Источники данных в задачах классификации запросов

Хоруженко Марина

Rambler

Обзор доклада

- Что такое классификация запросов?
- Типы источников данных
- Классификация на примере цитатных запросов
- Классификация на примере навигационных запросов

Rambler

Что такое классификация?

- Объединяем в классы запросы, которые имеют определённые признаки. Признаками может быть что угодно:
  - тема
  - типы
  - кластеризация по сессиям
  - частотность
  - длина
  - и т.п.

- Запросы разбиваются на классы ради чего-то. Иногда не имеет смысла создавать универсальную модель «ради науки», а следует решать конкретные задачи.

Rambler

Источники данных

- Сами запросы

  «*Мы все учились понемногу чему-нибудь и как-нибудь*» - интуитивно подозреваем, что это цитата. Даже если бы мы не знали этого заведомо.

- Внешние данные

  «*пижама всем*» - не зная, что есть такой сайт, трудно представить, что это навигационный запрос. Источники этого знания находятся вне запроса.

Rambler

Цитатные запросы: обзор

Попробуем классифицировать запросы без использования внешних знаний

- Определяем, что такое для нас цитата
- Создаём модель:
  - придумываем гипотезы-признаки
  - используем machine learning
  - убираем неэффективные гипотезы
- Оцениваем результаты

Rambler

Цитаты: придумываем гипотезы

Созерцаем:

каравай-каравай кого хочешь выбирай
Не уходи из сна моего. Сейчас ты так хорошо улыбаешься,
эй моряк ты слишком долго плавал
изгиб гитары желтой ты обнимаешь нежно
в поте лица твоего будешь есть хлеб свой
я знаю я буду лететь безумной вспышкой
и снова вижу где-то там вдали, летят с печальным криком журавли
теряю контроль над собой, когда ты улыбаешься Скажи мне, что это всё не сон!
Ты мне обязательно должна рассказать, как твоим родителям удалось сделать тебя такой прекрасной. Я тоже хочу попробовать. - Закрой глаза... ой нет открой-открой. Без них темно
ты лети лети лепесток через запад на восток через север через юг ты возвращайся сделав круг
люблю тебя как ангел бога, как любит розу соловей,как мать детя родного любит,а я тебя еще сильней.
Зачем его любить – не знаю, Он не преступник, но и не святой, Плохое в нем я вижу и воспринимаю, Но хочется пожить хоть миг мечтой…
"Ну да! Тебя Чалый сбросит!» – сказала она пренебрежительно
Завтра я еще не умру, но кто его знает
Ты покорил меня и я преклоняюсь. Но со мной ты убил и искусство, принадлежавшее всему миру

Цитаты: придумываем гипотезы

- Длина запроса
- Наличие знаков препинания
  - абсолютное количество
  - наличие конкретных знаков препинания (например, троеточие) и их количество
- Наличие личных местоимений
- Наличие глаголов с определенными морфологическими признаками (например, только финитные формы) и их количество
- Наличие определённой лексики: например, вводные слова
- «Минус» лексический признак: вряд ли цитаты содержат слова «*порно*», «*bmw*» и т.п.
- Запрос начинается с большой буквы
- Наличие повторяющихся слов
- …..

Rambler

Цитаты: обучаем

Можно посмотреть на информативность каждого признака

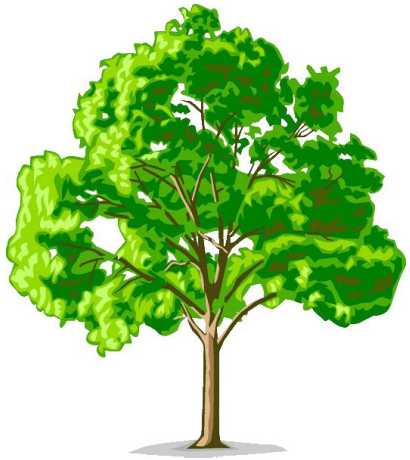
Повторяющиеся слова	test "repeated_words": information gain 0.117598 False: 470 quotes out of 96358 True: 40 quotes out of 445 повторяющихся слов нет в 96358 запросах, из них 470 - цитаты повторяющиеся слова есть в 445 запросах, 40 из них - цитаты
Местоимения	test "proname": information gain 0.286528 False: 469 quotes out of 50661 True: 41 quotes out of 46142 местоимений нет в 50661 запросах, 469 - цитаты местоимения есть в 46142 запросах,41 - цитаты
Троеточие	test "three_dots": information gain 0.177429 False: 478 quotes out of 96730 True: 32 quotes out of 73 Троеточие нет в 96730 запросах, 478 - цитаты Троеточие есть в 73 запросах, 32 - цитаты
Капитализация запроса	test "captal_letter": information gain 0.030874 False: 151 quotes out of 69515 True: 86 quotes out of 16714

