Методы очистки запросов к поиску

Карпенко Максим, Протасов Сергей

Rambler Research

May 2011

Обзор доклада

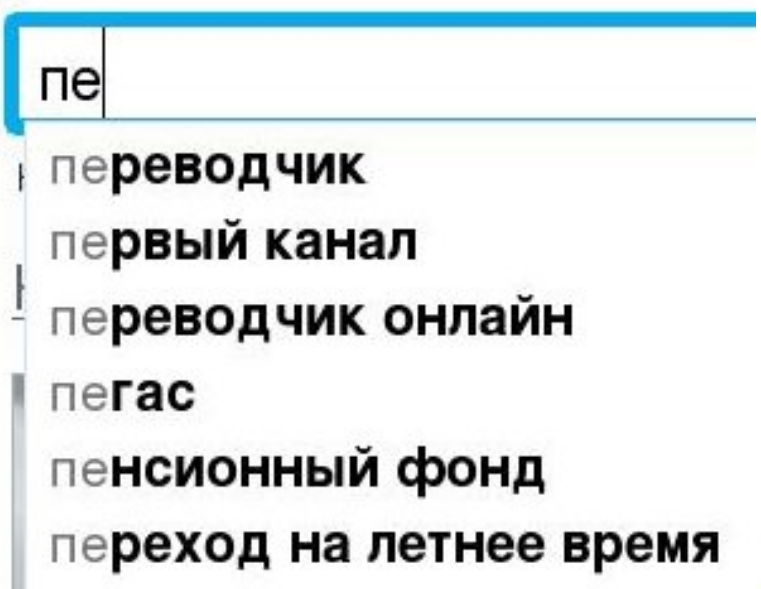
Rambler поиск

Обзор доклада

Rambler поиск

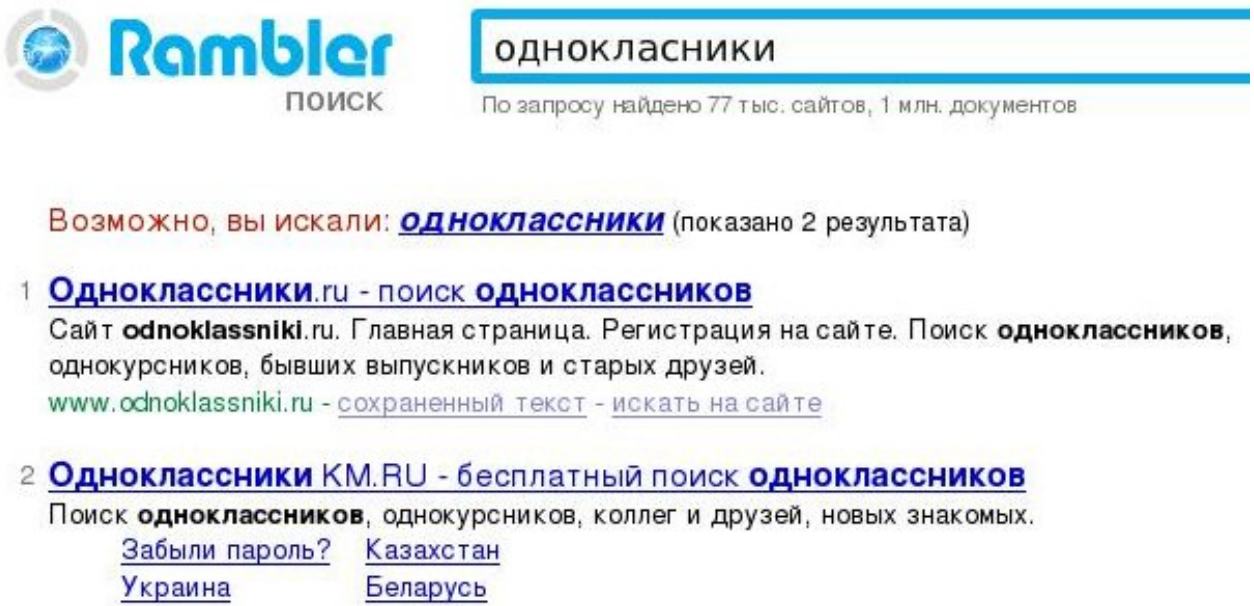
Подсказки в строке поиска

Половина запросов идёт через подсказки в строке поиска.



