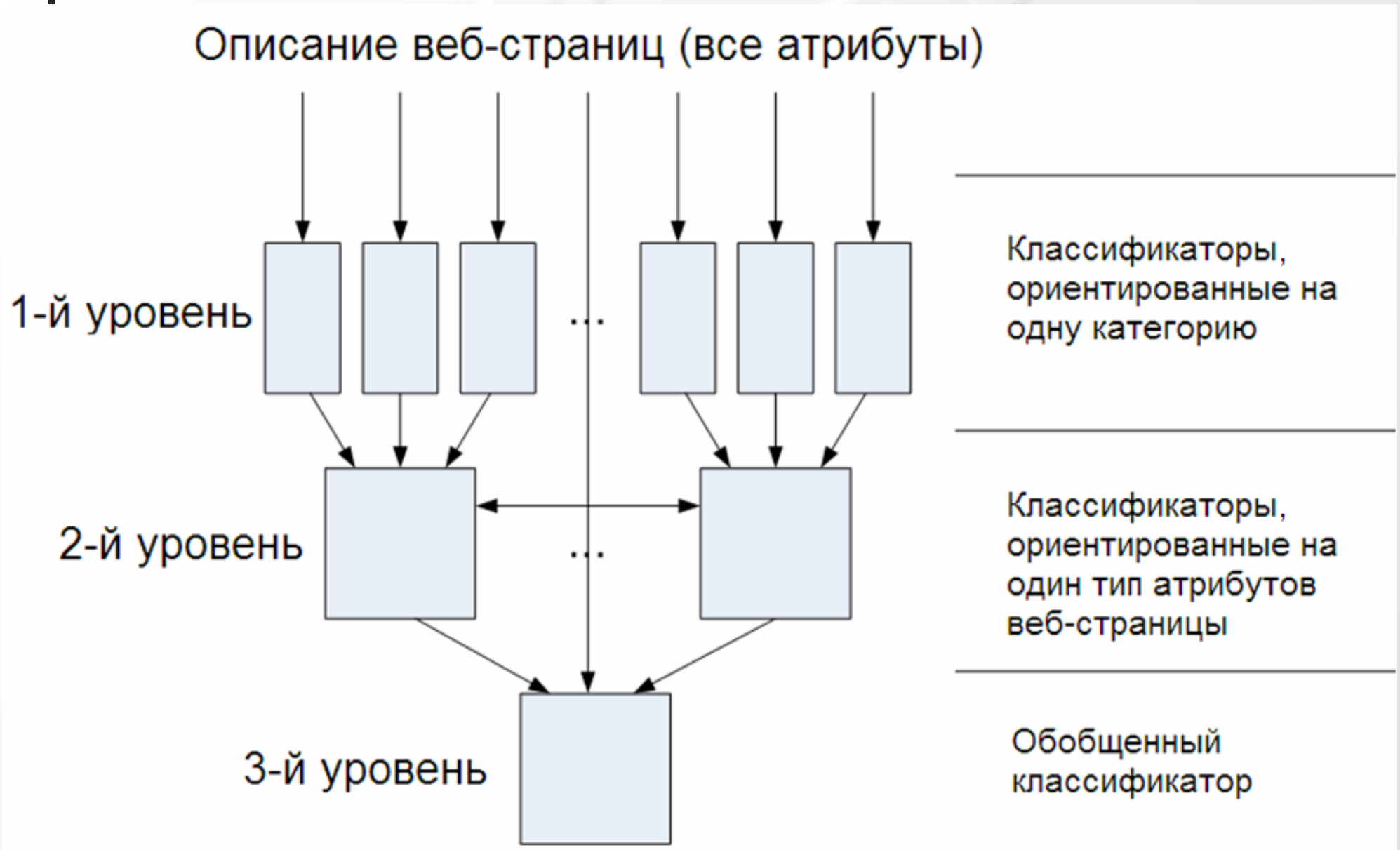
Релевантные работы (1/2)

- Описание типов текстовых исходных данных
  - Кузнецов Р.Ф. Классификатор веб-страниц на базе SVM-Multiclass // Труды РОМИП'2006. http://romip.narod.ru/romip2006/10_kuznecov.pdf
- Общее описание методик анализа данных
  - Han J., Kamber M. Data Mining: Concepts and Techniques // Elsevier, Morgan Kaufman publishers, 2006.
- Общее описание подхода к блокировке сайтов с неприемлемым содержимым
  - Зозуля Ю.В., Котенко И.В. Блокирование Web-сайтов с неприемлемым содержимым на основании выявления их категорий // РусКрипто'2010

Релевантные работы (2/2)

- Подходы к классификации веб-страниц.
  - Qi, X., Davison, B.D (2009). Web Page Classification: Features and algorithms, ACM Computing Surveys (CSUR). 2009. http://www.eecs.ucf.edu/~dcm/Teaching/COT4810-Spring2011/Literature/WebPageClassification.pdf
  - 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.