Rambler

Цитаты: обучаем

```
switch(proname)
case 0 :
    switch(noninf_verb)
    case 0 :
            switch(word_number)
            case 2 :return 0.000000;
            case 3 :switch(punct_all)
                        case 0 : switch(repeated_words)
                                    case 0 : return 0.000042;
                                    case 1 : return 0.018519;
                        case 1 : return 0.002685;
                        case 2 : return 0.025974;
            case 4 : switch(black_dict)
                        case 0 : switch(punct_all_f)
                                    case 0 : return 0.000771;
                                    case 1 : ……
                                    case 2 : return 0.000000;
                        case 1 : return 0.000000;
            case 5 : switch(black_dict)
                        case 0 :  switch(three_dots)
                                    {
                                    case 0 :
                                 case 1 : return 0.000000;
            case 6 :
    case 1 : ……
    case 2 : ……
    case 3 : return 0.269231;
    case 4 : return 0.368421;

case 1 :
    switch(black_dict)
    case 0 : ……
    case 1 : ……
```



Rambler

Цитаты: итоговые признаки

- Есть ли в запросе личные местоимения
- Число слов запроса (2, 3, 4, 5, 6 и больше), не считаем союзы и предлоги
- Число знаков препинания в запросе (0, 1, 2 и больше)
- Число финитных глаголов (0, 1, 2, 3, 4 и больше)
- Есть ли в запросе троеточие
- Есть ли в запросе слова из словаря, понижающие вероятность цитаты
- Есть ли повторяющиеся слова

Rambler

Цитаты: оцениваем результаты

Порог	Точность	Полнота	F-мера
0.01	16.14	89.04	27.33
0.05	33.03	85.38	47.64
0.1	43.21	78.53	55.75
0.15	57.40	70.77	63.39
0.2	58.93	70.77	64.31
0.25	61.11	70.31	65.39
0.3	64.13	69.40	66.66
0.4	70.52	61.18	65.52
0.5	71.97	59.81	65.33
0.6	77.49	56.62	65.43
0.7	80.13	53.42	64.10

Rambler

Навигационные запросы: обзор

- Проблемы
- Традиционные источники информации
- Навигационные запросы для suggest
- Создаём модель:
  - признак click entropy
  - лексические признаки запроса
  - структурные признаки подобранного url
  - использование переформулировок
- Оцениваем результаты

Rambler

Навигационные запросы: проблемы

- Навигационные запросы могут иметь видимые признаки:
  - url-like запросы: www.rambler.ru
  - специфическая лексика *официальный сайт МВД* и т.п.

- Однако большинство навигационных запросов таковыми признаками не обладают
  видеогурман - www.videogurman.ru
  жалуйтесь - jaluites.ru
  иван царевич - www.ivan-tzarevich.ru
  иди сюда - www.idisuda.ru
  кто если не я - ktoeslineya.ru

хотим уметь прогнозировать, какой сайт соответствует запросу

Rambler

Навигационные запросы: традиционные подходы

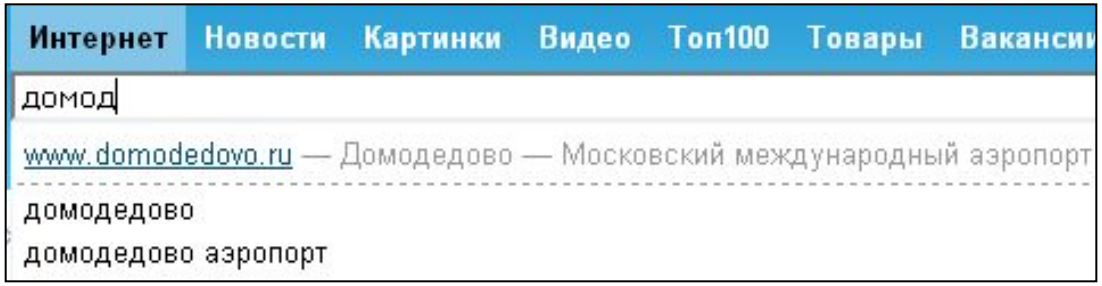
- Источник информации – клики. Например, click distribution
- Тексты ссылок: anchorlink distribution
- Признаки запроса
  - структурные
  - лексические
  - близость запроса к какому-либо существующему урлу