Подсказки опечаточника

10-15 % запросов - с опечаткой.



100 способов написать неправильно Бритни Спирс

- бритни спирз, бритни сприс, бритни сприрс, бритни спир, бритни спитс, ...
- бридни спирс, бритнти спирс, ритни спирс, вритни спирс, бритний спирс, бритней спирс, итни спирс, притни спирс, брини спирс, бринти спирс, бритри спирс, ...

Факты

- удаление (агенство/агентство) 8% случаев
- перестановка - 4% случаев
- вставка - 4% случаев
- замена - 80% случаев
- 90% опечаток с расстоянием 1

Обзор доклада

Алгоритм толпы

- Использование словарей - не подходит.
- Самый простой алгоритм - 28 строчек Норвига (norvig.com).

Алгоритм толпы

Расширения базового алгоритма

- модель ошибок, типы ошибок
- уверенность в результате
- модель языка (биграммы,очистка)

Обзор доклада

Rambler поиск

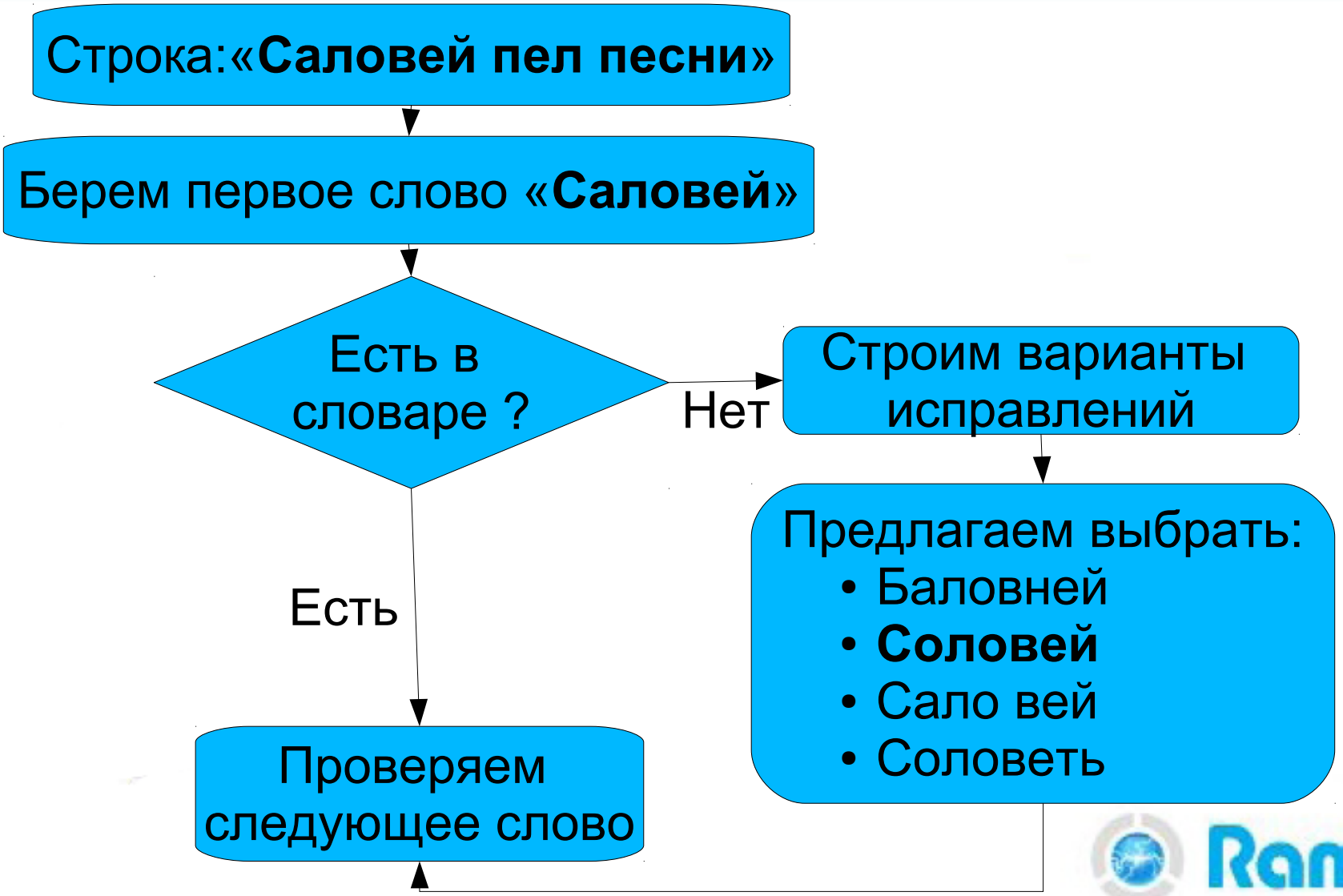
Некоторые методы очистки моделей языка запросов к поиску

