САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Математико-механический факультет
Кафедра системного программирования

Оценка распределения вероятностей поисковых целей для многозначных запросов

Артём Григорьев

Научный руководитель:
к.ф.-м.н., ст. преп. Л.В. Грауэр

Рецензент:
к.ф.-м.н., доц. Д.С. Шалымов

Оценка качества поиска

- Корзина запросов — выборка из потока запросов к поисковой системе
- Поисковые выдачи по запросам из корзины
- Сбор экспертных оценок релевантности документов в каждой выдаче
- Расчёт метрик по каждому запросу и усреднение по корзине

Многозначные поисковые запросы

- **Многозначный запрос** — запрос, поисковая выдача по которому может содержать документы, соответствующие различным целям (интентам) пользователя
  - **ягуар**: автомобиль, животное, напиток, фильм
  - **17 мгновений весны**: о фильме, скачать, смотреть
  - **Штирлиц**: персонаж, анекдоты, программа, игра

$$metrics_{IA}(q, D) = \sum_{i \in I} w_i \cdot metrics(q, D, i)$$

- **Вес интента** (w_i) — оценка вероятности того, что пользователь, задавший многозначный запрос, имел в виду данный интент

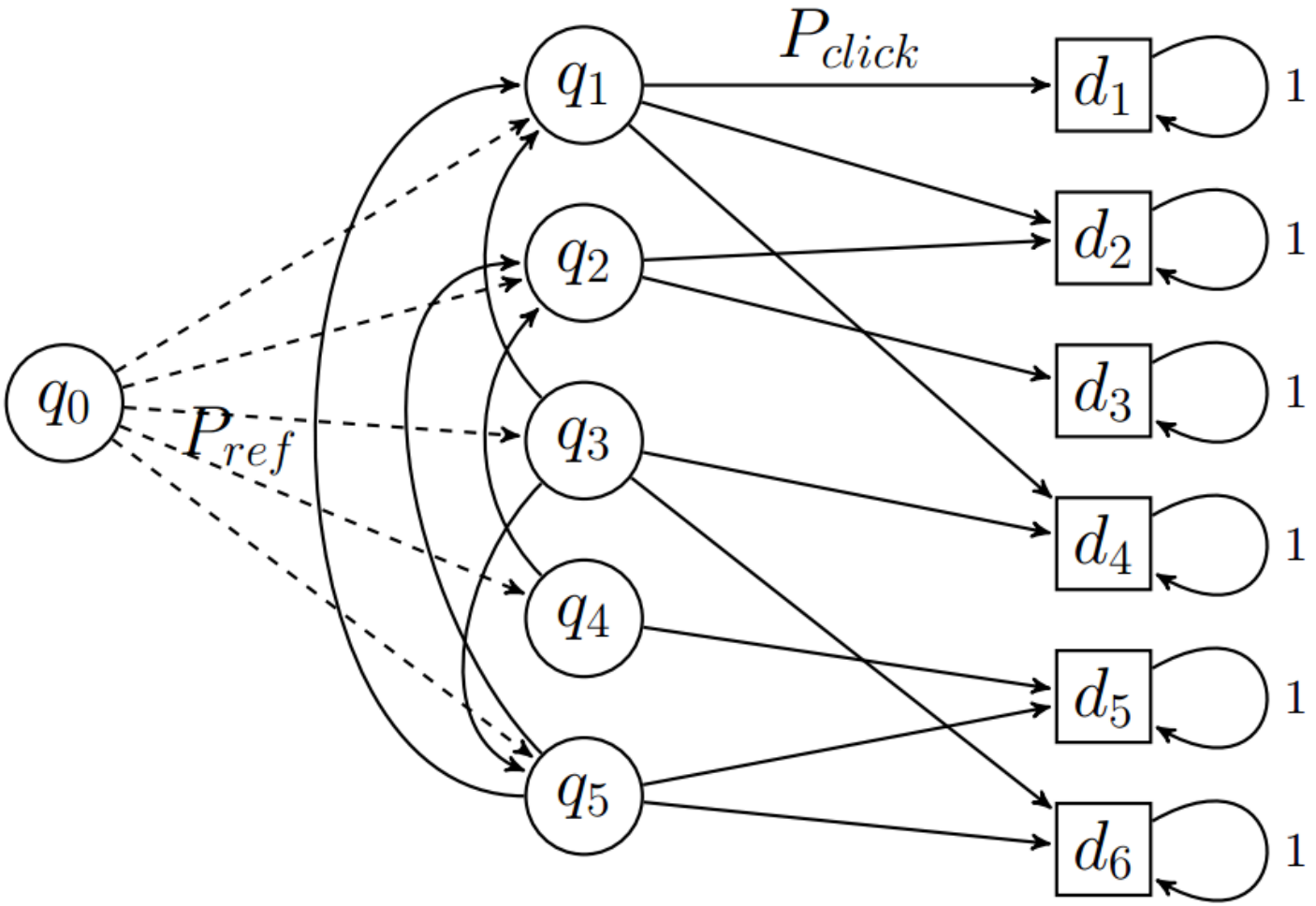
Постановка задачи

- Разработать алгоритм расчёта оценок весов интентов для многозначных запросов
- Создать систему для запуска и анализа расчётов
- Оценить качество полученных результатов

