Методика нейросетевой кластеризации корпуса текстов

Стулов Виктор Викторович

Научный руководитель:
Филиппович Андрей Юрьевич

МГТУ им. Баумана
Кафедра ИУ-5
НОК CLAIM

Цель работы и решаемые задачи

Цель исследования: разработка методики автоматического выявления групп семантически похожих документов.

Назначение работы: методика предназначена для разработчиков систем информационного поиска.

Задачи исследования:

- Анализ существующих методов кластеризации корпуса текстов;
- Разработка методики нейросетевой кластеризации;
- Разработка архитектуры программного комплекса, реализующего отдельные этапы методики;
- Реализация и отладка программного комплекса.

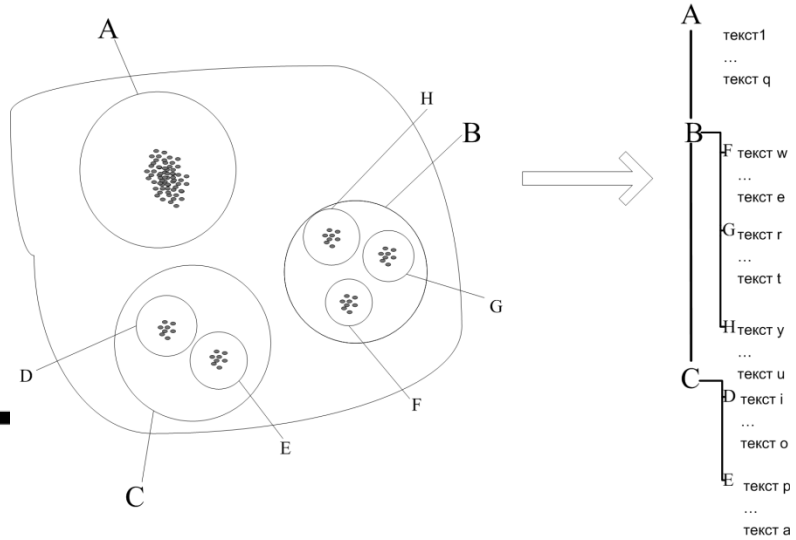
Классификация методов

1. По используемой модели текста
   a. Основанные на векторной модели текста;
   b. Основанные на модели суффиксного дерева;
   c. Основанные на графовой модели текста

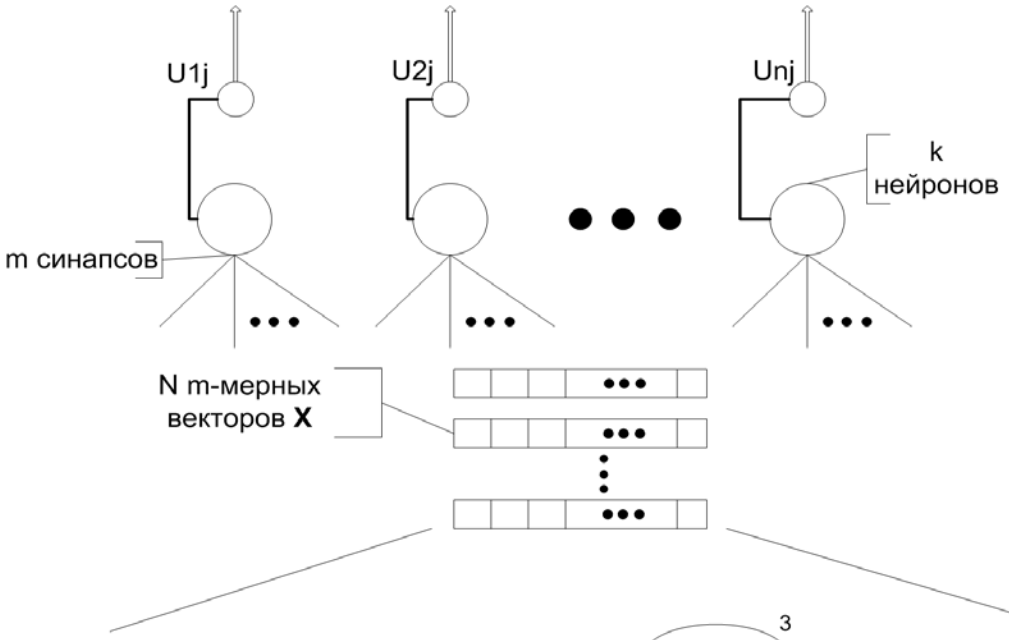
2. По принципу функционирования
   a. Использующие формулы близости текстов;
   b. Генеративные алгоритмы;

3. По функциональности
   a. Иерархические (дивизимные и агломеративные);
   b. Плоские;

4. По характеру получаемых кластеров
   a. Мягкие;
   b. Жесткие;