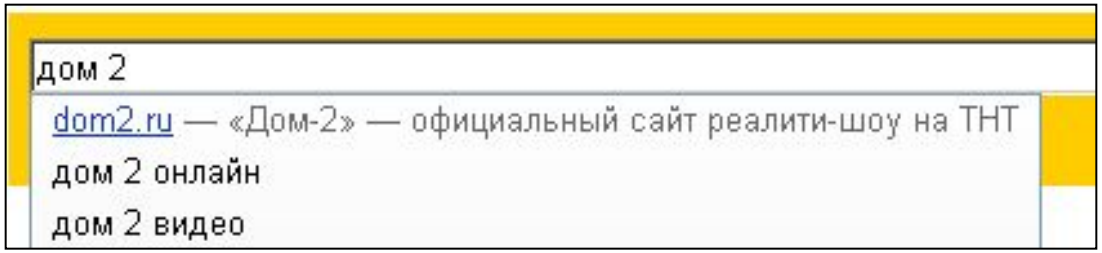
Для данной задачи хорошо подключить еще одни внешние данные: знания о переформулировках повышает полноту и точность

Слишком сложная модель! Упрощаем…

Rambler

Навигационные запросы:suggest

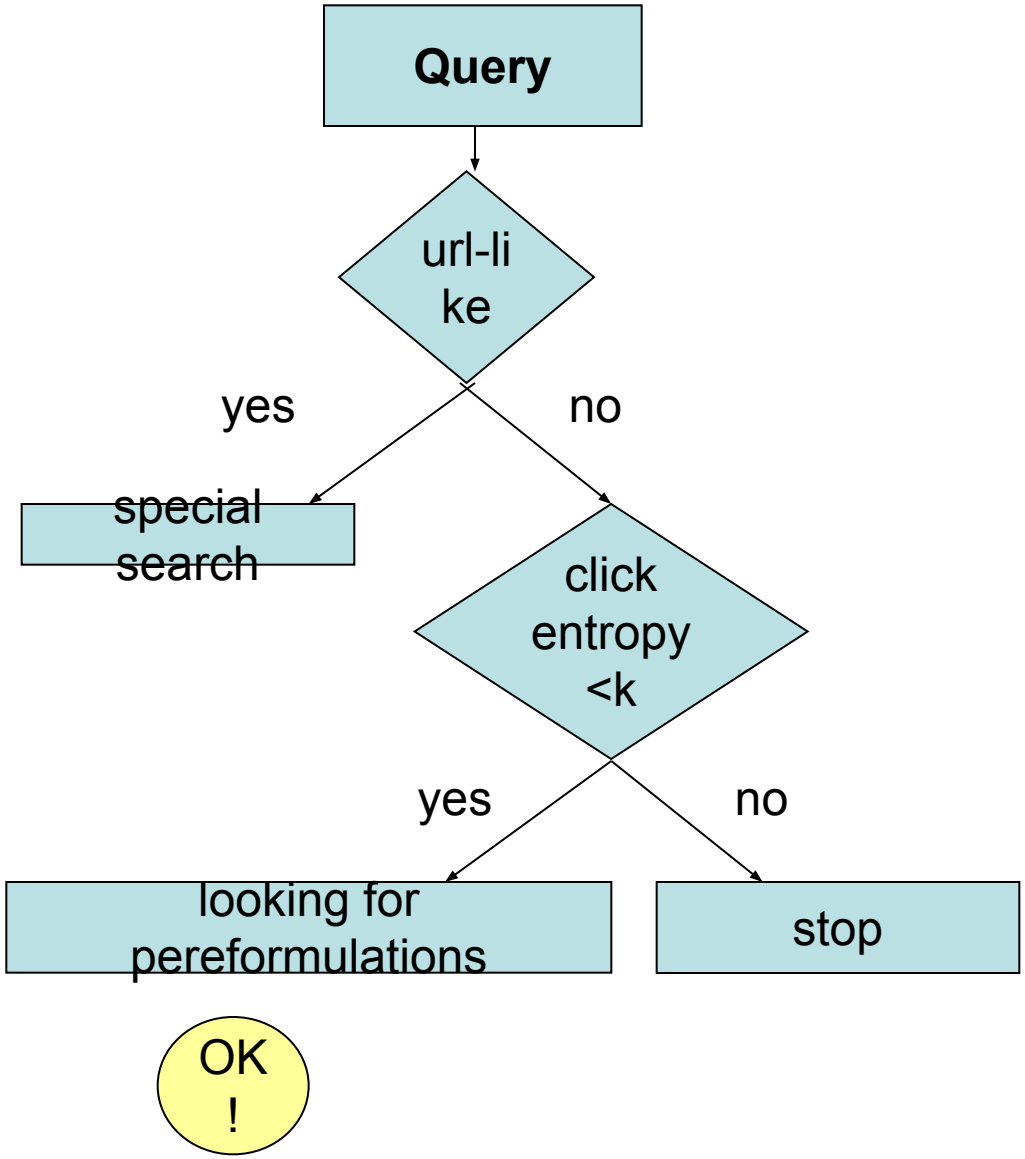




Важна точность!

Rambler

Навигационные запросы: модель



+ query structure
+ query lexical features
+ url structure

Rambler

Навигационные запросы: click entropy

- Кликовые признаки сильно зависят от качества поисковика. Если нужный результат поиска не попадает в «зону видимости» - то у нас нет статистики по кликам

- Хорошие результаты поиска могут «размывать» данные

```
банк русский стандарт | 3.2 | rs.ru
банк русский стандарт | rs.ru | 119
банк русский стандарт | banki.ru | 68
банк русский стандарт | rustinfo.ru | 39
банк русский стандарт | finlease.ru | 18
```

- Часто запросы ведут себя как «навигационные», таковыми не являясь. Частотный случай – запросы по Википедии: *шовинизм википедия, президенты сша список*

- Спам маскируется под нормальные ресурсы: *зайцев нет* - zajtsev.net

Rambler

Навигационные запросы: click entropy + lexical and url_features

click entropy даёт примерно 70% точности – мало!
Добавляем дополнительные признаки к парам <query,url>

- Лексические признаки запроса: слова «сайт», «магазин» и т.п
- Близость url и query: пижама всем ->***pijamavsem****.ru*
- Признаки подобранного урла в паре <query,url>:
  - наличие под-домена
  - длина пути
  - есть ли в урле get-параметры
  - и другие

