Московский Физико-Технический Институт

Технология персонализации на основе выявления тематических профилей пользователей и ресурсов Интернет

Выполнил студент 175 группы:
Лексин В.А.
Научный руководитель:
к.ф.-м.н. Воронцов К.В.

Задача АКС (анализа клиентских сред)

- Дано:
  - множество пользователей U
  - множество ресурсов R
  - выборка посещений $(u_t, r_t) \in U \times R$
- Требуется построить функции сходства:
  - пользователей $\rho_U(u, u')$
  - ресурсов $\rho_R(r, r')$

Конечная цель АКС

- Решение целого спектра прикладных задач:
  - персонализация контента
  - сегментация клиентской базы
  - каталогизация ресурсов
  - визуализация карт сходства
  - и др.

- 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

Задача восстановления тематических профилей

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

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

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

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

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

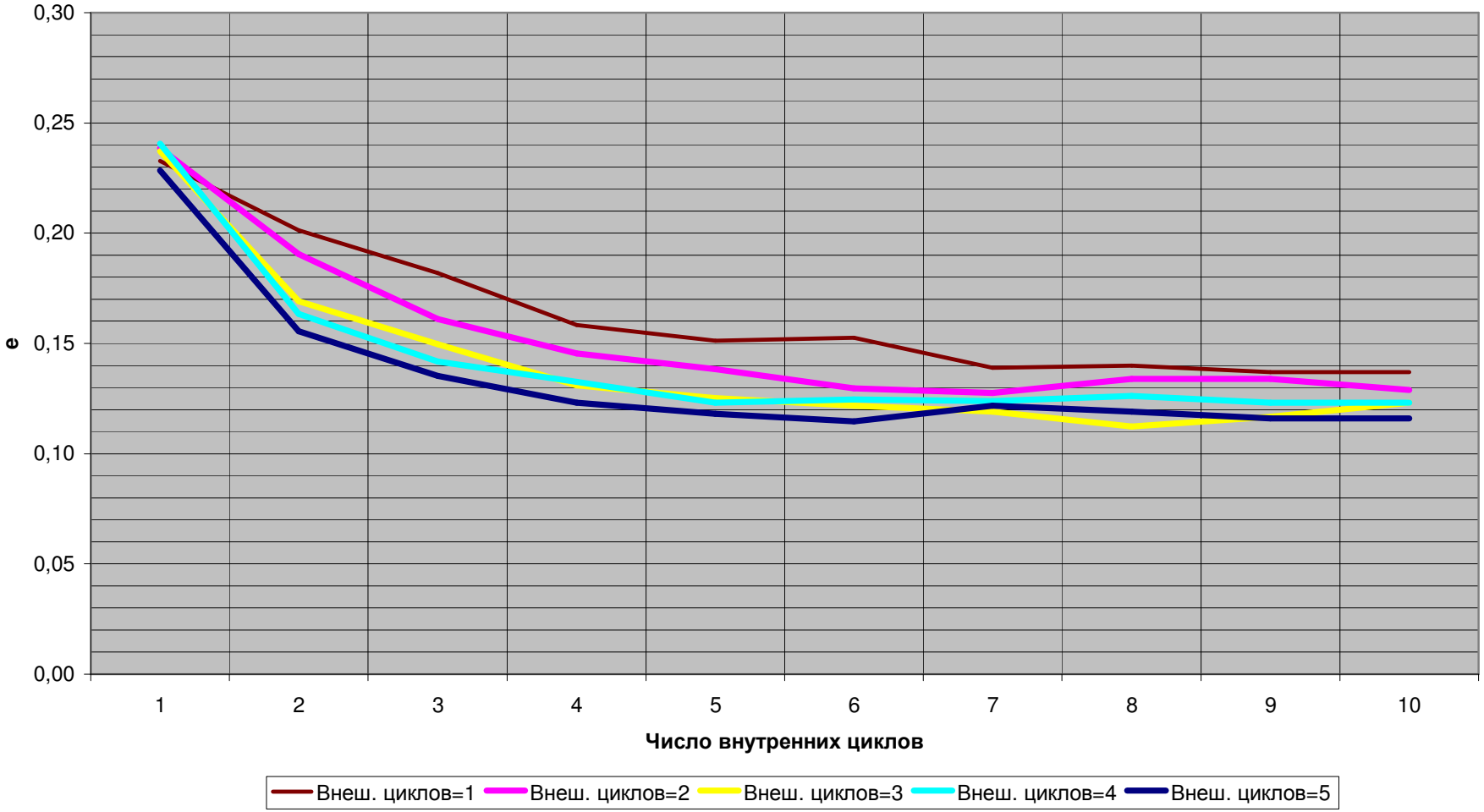
- Принцип максимума правдоподобия: $\ln \prod_{i=1}^{l} p(u_i r_i) \to \max_{p(t|u), q(t|r)}$

