rookee
основной инструмент оптимизатора!

Тематическая классификация WEB-страниц и WEB-сайтов

С.В. Панков, С.П. Шебанин, А.А. Рыбаков

Немного о нас

проект & команда

- R&D подразделение компании Ingate
- Специализация – создание массовых сервисов для интернет рекламы
- ROOKEE – первый продукт, SEO

Цели исследования

- Опробовать методики предварительной обработки WEB-ресурсов
- Построить и оценить классификатор WEB-ресурсов, приемлемой точности и производительности
- Получить опыт участия в РОМИП

Задача классификации

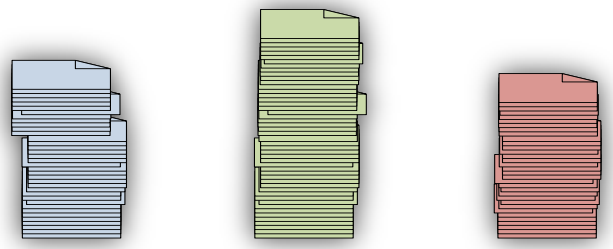
Классическая постановка

$D = \{d_1 \dots d_{|D|}\}$

множество документов

$C = \{c_1 \dots c_{|C|}\}$

множество тематик



$$\Phi: D \times C \to \{0, 1\}$$

неизвестная целевая функция

$$\Phi' : D \times C \to [0, 1]$$

Построение классификатора

Использованные методики

- АОТ для работы с морфологией (aot.ru)
- Линейный метод для обучения
- Векторное представление тематик и документов
- Косинус угла для оценки тематической близости

Но!

Есть проблемы, связанные со спецификой

Специфика задачи

Проблема

Коллекция классифицированных документов (коллекция DMOZ) зашумлена

$$\boldsymbol{\Omega} = \{\boldsymbol{d_1} \ldots \boldsymbol{d_{|\Omega|}}\} = \boldsymbol{\Omega'} \cup \boldsymbol{\Omega''}$$

Специфика задачи

Проблема

Коллекция классифицированных документов (коллекция DMOZ) зашумлена

$$\boldsymbol{\Omega} = \{\boldsymbol{d_1} \ldots \boldsymbol{d_{|\Omega|}}\} = \boldsymbol{\Omega'} \cup \boldsymbol{\Omega''}$$

Проблема 1

Наполнение WEB-страниц

Не весь текст WEB-страницы информативен

Главное
В России
В СССР
В мире
Америка
Германия
Экономика
Финансы
Бизнес
О рекламе
Недвижимость
Авто
Преступность
Масс-медиа
О высоком
Кино
Музыка
Спорт
Прогресс
Интернет
Технологии
Игры
Оружие
Медицина
Из жизни
Пентаедия
Онлайны
Письма
Пресс-релизы
@lenta.ru
RSS
поиск
в Ленте везде

08.10.2010, 12:44:58

Версия для печати | PDA/WAP

Фото AFP

Хакеры анонсировали инструмент для взлома iPhone 4

Хакеры из группы Chronic Dev Team анонсировали выход инструмента для взлома смартфонов iPhone 4 и других устройств на платформе iOS 4.1, пишет Redmond Pie. Процедура взлома ("джейлбрейка") позволяет устанавливать на смартфоны сторонний контент не из официального магазина приложений Apple App Store.

"Джейлбрейк", разработанный для последнего поколения смартфонов iPhone, плееров iPod touch, а также планшетов iPad, будет выпущен 10 октября в 10 часов 10 минут и 10 секунд по британскому времени. Отмечается, что подвергнуть "джейлбрейку" можно будет только устройства на базе процессора A4. Таким образом, он не будет работать на смартфонах и плеерах Apple предыдущих поколений.

Сообщается, что "джейлбрейк" основан на уязвимости SHAtter, выявленной Chronic Dev Team в начале сентября. При этом "джейлбрейк" будет полным и не потребует подключения устройства к компьютеру при включении аппарата после перезагрузки.