Связанные исследования

- Выделение логических сессий (сегментация)
  - Yury Ustinovskiy et al., 2013
- Исследования переформулировок
  - Rosie Jones et al., 2006
  - Wang Xuanhui et al., 2008
  - Paolo Boldi et al., 2009
- Кластеризация запросов
  - Beeferman Doug et al., 2000
  - Eldar Sadikov et al., 2010
  - Radlinski Filip et al., 2010

Общий алгоритм: марковская модель



Общий алгоритм: кластеризация запросов

- Случайное блуждание: **n** шагов
- Представление запросов в векторном пространстве:
  - с координатами-документами
  - с координатами-словами из документов
- Функция сходства:

$$Sim(x, y) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

- Кластеризация Complete Linkage до фиксированного порога сходства **θ**

Общий алгоритм: расчёт весов

- Кластер: запросы и документы с весами — соответствуют определённому интенту
- Сессия: последовательность запросов и переходов на документы, ограниченная некоторым периодом времени
- Распределяем выборку сессий, содержащих исходный многозначный запрос, по полученным кластерам
- Доля сессий, попавших в определённый кластер, — вес этого интента

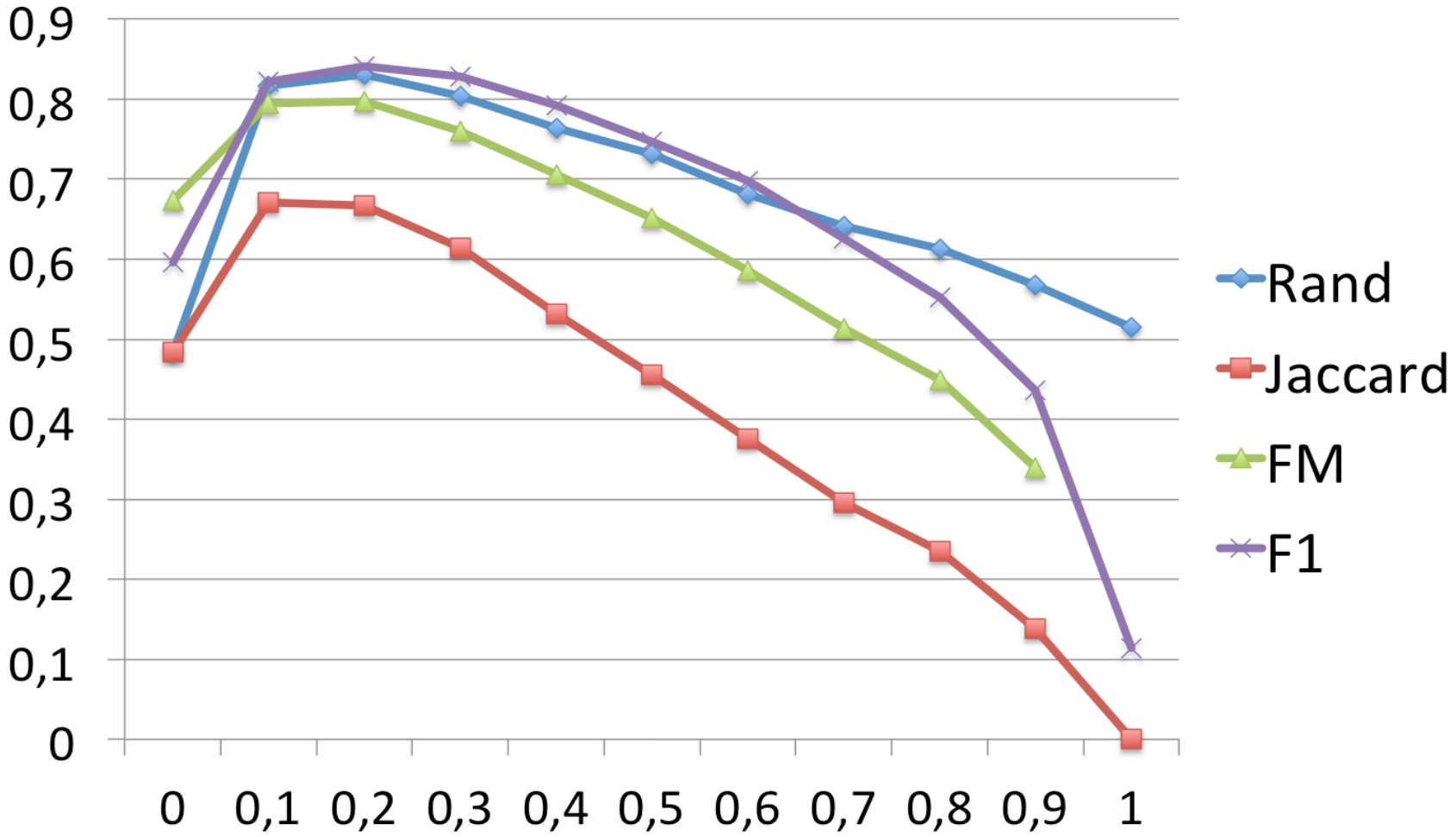
Общий алгоритм

- **На входе:** многозначный запрос
- Построение набора связанных запросов
- Построение цепи Маркова, случайное блуждание
- Кластеризация запросов
- Обработка сессий для подсчёта весов кластеров
- **На выходе:** вес для каждого кластера, кластер соответствует интенту

Эксперименты и оценки

- Качество кластеризации
  - Метрики: Rand, Jaccard, FM, F1-мера
  - Сравнение со случайной кластеризацией (критерий перестановок)
  - Сравнение двух способов представления запроса в векторном пространстве (бутстреп)
  - Зависимость от количества шагов **n** и порога **θ**
- Качество расчёта весов
  - Как нечёткая кластеризация: Fuzzy Rand
  - Как классификатор наличия интента: точность и полнота

Качество кластеризации



