Алгоритм построения оценок весов интентов для многозначных запросов

Артём Григорьев
445-ая группа

Кафедра Системного программирования
Математико-механический факультет
СПбГУ

Научный руководитель: к. ф.-м. н. Грауэр Л. В.
ООО «Яндекс»

Предметная область

Многозначный запрос – запрос, по которому возможны несколько пользовательских интентов (намерений пользователя).

Примеры: ягуар, наполеон, титаник…

IA-метрики (intent-aware) – метрики качества поиска, учитывающие различные интенты по многозначным запросам.

$$metrics = \sum_{i=1}^{k+1} w_i \cdot metrics(O_i)$$

Постановка задачи

Сейчас: Расчёт весов для IA-метрик производится вручную. Асессоры получают небольшой набор случайных сессий, должны определить по сессии интент. Доля сессий с данным интентом = вес.

Минусы: ограниченные возможности асессоров, => малое количество сессий по запросу, редкие обновления.

Задача: Придумать и реализовать алгоритм, вычисляющий по заданному на вход запросу набор

Алгоритм

- Формирование множества связанных запросов
- Кластеризация
  - Построение графа запросов и документов
  - Случайное блуждание по графу
  - Кластеризация по векторам предельных вероятностей документов
- Распределение сессий по кластерам и расчёт весов

Построение графа

- Вершины – запросы (Q) и документы (D)
- Рёбра:
  - Q_1 -> Q_2 (вес = вероятность переформулировки)
  - Q -> D (вес = вероятность клика)
  - Петли D -> D (вес = 1)
- Полученный граф – марковская цепь
  - Документы – конечные состояния

Результаты

- Разработан алгоритм
  - 65% наборов интентов найдено полностью
  - 94% без одного интента
  - Ошибки в точности в среднем <= 0.17
- Создан веб-инструмент для запуска и анализа результатов
- Утилита для расчёта данных по переформулировкам на кластере MapReduce

Дальнейшая работа

- Создание полуавтоматической системы проверки точности и полноты
- «Правильная» фильтрация «мусорных» данных
- Использование лингвистических данных при распределении сессий по кластерам
- Определение интентов из коротких, малоинформативных сессий
- Другие алгоритмы кластеризации и функции сходства