Общая архитектура системы



Используемые данные

- URL;
- HTML;
  - Текст;
  - Статистика встречаемости тегов;
  - Текст из определенных тегов (H1,..,H6, META, …);
- Данные из внешних источников;
  - Ответы WhoIs серверов;
  - Существующие списки категоризированных сайтов;
  - История категоризаций.

Используемое ПО

- Для сбора и первичной обработки данных
  - Cobra HTML Parser 0.98.4 (http://lobobrowser.org/cobra/java-html-parser.jsp);
  - NetBeans IDE 6.8 (http://netbeans.org/);
- Для хранения данных
  - Postgresql 8.4 (http://www.postgresql.org/);
  - pgAdmin 1.10.2 (http://www.pgadmin.org/);
- Для проведения экспериментов
  - RapidMiner 5.0 (http://rapid-i.com/);
  - Amazon Web Services (http://aws.amazon.com/).

Источники тестовых и обучающих выборок

- Примеры списков категоризированных веб-страниц
  - Open Directory RDF Dump (DMOZ) (http://rdf.dmoz.org/);
  - Shalla's Blacklists (http://www.shallalist.de/);
  - URL blacklist (http://urlblacklist.com/);
- Загружено около 900 тысяч URL по 23 категориям.
- Из них загружен и обработан контент около 70 тысяч сайтов;

Особенности загрузки данных

- Пересечение категорий
  - В категориях Shopping и Phishing оказалось около 400 общих веб-страниц, в категориях Phishing и Gambling – около 300;
- Количество веб-страниц, выдавших ошибку при загрузке
  - Больше всего: Phishing: 89%, Gambling: 58%.
  - Меньше всего: Travel: 10%, Health: 13%;

Особенности загрузки данных

- Суммарный объем текстов
  - Больше всего: Blogs: около 200 Мб, Forum: около 140 Мб;
  - Меньше всего: Warez: около 40 Мб;
- Страницы меньше 500 байт
  - Больше всего: Phishing: около 900 страниц;
  - Меньше всего: Games: около 300;

Особенности загрузки данных

- Суммарный объем текстов
  - Больше всего: Blogs: около 200 Мб, Forum: около 140 Мб;
  - Меньше всего: Warez: около 40 Мб;
- Страницы меньше 500 байт
  - Больше всего: Phishing: около 900 страниц;
  - Меньше всего: Games: около 300;