Схема алгоритма двухуровневая

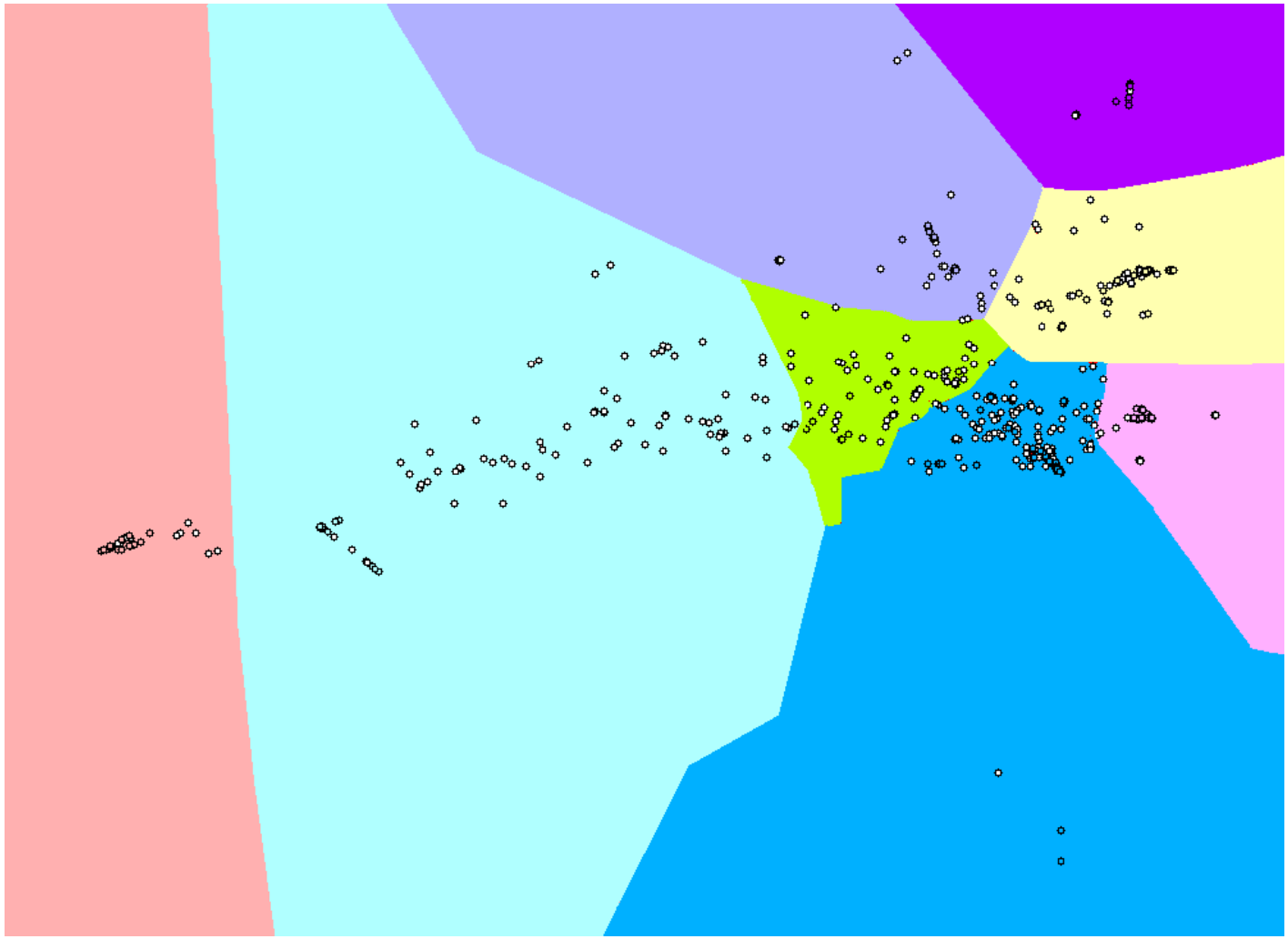
Повторять, пока не сойдется:

- Оптимизировать p_{tu} при фиксированных q_{tr}
  - E-шаг: $H_{tr}(u) = \frac{p_{tu} q(r \mid t)}{\sum_s p_{su} q(r \mid s)}$ -скрытые переменные
  - M-шаг: $p_{tu} = \frac{\sum_{r \in D_u} H_{tr}(u)}{\sum_{r \in D_u} 1}$ -профиль пользователя
- Оптимизировать q_{tr} при фиксированном p_{tu}
  - E-шаг: найти скрытые компоненты
  - M-шаг: найти профиль ресурса

Оптимизация числа итераций на внутренних и внешнем циклах

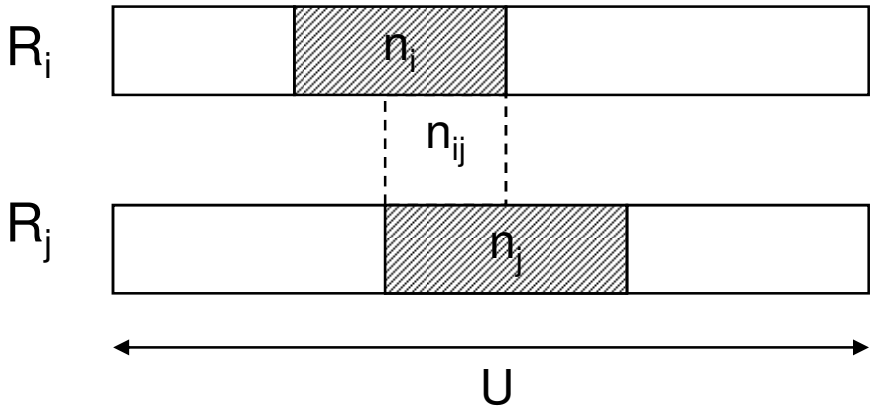


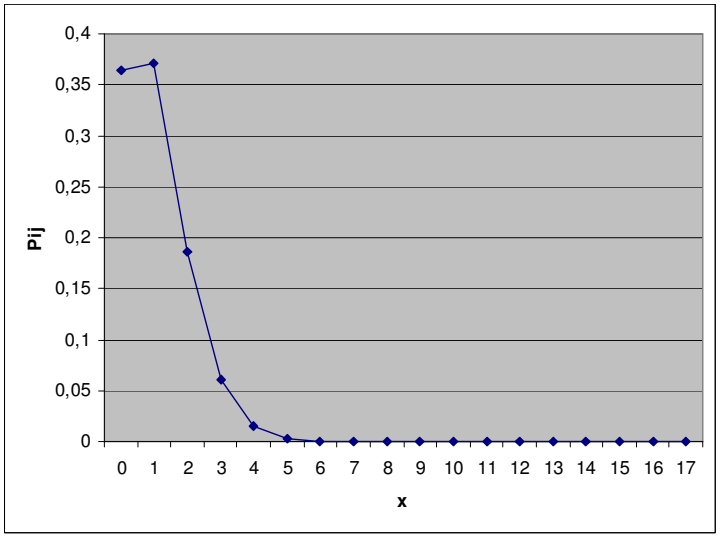
Карта сходства по EM-алгоритму



Точный тест Фишера

Гипергеометрическое распределение:

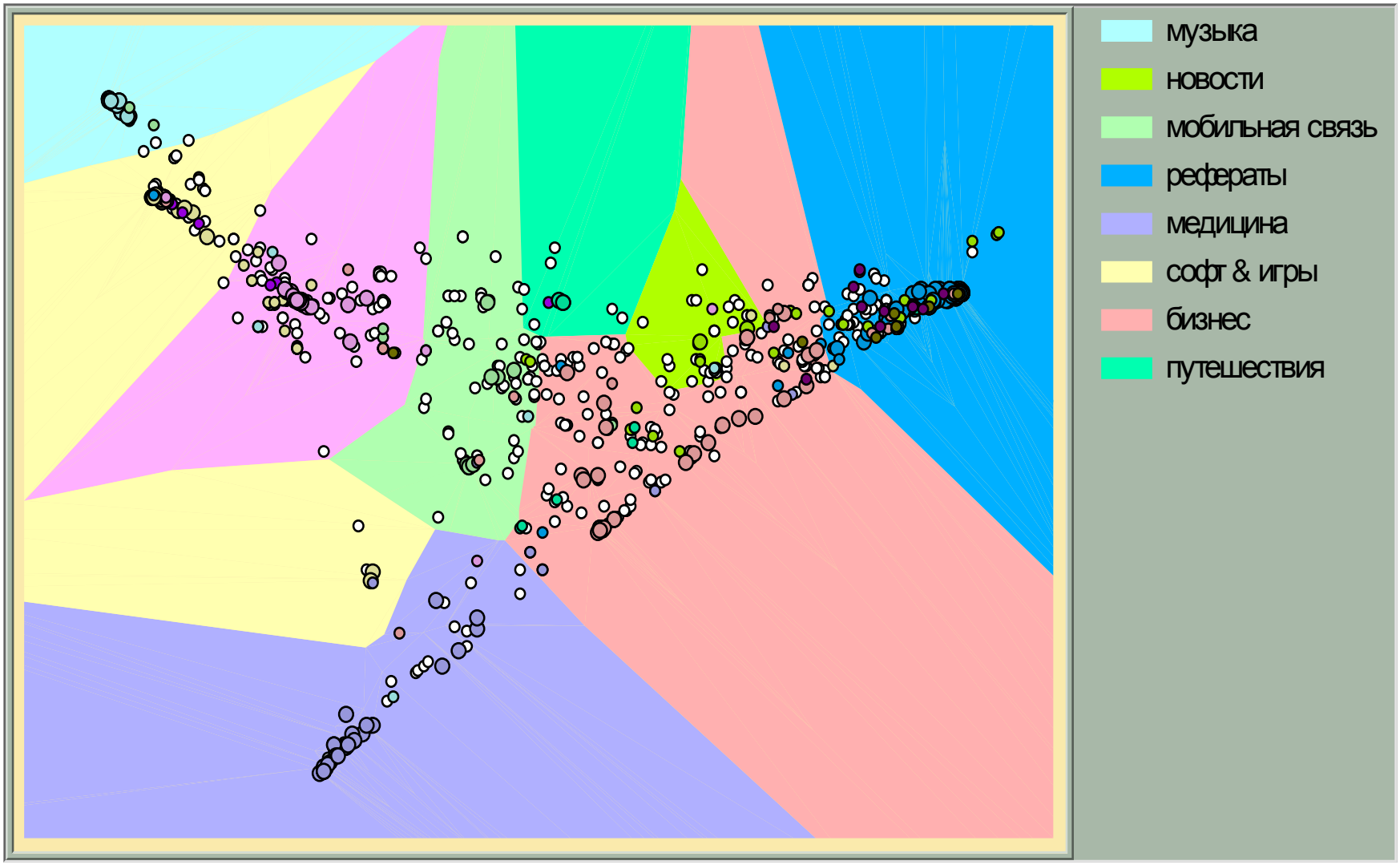




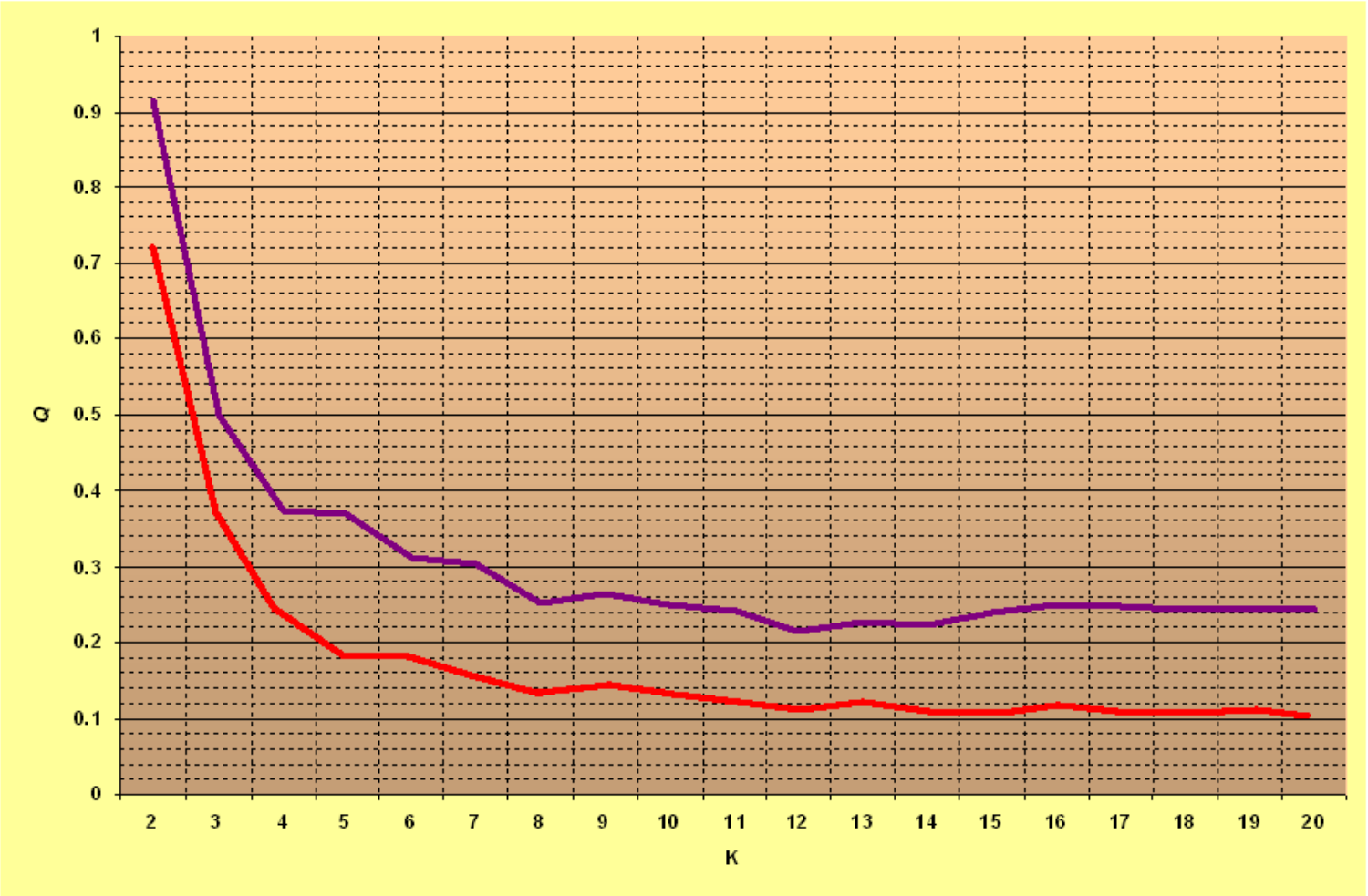
$$P_{ij} = P(n_{ij} = x) = \frac{C_{n_i}^{x} C_{U-n_i}^{n_j - x}}{C_U^{n_j}}$$

$$\rho(i, j) = \left(\frac{|\ln \alpha|}{|\ln P_{ij}|} \right)^3$$

Карта сходства ресурсов по тесту Фишера



Оптимизация количества соседей в методе kNN и сравнение алгоритмов



Выводы

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

Направления дальнейших исследований

- Построение иерархических профилей
- Оптимизация тематической структуры профиля
- Построение обобщенных профилей