Московский Физико-Технический Институт
Вычислительный Центр РАН

Оценивание сходства пользователей и ресурсов путем выявления скрытых тематических профилей

Лексин В.А.
vleksin@mail.ru

Воронцов К.В.
voron@ccas.ru

Задача АКС (анализа клиентских сред)

- Дано:
  - множество пользователей U
  - множество ресурсов R
  - выборка посещений $\{u_i, r_i\}_{i=1}^{l} \in U \times R$

- Требуется построить функции сходства:
  - пользователей $\rho_U(u, u')$
  - ресурсов $\rho_R(r, r')$

Конечная цель АКС

- Решение целого спектра прикладных задач:
  - поиск схожих ресурсов и схожих пользователей
  - персонализация контента
  - сегментация клиентской базы
  - каталогизация ресурсов
  - визуализация карт сходства
- Основная идея АКС: ρ_U и ρ_R должны быть взаимосогласованными:
  - клиенты схожи, если они пользуются схожим набором ресурсов
  - ресурсы схожи, если ими пользуются схожие клиенты

Вероятностная постановка задачи, понятие профиля

- У каждого пользователя $u \in U$ имеется некоторое множество интересов или потребностей (тем).
- Множество всех тем обозначим через T.
- Профиль ресурса r — вектор вероятностей

$$q_{tr} = q(t \mid r),\ t = 1,\ldots,|T|,\ \sum_{t \in T} q_{tr} = 1.$$

- Профиль пользователя u — вектор вероятностей

$$p_{tu} = p(t \mid u),\ t = 1,\ldots,|T|,\ \sum_{t \in T} p_{tu} = 1.$$

Задача восстановления тематических профилей

- p-формула: $p(u,r) = \sum_t p(u)\, p_{tu} q(r \mid t,u)$

  По Байесу: $q(r \mid t) = \dfrac{q_{tr} q(r)}{\sum_{s \in R} q_{ts} q(s)}$

- q-формула: $p(u,r) = \sum_t q(r) q_{tr}\, p(u \mid t,r)$

  По Байесу: $p(u \mid t) = \dfrac{p_{tu} p(u)}{\sum_{s \in U} p_{ts} p(s)}$

- Выборка посещений: $D = \{u_i, r_i\}_{i=1}^{l}$

- Принцип максимума правдоподобия:

$$L(D; \{p_{tu}\}, \{q_{tr}\}) = \ln \prod_{i=1}^{l} p(u_i r_i) \to \max_{\{p_{tu}, q_{tr}\}}$$

Схема алгоритма двухуровневая

Повторять, пока не сойдется:

- Оптимизировать p_{tu} при фиксированных q_{tr}
  - E-шаг: $H_{tr}(u) = \dfrac{p_{tu} q(r \mid t)}{\sum_s p_{su} q(r \mid s)}$ -скрытые переменные
  - M-шаг: $p_{tu} = \dfrac{\sum_{r:(u,r)\in D} H_{tr}(u)}{\sum_{r:(u,r)\in D} 1}$ -профили пользователей
- Оптимизировать q_{tr} при фиксированном p_{tu}
  - E-шаг: найти скрытые компоненты
  - M-шаг: найти профили ресурсов

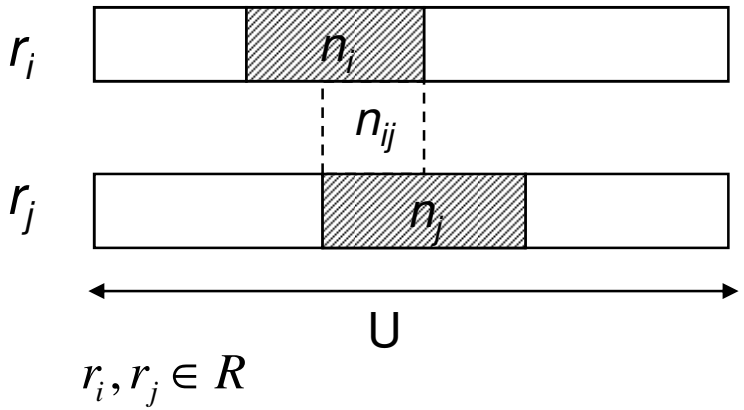
Алгоритмы вычисления метрики

- По профилям:

$$\rho_R(r, r') = \rho(p_{tr}, p_{tr'}) = \sqrt{\sum_{t=1}^{|T|} (p_{tr} - p_{tr'})^2}, r \in R, r' \in R,$$

- Вычисление корреляции:

$$\rho(r_i, r_j) = \left(\frac{n_i + n_j - 2n_{ij}}{n_i + n_j - n_{ij}} \right)^{\alpha}$$



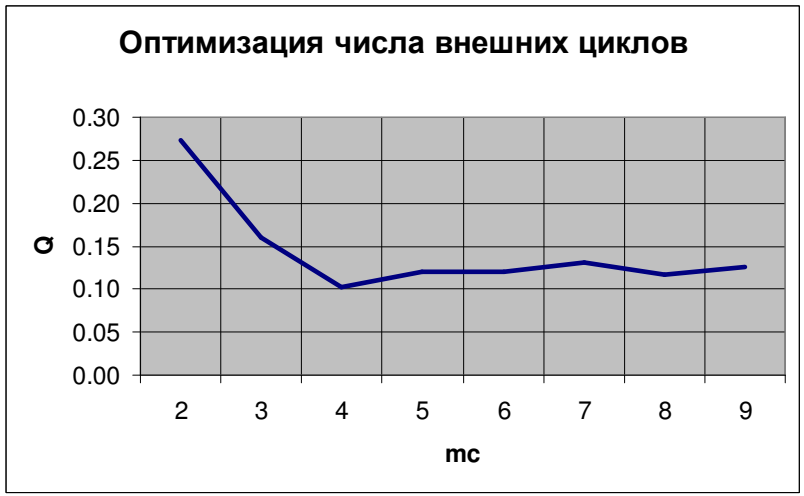
- Точный тест Фишера:

Гипергеометрическое распределение: $\rho(r_i, r_j) = P\{n_{ij} = x\} = \frac{C_{n_i}^{x} C_{|U|-n_i}^{n_j - x}}{C_{|U|}^{n_j}}$