Зависимость метрик качества от порогового значения θ при кластеризации, n = 16, вектора с координатами-словами

Качество расчёта весов

- Размечены 30 запросов по 100 сессий
  - Для 64% из них определён интент

- 95% бутстреп доверительные интервалы:
  - Fuzzy Rand (0.685, 0.717)
  - Точность (0.864, 0.905)
  - Полнота (0.413, 0.466)

Результаты

- Построен алгоритм расчёта оценок весов интентов для многозначных запросов
  - Реализация на языке программирования Java с использованием утилит на C++ для подготовки и обработки данных на кластере MapReduce
- Реализована система для запуска расчётов, анализа результатов, ручной разметки тестовых данных и оценки качества алгоритма
  - В виде веб-сервиса
- Проведена оценка качества кластеризации и расчёта весов, а также сравнительный анализ запусков алгоритма с различными его параметрами
  - На размеченных вручную 30 многозначных запросах (более 5000 связанных) и 3000 сессиях

Метрики качества кластеризации

$$Rand = \frac{SS + DD}{SS + SD + DS + DD}$$

$$Jaccard = \frac{SS}{SS + SD + DS}$$

$$FM = \sqrt{\frac{SS}{SS + SD} \cdot \frac{SS}{SS + DS}}$$

Метрики качества кластеризации

- n_{ij} — количество запросов, попавших в c_i и g_j одновременно,
- n_i — количество элементов кластера c_i,
- n_j — количество элементов класса g_j,
- $Precision(i,j) = \frac{n_{ij}}{n_i}$ — точность,
- $Recall(i,j) = \frac{n_{ij}}{n_j}$ — полнота,

$$F1(i,j) = \frac{2 \cdot Precision(i,j) \cdot Recall(i,j)}{Precision(i,j) + Recall(i,j)}$$

$$F1 = \sum_j \frac{n_j}{N} \max_i F1(i,j)$$

Fuzzy Rand

$$E_P(S, S') = 1 - ||P(S) - P(S')||$$

$$d(P, Q) = \frac{\sum_{S,S' \in \mathbb{S}_{marked} \cap \mathbb{S}_{matched}, S \neq S'} |E_P(S, S') - E_Q(S, S')|}{C_n^2}$$

$$Rand_{fuzzy} = 1 - d(P, Q)$$

Точность и полнота

$$Precision_{sessions} = \frac{|\mathbb{S}_{marked} \cap \mathbb{S}_{matched}|}{|\mathbb{S}_{matched}|}$$

$$Recall_{sessions} = \frac{|\mathbb{S}_{marked} \cap \mathbb{S}_{matched}|}{|\mathbb{S}_{marked}|}$$