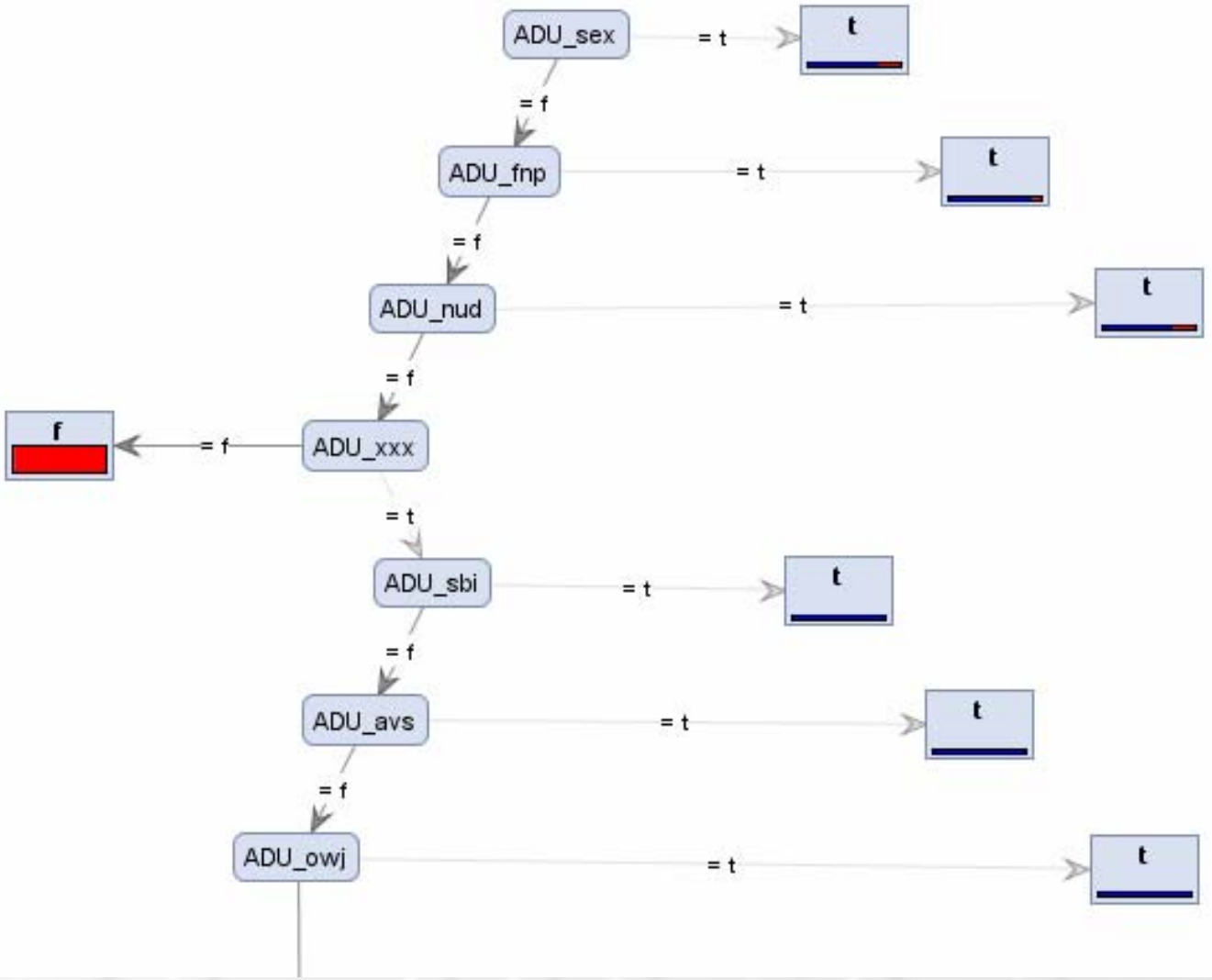
Метрики оценки качества классификации

- 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 (r)* = $\frac{a}{a+c}$; *Точность (r)* = $\frac{a}{a+b}$;
  - *F-мера (Fmeasure)* = $\frac{2pr}{p+r}$

Эксперименты по URL данным. Результаты.

- Около 40% категорий не могут быть достаточно хорошо классифицированы (например, Phishing, Shopping в силу своей разнородности; Anonymizers, Social Networks в силу малого количества данных для обучения; сходные по тематике категории, как Forum – Chat, и т.д.);
- Применение комбинированный схемы классификации, объединяющей базовые классификаторы, дает улучшение как в точности, так и в полноте практически по всем категориям и составляет 2-3%.

Эксперименты по URL данным. Пример фрагмента дерева решений



Эксперименты по URL данным. Особенности

- Результаты экспериментов показывают, что наиболее сильной связью обладают категории Banking и Phishing – по количеству ошибок между этими категориями;
- Некоторые категории нуждаются в дополнительном семантическом разделении – например Shopping, Chats и т.д. Как правило, содержимое URL не несет достаточной смысловой нагрузки для их разделения.

Эксперименты по текстовым данным. Словари

- На основе показателей TF и IDF и на основе модификаций этих показателей;
- Примеры словарей:
  - Warez: xvid, torrent, crack, german, dvdrip, serial, dvd, film, warez, keygen, …;
  - Phishing: domain, traffic, firefox, influenc, nokia, qip, infium, symbian, oneworld, zoomumba, …

