Влияние представления веб-сайта на качество классификации

Кондратьев Михаил
Mikhail.Kondrayev@sun.com

Классификация веб-сайтов

Автоматическая классификация

- построение профиля сайта обучающей коллекции
- построение профиля сайта рабочей коллекции
- обучение классификатора
- классификация профилей

Как строить профили ?

- какое количество информации необходимо?
- каков алгоритм создания профиля?

Экспериментальное окружение

Классификатор

- основан на пакете Rainbow
- простой байесовский классификатор
- отсутствие поддержки русского языка

Предварительные эксперименты

- использовалась коллекция РОМИП’2003
- метрики вычислялись по категориям, оцененным асессорами

Типы профилей

Стартовая страница сайта

- идея: вся основная информация доступна со стартовой страницы

Стартовая страница и ее окрестность

- идея: страницы, на которые ссылается главная, должны раскрывать основные темы, упомянутые на стартовой странице

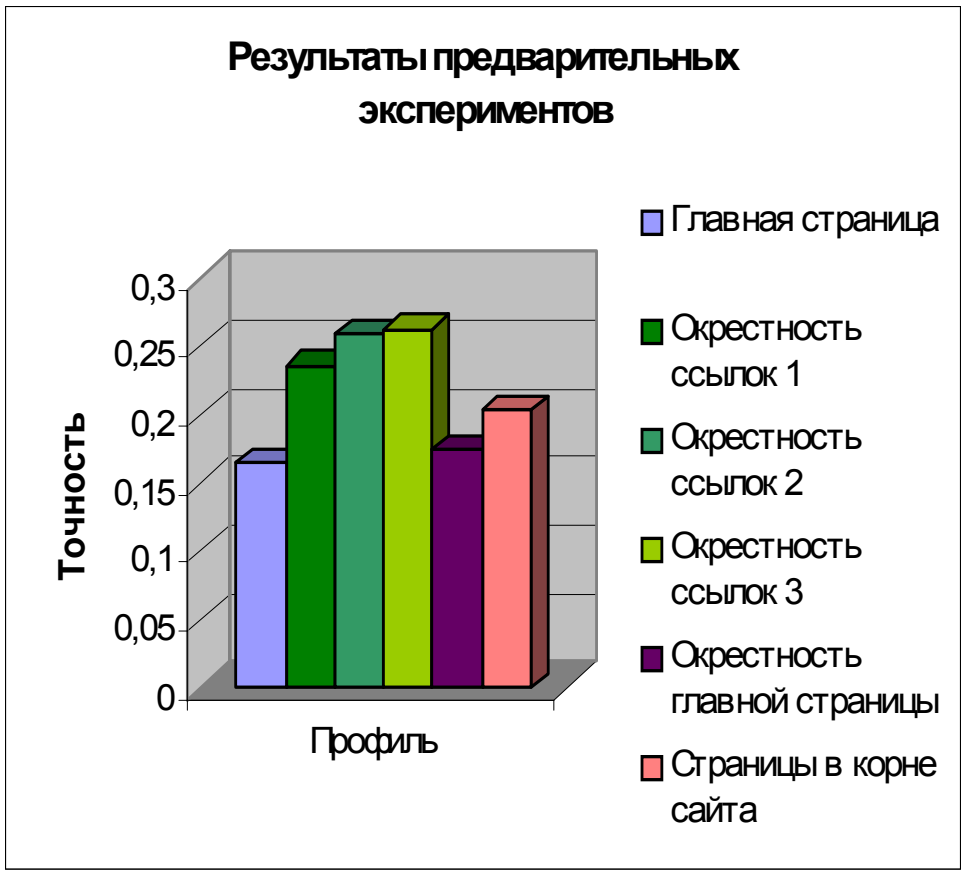
Страницы в корне сайта

- идея: страницы, посвященные основной тематике, находятся в одном каталоге – в корне сайта

Текст ссылок и текст в окрестности ссылок

- идея: ссылка должна содержать необходимую информацию о странице, на которую она ссылается

Результаты предварительных опытов



Точность классификации

- наилучший результат – окрестность ссылок
- худший результат– главная страница

Размер профиля

- наименьший – окрестность ссылок
- наибольший – страницы в корне

Предварительная оценка профилей

Стартовая страница

- отсутствие текстовой информации
- минимальная текстовая информация

Стартовая страница и ее окрестность

- зашумленность данных
- б**о**льшее время построения

Страницы в корне сайта

- зашумленность данных
- большой объем профиля

Текст ссылок и текст окрестности ссылок

- малый объем профилей
- навигационная информация

Участие в РОМИП’2004

Использованные профили:

- стартовая страница и ее окрестность (глубина 1)
- текст ссылок и их окрестностей
  (глубина 3, текстовая окрестность радиусом 10 слов)

Выполненные прогоны:

	обучающая коллекция	**рабочая коллекция**	F1
A	окрестность ссылки	окрестность ссылки	0.23
B	окрестность страницы	окрестность ссылки	0.25
C	окрестность страницы	окрестность страницы	0.19

Результаты РОМИП’2004

Результаты примерно в 2 раза хуже чемпиона

- оценка OR по $F1_{\mathbf{macro}}$: 0.16 < **0.25 <** 0.3 < 0.51
- в 9 категориях не найдено ни одного правильного ответа (наилучший результат– 0, наихудший – 20)

В ряде категорий получены наилучшие результаты

- 6 категорий в прогоне B по метрике F1

Время выполнения прогонов достаточно мало

- потребовалось 8 часов на все прогоны

Результаты РОМИП’2004

Прогоны A и B

- результаты приблизительно равны, существует небольшое превосходство прогона B
- прогон B показал лучшие значения полноты (28 категорий)
- прогон A показал лучшие значения точности (33 категории)
- оьъем профилей на основе текста ссылок на порядок меньше

Прогон C

- результаты прогона стабильно уступают другим прогонам (29 категорий для прогона B по F1)
- классификация заняла наибольшее время