Rambler

Навигационные запросы: расширяем переформулировками

По пользовательским сессиям объединяем запросы в кластеры, которые

1. Содержат query в качестве запроса, по которому был клик
2. Содержат запросы, которые были вместе с query в n-количестве сессий
3. Имеют общие слова с query

В такие кластеры могли попасть и такие запросы
query = **погода** гисметео - **погода** в москве (общее слово погода)

Проводим фильтрацию!

Rambler

Навигационные запросы: переформулировки + фильтрация

Входные данные:

<query, $pereform_1$, $pereform_2$…$pereform_k$, url>

- число таких переформулировок
- общие слова (минус география)
- среднее число общих слов (чем больше, тем лучше)

макс 2009 официальный сайт - официальный сайт макс 2009

- энтропия по url для запросов с такими общими словами
  Например, большая энтропия по url у слова «зао», т.е. часто является общим словом, значит, оно не значимо и следует внимательно смотреть на совпадение остальных слов
- число запросов с общими словами
- то же самое для различных слов

Rambler

Навигационные запросы:результаты

Ветка	Точность	Полнота	F-мера
Весь алгоритм	86.47 %	29.64 % ☹	44.15 %
click entropy	70%		
click entropy + query/url features	83%		
reformulations	59 %		

Rambler

Итого

- В задачах классификации выбор данных и модели зависят от задачи
- Очень интересные возможности предоставляют «пользовательские» данные
- Machine learning нам в помощь

Rambler

Спасибо за внимание!

Вопросы?

Хоруженко Марина
m.horuzhenko@rambler-co.ru

Rambler