"Джейлбрейк" отличается от процедуры разблокировки SIM-карты (SIM unlock), которая позволяет лишить аппараты привязки к сети определенного мобильного оператора. Отметим, что iPhone с привязкой к оператору продаются в некоторых странах, например в США. В России iPhone 4 продается без привязки.

Ссылки по теме
- BREAKING: GreenPois0n iOS 4.1 Jailbreak is Confirmed to Release on Sunday, 10th October ! - Redmond Pie, 08.10.2010
- Найден способ взлома последней версии Apple iOS – Lenta.ru, 09.09.2010
- Хакеры продемонстрировали способ взлома iPhone 4 – Lenta.ru, 02.08.2010
- Библиотека Конгресса США разрешила "джейлбрейк" iPhone – Lenta.ru, 27.07.2010

Сайты по теме
- Apple iPhone

[Обсудить с другими читателями]

[Сообщить о найденной опечатке]
[Письмо в редакцию]
URL: http://lenta.ru/news/2010/10/08/jailbreak/

Последние новости

08.10 16:51 Кремль обсудил с единороссами возможных преемников Лужкова

08.10 16:36 У песни "Мерседес Бенц" появилось продолжение

08.10 17:03 Главная страница "Яндекса" перестала открываться

08.10 17:02 Количество подписчиков World of Warcraft превысило 12 миллионов

08.10 17:00 "Врачи без границ" объявили войну европейской фарминдустрии

08.10 16:29 Китай пригрозил Норвегии ухудшением отношений из-за премии мира

08.10 16:05 Акции Adobe резко выросли из-за слухов о сделке с Microsoft

Аутсайд

iFixit: Nokia N8 Teardown
Nokia N8 разобрали до винтиков.

Ubergizmo: Docomo Phone with integrated air purifier
В Японии показали телефон, оснащенный встроенным ионизатором воздуха.

Mozilla: Firefox 4 Beta for Android and Maemo is Now Available
Вышла первая бета-версия мобильного браузера Firefox 4 для Android.

Gdgt: With Antennagate over, is Gizmodo next for the iPhone 4?
Техноблог Gdgt рассказал о еще одном недостатке iPhone 4, выявленном инженерами Apple.

Designs: Nokia E-Cu by Patrick Hyland
Британский дизайнер создал концепт телефона, заряжающегося в кармане от тепла тела.

Технологии

08.10 13:10 ОАЭ отказались от введения запрета на BlackBerry

07.10 14:17 Motorola подала в суд на Apple из-за технологий в iPhone и iPad

07.10 17:52 Samsung представил стартовую модель смартфона на ОС bada

07.10 12:57 В Android Market вернут оплату с помощью PayPal

07.10 15:19 Sony выпустила 16-мегапиксельную камеру для телефонов

07.10 11:06 В США до конца года запустят крупнейшую в мире LTE-сеть

// Офтопик

Видеобудущее
08.10 15:31
О том, как видеозвонки постепенно выходят "в массы"

Первое касание
08.10 14:30
Обзор нового сенсорного iPod nano

// Комментарии

Двойное касание
06.10 16:22
Обзор нового iPod touch

Главное
В России
В СССР
В мире
Америка
Германия
Экономика
Финансы
Бизнес
О рекламе
Недвижимость
Авто
Преступность
Масс-медиа
О высоком
Кино
Музыка
Спорт
Прогресс
Интернет
Технологии
Игры
Оружие
Медицина
Из жизни
Пентаедия
Онлайны
Письма
Пресс-релизы
@lenta.ru
RSS

поиск
в Ленте везде

08.10.2010, 12:44:58

Версия для печати | PDA/RTK

Фото AFP

Хакеры анонсировали инструмент для взлома iPhone 4

Хакеры из группы Chronic Dev Team анонсировали выход инструмента для взлома смартфонов iPhone 4 и других устройств на платформе iOS 4.1, пишет Redmond Pie. Процедура взлома ("джейлбрейка") позволяет устанавливать на смартфоны сторонний контент не из официального магазина приложений Apple App Store.