Карпенко М.
Протасов С.

Рамблер

Rambler
ПОИСК

Простейший опечаточник



Rambler поиск

Словарь опечаток

лиkимии → лейкемии
лекимии → лейкемии
ликимия → лейкемия
...

Плюсы:

- Высокая точность

Минусы:

- Недостаточная полнота
- Высокая трудоемкость
- Не исправляет смысловые опечатки.

  Небо сырое, но дождя нет → Небо серое, но дождя нет

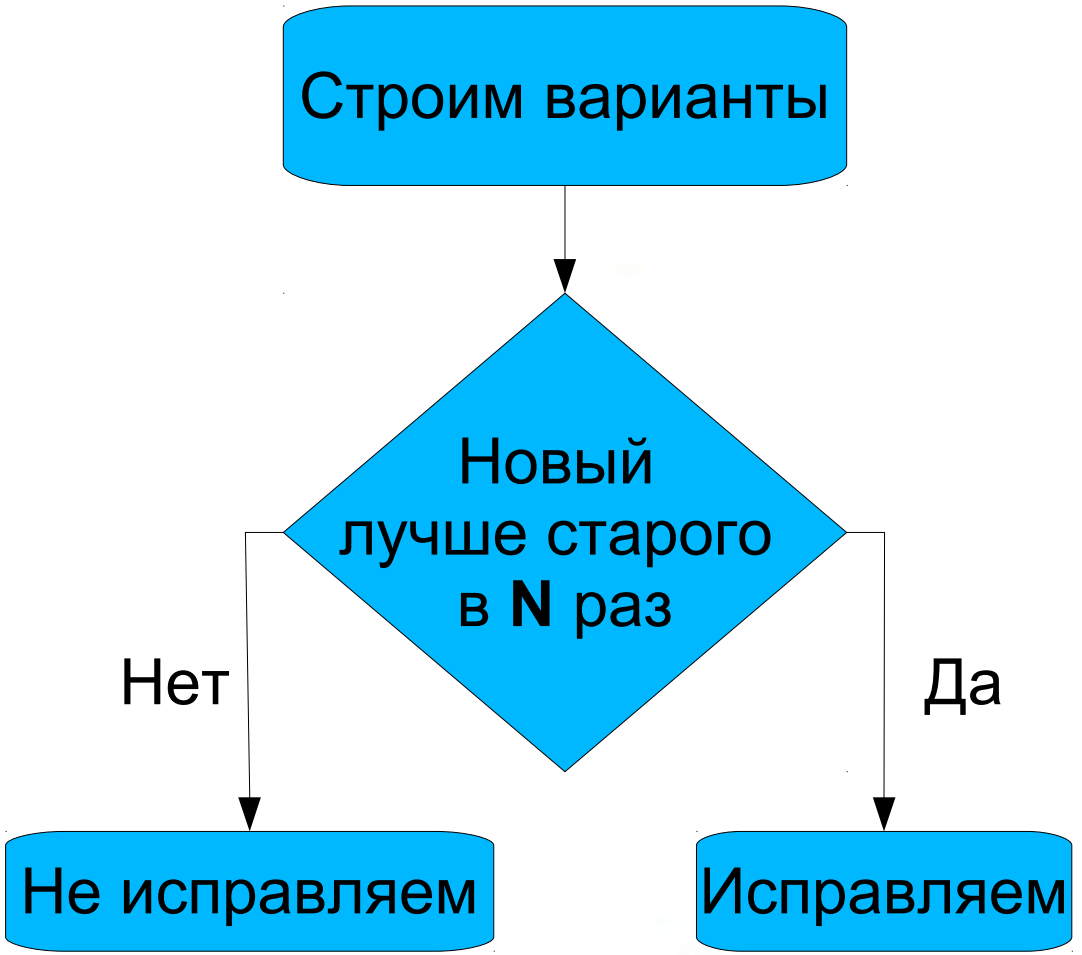
- Плохо исправляет ошибки склейки и разбиения слов.

