Разработка системы подсказок и исправления ошибок для тематических и географических поисковых систем.

Автор: Романова Т. С.
Руководитель: Калинин А. Л.

Московский авиационный институт

2010

Ошбики в запросах

Около 10 % запросов содержат ошибки.

- Орфографические ошибки.
- Клавиатурные опечатки.
- Ошибки в заимствованных словах.
- Ошибки в фамилиях и названиях брендов.
- Искажения в последовательности слов.
- Контекстные ошибки.
- Неправильная раскладка клавиатуры.

Данные для исправления

Основной источник информации — логи запросов.
Преимущества:

- постоянно обновляются;
- содержат и ошибочные запросы, но частотность правильных запросов выше;
- содержат информацию о сочетаемости слов;
- соответствуют специфике задачи.

Общая идея

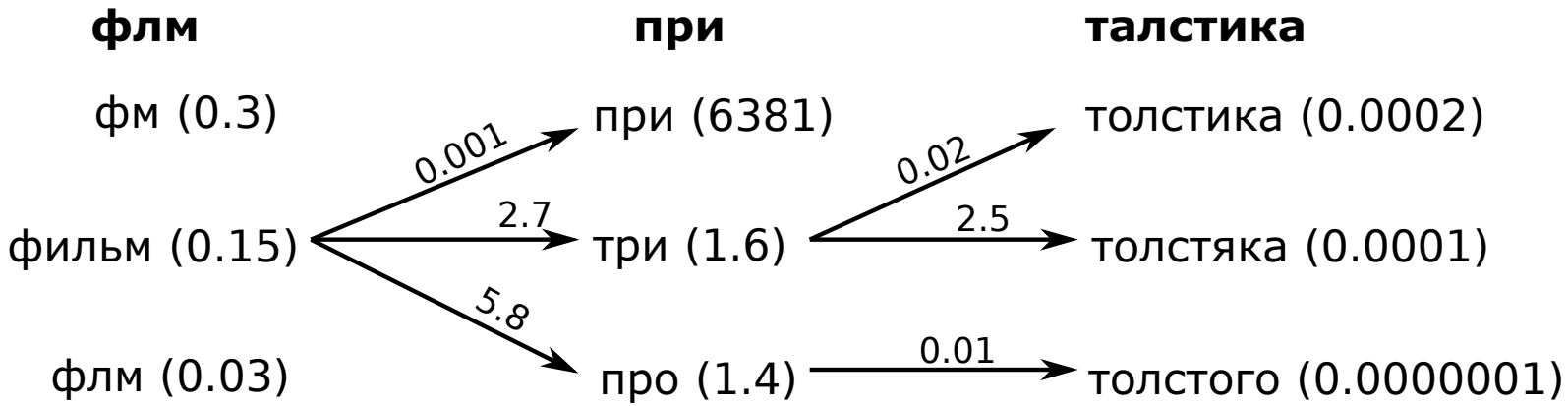
Итеративное исправление запроса:

растояие левиштейна → растояние левинштейна → расстояние левенштейна.

Шаги одной итерации:

1. Разбить запрос на части.
2. Для каждой части составить список вариантов замен.
3. Оценить вес каждой замены.
4. Составить граф слов: вершины — варианты замены, веса дуг — вероятности перехода от слова к слову.
5. Найти оптимальный путь в графе — алгоритм Витерби.

Пример графа слов



Разбиение запроса на части

По словам и знакам препинания.

Достоинства:

- Единственный способ разбиения.
- Простота реализации.

Недостатки:

- Невозможно «склеить» две части слова.

  шварц негер $\nrightarrow$ шварценеггер

Оценка близости слов

1. Взвешенное расстояние Дамерау-Левенштейна.
   - Операции редактирования: вставка, удаление, замена, перестановка.
   - У каждой операции — свой вес, зависящий от аргументов операции.
   - Сложность алгоритма — $O(mn)$.
2. Вероятностная модель.

$$P(w|t) = \frac{P(t|w)P(w)}{P(t)},$$

$P(w)$ — вероятность слова из словаря,
$P(t)$ одинаковая для всех вариантов $\Rightarrow$ можно не вычислять,
$P(t|w)$ вычисляется методом динамического программирования с помощью экспериментально собранной статистики ошибок.

Генерация вариантов замены

1. Метод Soundex.
   Сигнатура: убрать повторяющиеся буквы → оставить только согласные → оглушить звонкие согласные.

Достоинства
Учет больнистваа фонетических ошибок.
Sign(режесер) = ршср
Words(ршср) = [режиссер, рыжая серии, оружие зорро...]

Недостатки
Пропуск непроизносимой согласной.
Sign(сонце) = снц
солнце $\notin$ Words(снц)

2. Генерация всех слов, находящихся на расстоянии Д.-Л. 1, и выборка тех, что есть в словарях.

Замечание: словари слов и словосочетаний используются вместе $\Rightarrow$ возможно разбиение слова на два.

Размеры словарей

Логи:

1. Большой лог запросов — 11 млн. запросов.
2. Тематические логи запросов — 10–40 тыс. запросов.
3. Частотный словарь русского языка — 30 тыс. слов.

Словари:

1. Словарь юниграмм — 1 млн. слов.
2. Словарь биграмм — 7 млн. словосочетаний.

Оценка качества работы

Три файла $f1, f2, f3$ по 500 случайных запросов в каждом.

Точность — сколько было исправлено из содержащих ошибку запросов.
Полнота — сколько правильных запросов было оставлено без изменения.

	Ошибки	Точность	Полнота
f1	8.5 %	54.7 %	80 %
f2	6.8 %	61.8 %	79.6 %
f3	3.8 %	58 %	93.6 %

Удачные подсказки

свиной грип → свиной грипп, но
белый грип → белый гриб
гари потер → гарри поттер
камень в ерусалиме → камень в иерусалиме
нтвсмотреть онлайн → нтв смотреть онлайн
женская приступность → женская преступность
деревянный плод → деревянный плот
грибница фараона → гробница фараона
аднакласнеки → одноклассники.

Неудачные подсказки

швартцнегер $\longrightarrow$ шварцнегер
болезнь людмилы гурченко $\longrightarrow$ болезнь у людмилы гурченко
система 3 $\longrightarrow$ система в
задача про черепах $\longrightarrow$ задача про черепаху
аристотель и работа о душе $\longrightarrow$ аристотель и работа в душе
то и джери $\longrightarrow$ то и джерри
отстрел воров $\longrightarrow$ отстрел ворон
бьянка за тобой $\longrightarrow$ пьянка за тобой
баров юрий $\longrightarrow$ шаров юрий
vfnx htdfyi $\nrightarrow$ матч реванш

Результаты

- Разработана вероятностная математическая модель для вычисления близости двух слов языка.
- Исследована возможность использования сильно зашумленных данных, хранящихся в логах запросов, для итеративного исправления запроса.
- Разработана модификация метода Soundex для сужения множества вариантов замен.
- Реализована программа, демонстрирующая применимость этих решений.