"Джейлбрейк", разработанный для последнего поколения смартфонов iPhone, плееров iPod touch, а также планшетов iPad, будет выпущен 10 октября в 10 часов 10 минут и 10 секунд по британскому времени. Отмечается, что подвергнуть "джейлбрейку" можно будет только устройства на базе процессора A4. Таким образом, он не будет работать на смартфонах и плеерах Apple предыдущих поколений.

Сообщается, что "джейлбрейк" основан на уязвимости SHAtter, выявленной Chronic Dev Team в начале сентября. При этом "джейлбрейк" будет полным и не потребует подключения устройства к компьютеру при включении аппарата после перезагрузки.

"Джейлбрейк" отличается от процедуры разблокировки SIM-карты (SIM unlock), которая позволяет лишить аппараты привязки к сети определенного мобильного оператора. Отметим, что iPhone с привязкой к оператору продаются в некоторых странах, например в США. В России iPhone 4 продается без привязки.

Ссылки по теме
- BREAKING: GreenPois0n iOS 4.1 Jailbreak is Confirmed to Release on Sunday, 10th October ! - Redmond Pie, 08.10.2010
- Найден способ взлома последней версии Apple iOS – Lenta.ru, 09.09.2010
- Хакеры продемонстрировали способ взлома iPhone 4 – Lenta.ru, 02.08.2010
- Библиотека Конгресса США разрешила "джейлбрейк" iPhone – Lenta.ru, 27.07.2010

Сайты по теме
- Apple iPhone

[Обсудить с другими читателями]
[Сообщить о найденной опечатке]
[Письмо в редакцию]

URL: http://lenta.ru/news/2010/10/08/jailbreak/

Последние новости

08.10 16:51 Кремль обсудил с единороссами возможных преемников Лужкова

08.10 16:36 У песни "Мерседес 8666" появился продолжатель

08.10 17:03 Главная страница "Яндекса" перестала открываться

08.10 17:02 Количество подписчиков World of Warcraft превысило 12 миллионов

08.10 17:00 "Врачи без границ" объявили войну европейской фарминдустрии

08.10 16:29 Китай приглушил Норвегии упущением отношений из-за премии мира

08.10 16:05 Акции Adobe резко выросли из-за слухов о сговоре с Microsoft

Аутсайд

iPad: Nokia N8 Teardown
Nokia N8 разобрали до винтиков.

Ubergizmo: Docomo Phone with integrated air purifier
В Японии показали телефон, оснащенный встроенным кондиционером воздуха.

Mozilla: Firefox 4 Beta for Android and Maemo is Now Available
Вышла первая бета-версия мобильного браузера Firefox 4 для Android.

Gdgt: With Antennagate over, is Glasspaper next for the iPhone 4?
Технопортал Gdgt рассказал о еще одном недостатке iPhone 4, вычеканном инженерами Apple.

Dezeen: Nokia E-Co by Patrick Hyland
Британский дизайнер создал концепт телефона, заряжающегося в кармане от тепла тела.

Технологии

08.10 13:10 ОАЭ отказались от введения запрета на BlackBerry

07.10 14:17 Motorola подала в суд на Apple из-за технологий в iPhone и iPad

07.10 17:12 Samsung представит сенсорную модель смартфона на ОС bada

07.10 12:57 В Android Market вводят оплату с помощью PayPal

07.10 15:19 Sony выпустила 16-мегапиксельную камеру для телефонов

07.10 11:56 В США до конца года запустят крупнейшую в мире LTE-сеть

// Офтопик

Видеобудущее
08.10 15:31
О том, как видеозвонки постепенно выходят "в массы"