Rambler
поиск

Модель языка.

укусила сабака - 18
укусила собака - 3571

группа собака - 11
группа сабака - 740



Rambler поиск

Опечаточник = Модель ошибок + Модель языка

$$argmax_c P(w|c) P(c)$$

- “w” - изначальный вариант (слооово)
- “c” - вариант исправления (слово)
- P(c) – языковая модель
  - P(слово) = 78669
  - P(слооово) = 487
- P(w|c) – модель ошибок
  - P(слооово|слово) = 0.153
  - P(слооово|слооово) = 0.99

P(слооово|слово)P(слово) > P(слооово|слооово)P(слооово)

*0.153*78 669 > 0.99*487*

12 036 > 482

Rambler
поиск

Расширенный опечаточник

Используется:

- Степень уверенности в исправлении
- Разные модели языка.
  - *Разные модели для однословных и многословных запросов*
  - *Тематические модели.*
- Поведенческие метрики.

Проблемы, создаваемые ошибками в модели.

$$argmax_c P(w|c)P(c)$$

- Частототные опечатки.
  - Не исправляются
    - *Инцефалит / энцефалит*
    - *Интелегенция / интеллигенция*
  - 2 ошибки исправляются на одну частотную
    - *Хандраксит → Хондраксит → Хондроксит*

- Редкие опечатки.
  - Снижается скорость работы.
  - Модель занимает много места.
    - *Вентилятор (112 опечаток): Вентиолятор, Вентилатро, Вентиятор …*
    - *География (84 опечатки): Георафия, Географффия ..*

Rambler
поиск

Удаление низкочастотных ошибок в модели

- Удаляем “невероятные” слова языковой модели.
  - Качество не меняется
  - Скорость +58 %
- Самоочистка
  - Качество не меняется
  - Скорость +9 %
- Результаты поиска.
  - Качество + 1,2 %
  - Скорость + 5 %

Rambler
поиск

Выявление и удаление частотных ошибок.