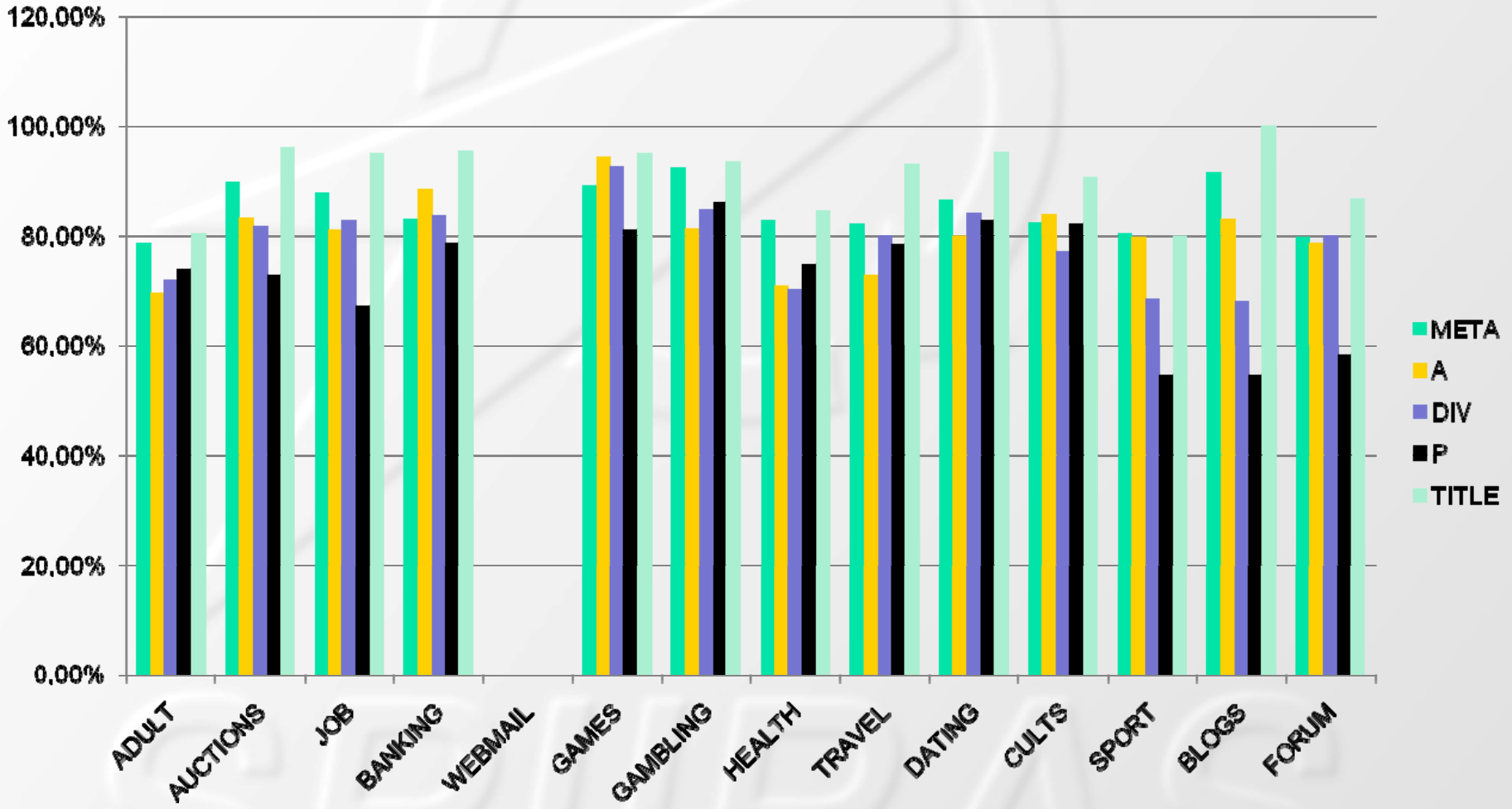
Эксперименты по текстовым данным. Результаты

- Классификатор Naïve Bayes. Аккуратность: 49,04%; Ошибки: 50,96%
  - Лучше всего классифицируются: Gambling, Banking и Auctions с F-мерой около 0,67;
  - Хуже всего классифицируются: Phishing, Chat и Social_Networking с F-мерой около 0,18;
- Классификатор Decision Tree. Аккуратность: 38,86%; Unknown: 47,36%; Ошибки: 13,78%.
  - Лучше всего классифицируются: Gambling, Games и Auctions с F-мерой от 0,65 до 0,72;
  - Хуже все классифицируются: Shopping, Phishing и Chat с F-мерой от 0,08 до 0,11.

Эксперименты по тегам. Результаты.

- В отличие от классификации по общему тексту, результаты в среднем лучше и по точности и по полноте;
- Существуют свои проблемные категории, например WebMail. Проблемы в данном случае связаны с достаточно размытой дефиницией категории и относительно небольшим количеством примеров, относящихся к ней.

Эксперименты по тегам. Точность базовых классификаторов



Эксперименты по тегам. Особенности.

- Две группы категорий:
  - Первая категория направлена на агрессивное привлечение пользователей, и, как следствие, имеет относительно высокую точность результатов классификации по тегам, традиционно используемым для формирования краткого смыслового описания ресурса (например теги TITLE, META). Примеры: Gambling, Travel.
  - Вторая категория ориентирована в большей степени на тематическое информационное обеспечение пользователя и, как следствие, демонстрирует повышенную точность по «контентным» и «ссылочным» тегам (например, A, DIV и т.д.). Примеры: Games, Banking.

Заключение

- Предложен иерархический подход к категоризации веб-странц;
- Собраны исходные данные и выделены основные атрибуты по которым, может проводиться категоризация;
- Проведена серия экспериментов.

Дальнейшие исследования

- Объединение аспектных классификаторов в единую схему;
- Расширение списка категорий;
- Добавление новых типов классификаторов;
- Добавление новых типов исходных данных;
- Проведение новых серий экспериментов.

Контактная информация

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

Благодарности

Работа выполняется при финансовой поддержке РФФИ (проект 10-01-00826-а), программы фундаментальных исследований ОНИТ РАН (проект 3.2) и при частичной финансовой поддержке, осуществляемой в рамках проектов Евросоюза SecFutur и MASSIF.