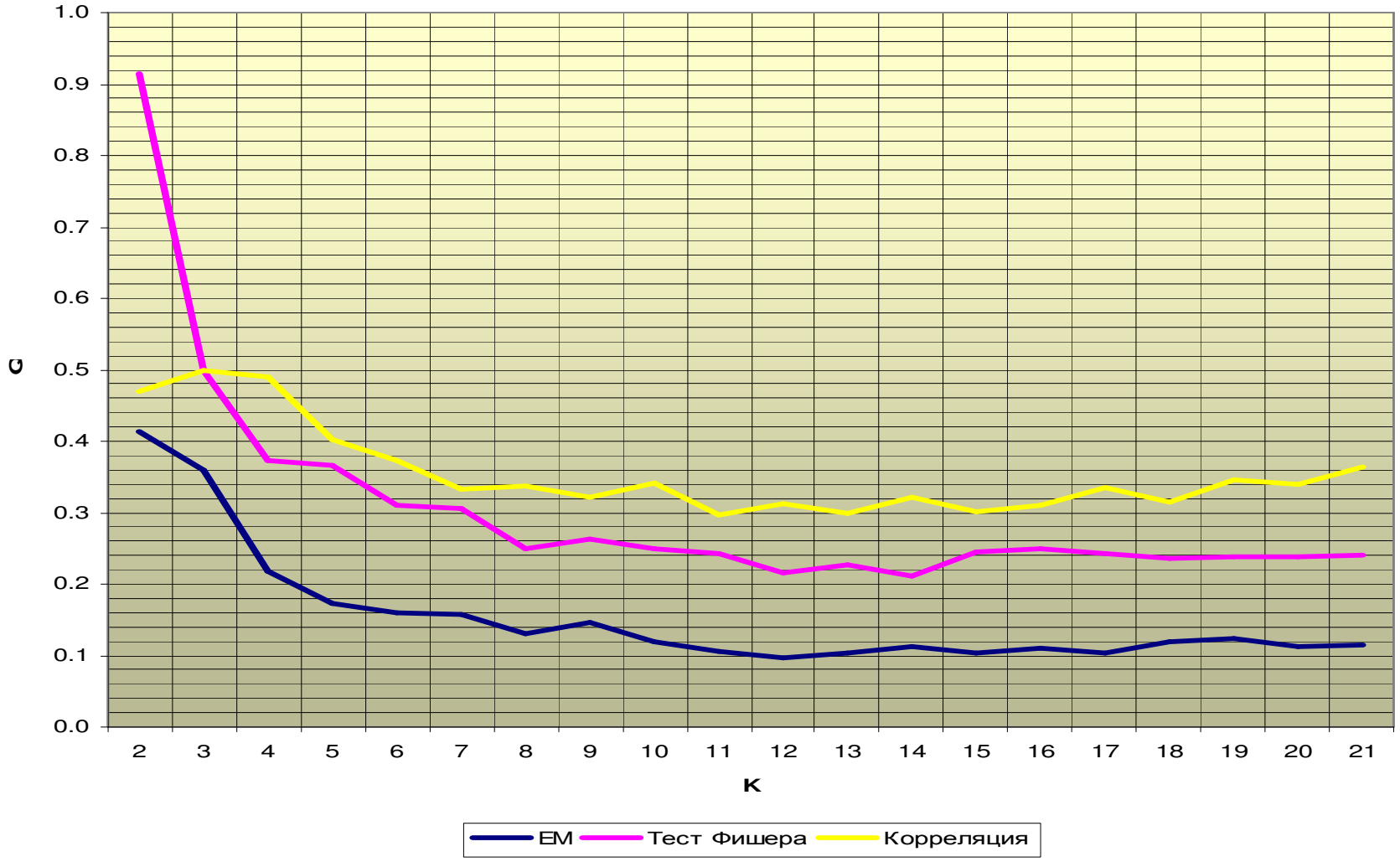
Методика оценивания качества метрики

- Данные поисковой машины Яндекс:
  - объем лога 4Гб
  - 14 606 пользователей
  - 1 972 636 ресурсов (из них 129 600 были выбраны)
  - интервал времени: 1 неделя работы поисковой системы
- Классифицированные экспертом ресурсы:
  - 396 сайтов
  - 8 классов
- Критерий качества построенной метрики:
  - количество ошибок при попытке классифицировать точки методом kNN, используя частичную классификацию ресурсов