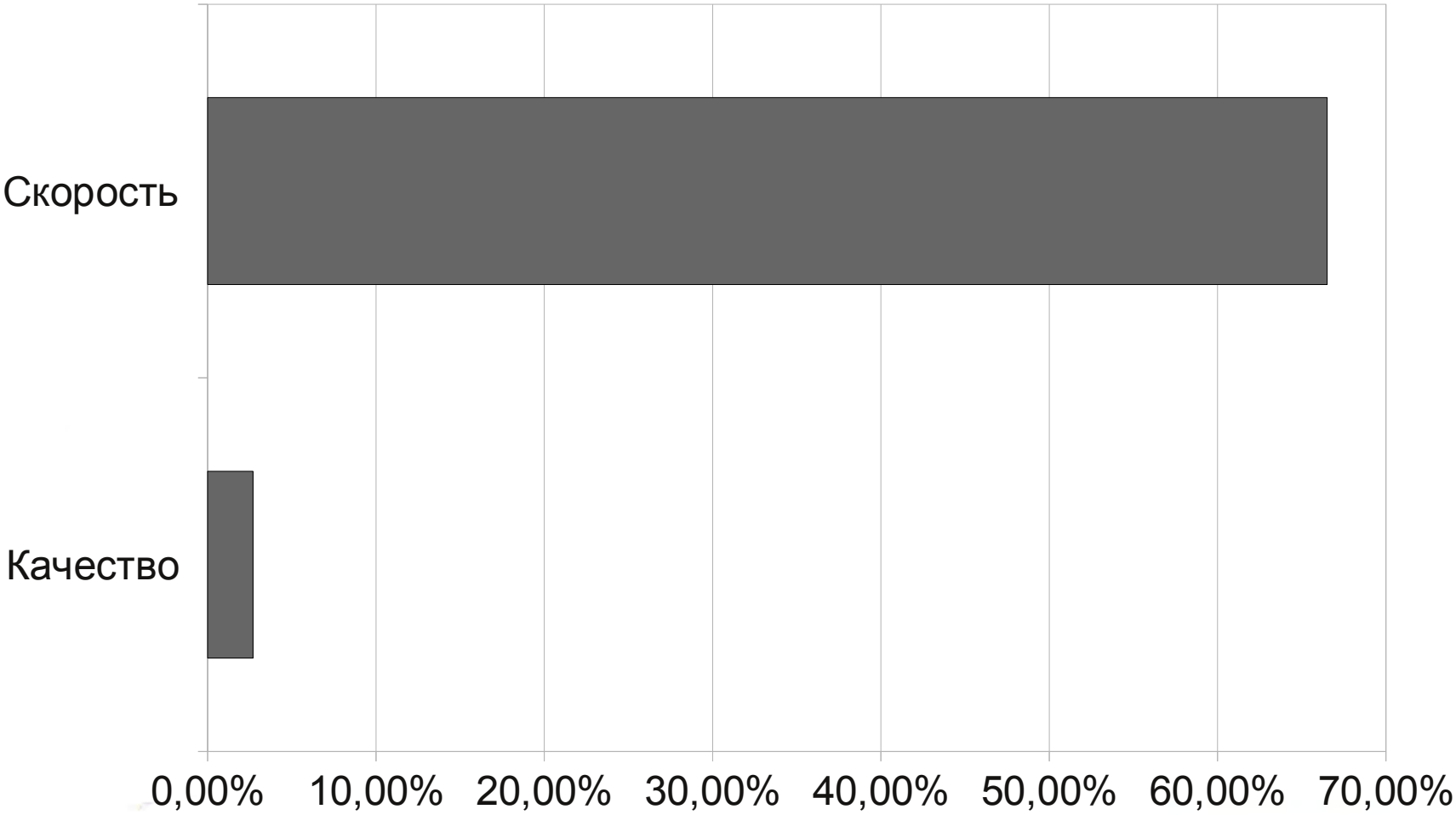
- Соотношение частотности похожих слов.
  - "бессо**н**ица" 33735
  - "бессонница" 34471

- Использование общего контекста.
  - программы ***для android***
  - програ**м**ы ***для android***
  - п**р**аграммы ***для android***

- Результаты поиска. (Уровень цитируемости)
  - “гидропередача” 28376
  - “гидро передача” 52

Rambler
поиск

Результат удаления ошибок.



Rambler
поиск

Проблемы очистки

- Имена и фамилии.
- Названия: лекарств, компаний, населённыхпунктов.

Система выделения признаков запросов:

Тематика лекарства (1000 признаков)

- *карбуситрат горелутамин*
- *ниналонин оксацтозан пициллин репитоксодин фитотиланадат строин тубутатин*
- *ликсферид питиазолол ягмониламид карбозыкол фенамещенный риптомазоин*

Спасибо за внимание.

Вопросы ?

m.karpenko@rambler-co.ru
s.protasov@rambler-co.ru

Rambler
ПОИСК