Первое касание
08.10 14:30
Обзор нового сенсорного iPod nano

// Комментарии

Двойное касание
06.10 16:22
Обзор нового iPod touch

rookee

Проблема 2

Разнородность тематик

На сайте присутствуют WEB-страницы, не относящиеся к тематике сайта

Проблема 2

Разнородность тематик

rookee

Проблема 2

Выделение тематического ядра

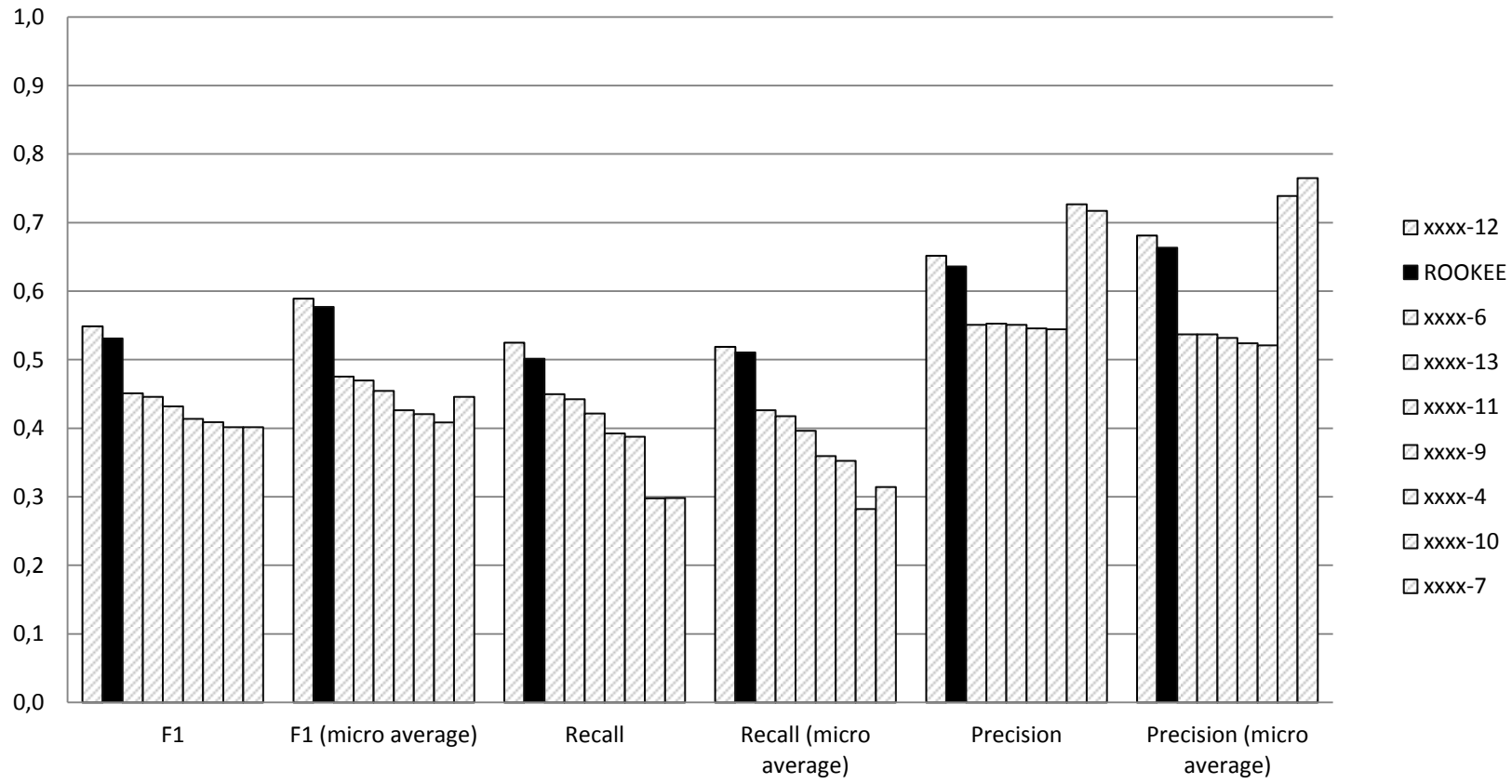
Гипотеза:

Большая часть страниц соответствует тематике сайта

Отсекаем девиантные отклонения

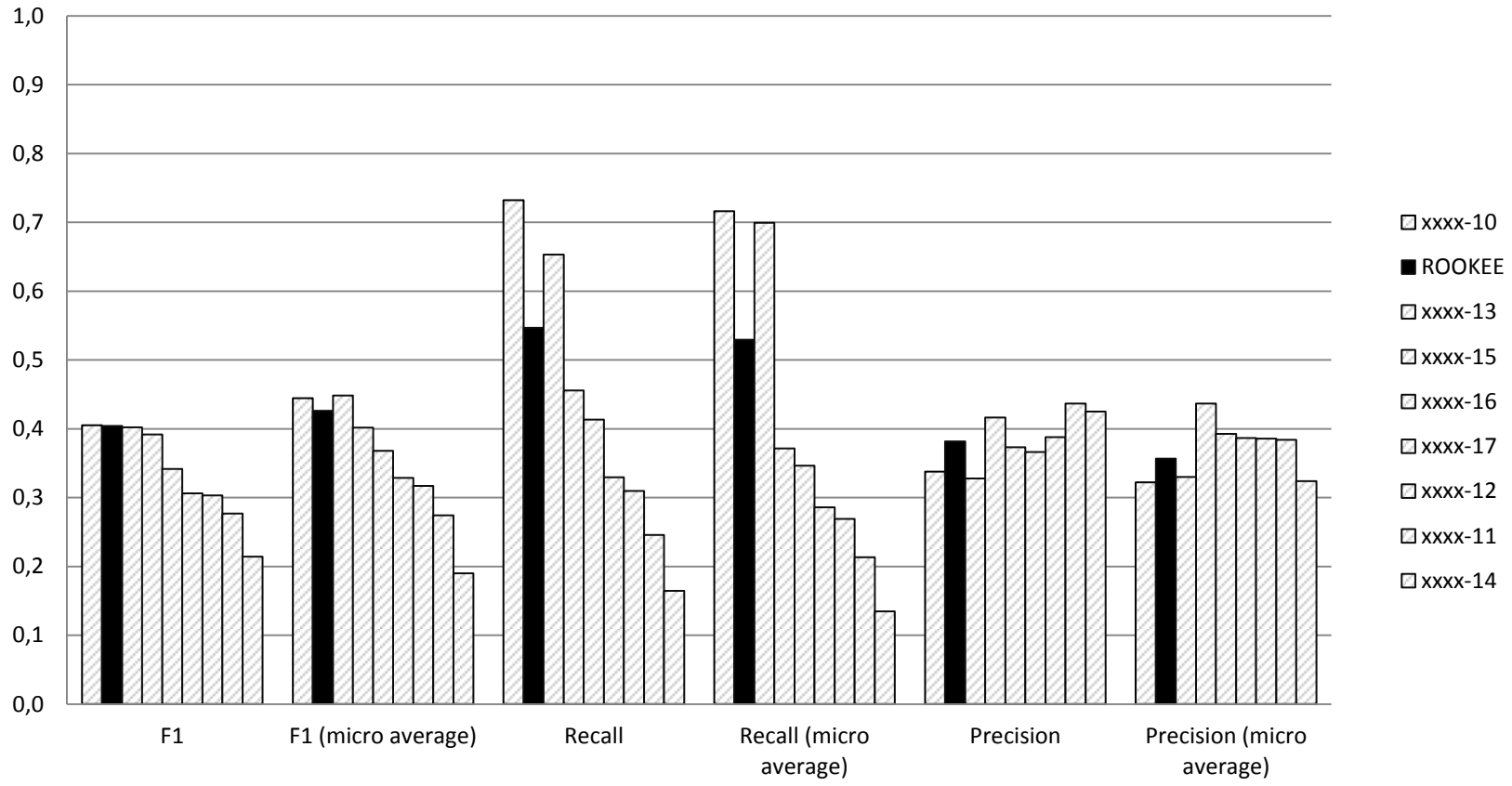
Результаты

Классификации страниц (AND)



Результаты

Классификации сайтов (AND)



Результаты

Выводы

- Использованная методика позволяет получать неплохие результаты рубрикации
- Методика выделения значимой информации на странице может получить широкое применение

rookee

Перспективы

Что еще предстоит сделать?

- Учитывать позицию текста на странице (заголовки, html-теги и т.д.)
- Корректно обрабатывать случаи мультитематичных сайтов
- Протестировать использование прочих способов для оценки тематической близости
- …

Вопросы

Спасибо за внимание!

Вопросы?