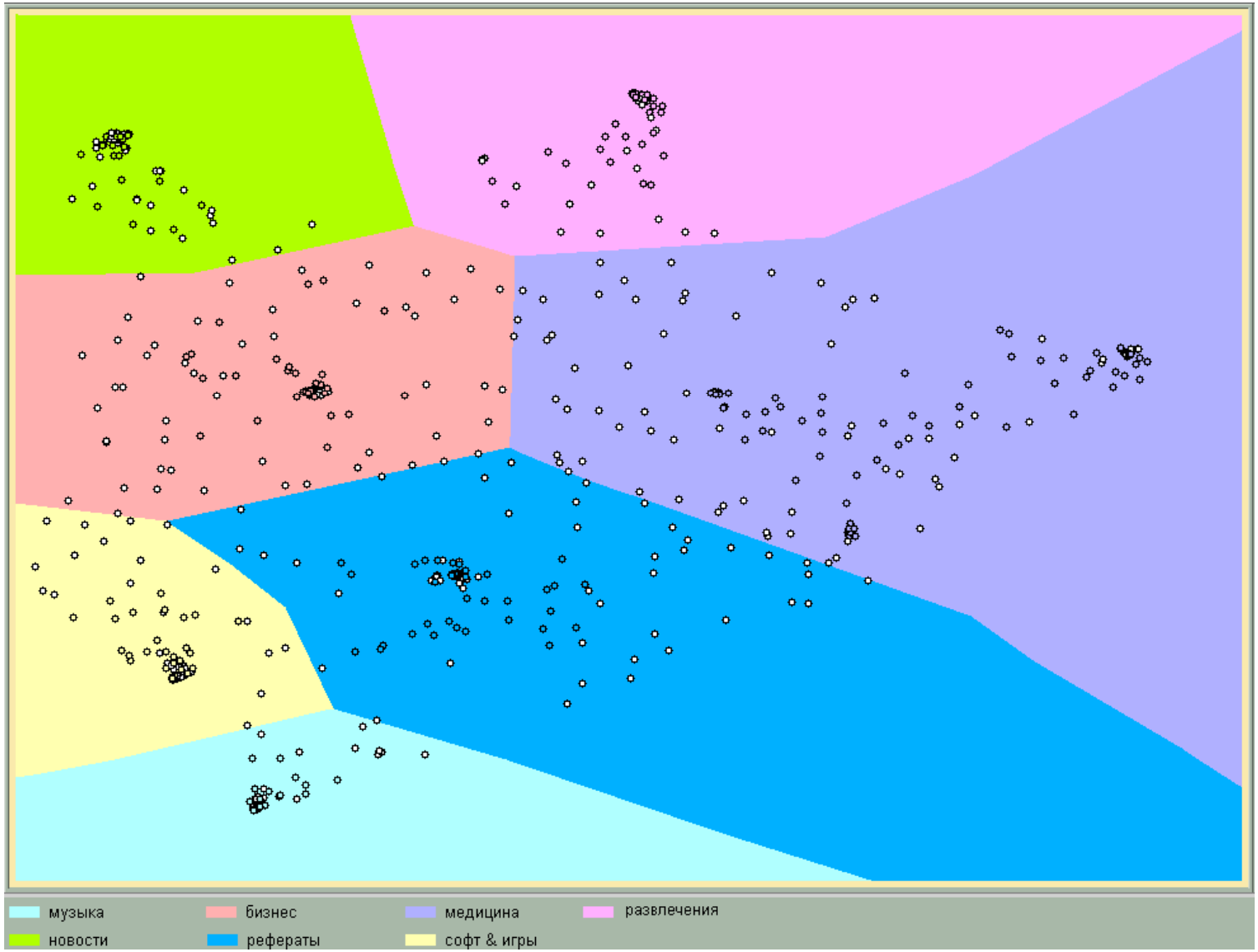
Настройка параметров алгоритма на данных поисковой машины



Сравнение алгоритмов и оптимизация параметра метода kNN



Карта сходства ресурсов по профилям



Выводы

- Улучшается качество метрик
- Восстанавливаются профили поддающиеся содержательной интерпретации
- Уменьшается объем хранимых данных, повышается скорость обработки
- Легко учитывается априорная информация
- Решается проблема «холодного старта»
- Широкий спектр применений

Список литературы

- [2] К. В. Воронцов, В. А. Лексин, Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов, тезисы ММРО-13, 2007
- [1] К. В. Воронцов, К. В. Рудаков, В. А. Лексин, А. Н. Ефимов, Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет, 2006
- [3] Resnik et al., Statistical collaborative filtering, 1994
- [4] Schein et al., Generative Models for Cold Start Recommendations, 2002
- [5] Jun Wang et al., A User-Item Relevance Model for Log-based Collaborative Filtering, 2006