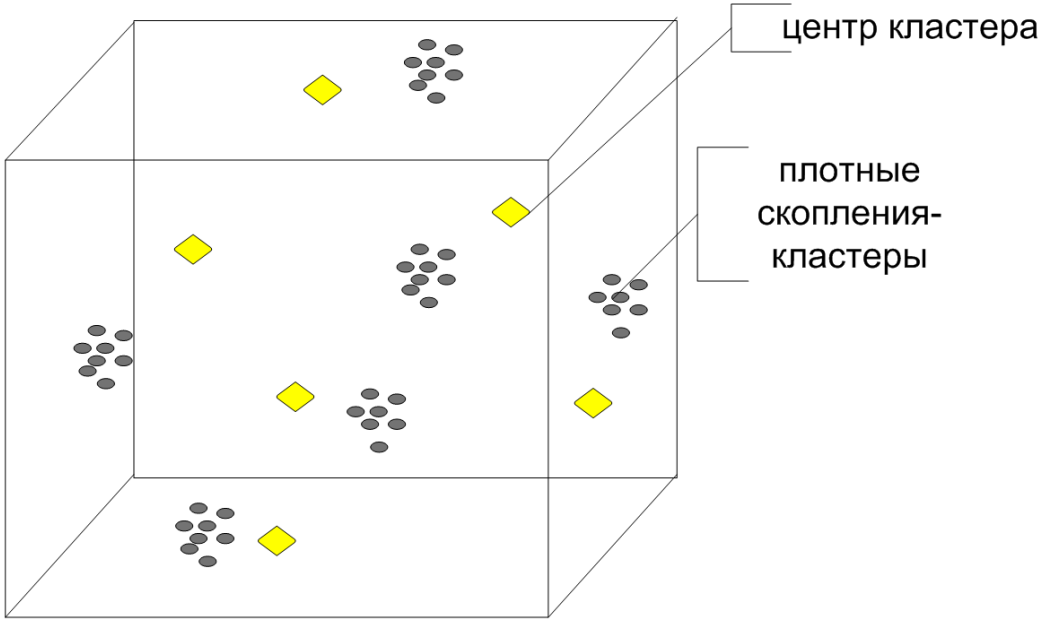
Нечеткая сеть Кохонена



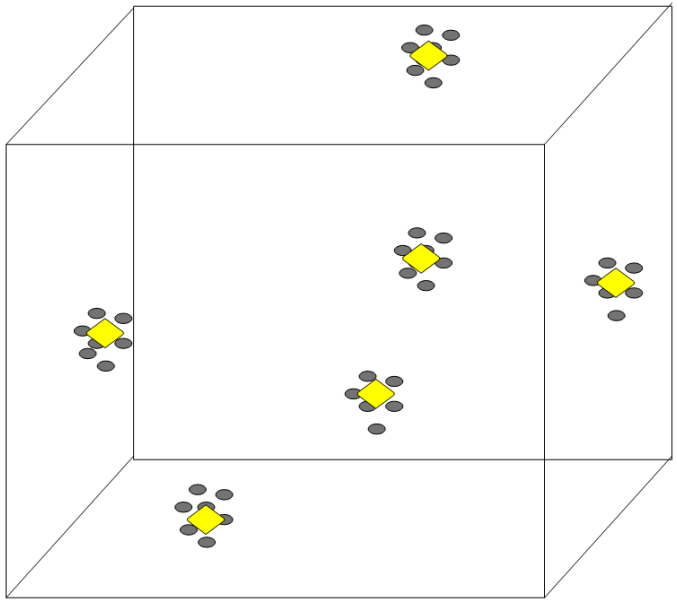
1. Используется ориентированная на семантику векторная модель;
2. Для сокращения размерности пространства признаков - латентный семантический анализ;
3, Мера сходства текстов – скалярное произведение нормализованных векторов

Нечеткая сеть Кохонена - результат

До обучения



После обучения

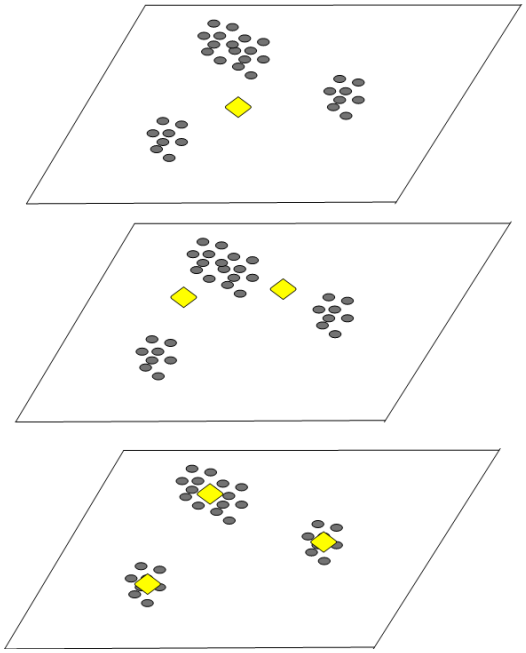


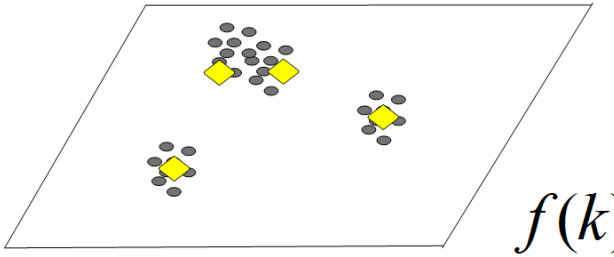
$$\mu_{ij} = \begin{cases} \left(\dfrac{1}{dist(\mathbf{X_j}, \mathbf{W_i}) \sum_{p=1}^{k} \dfrac{1}{dist(\mathbf{X_j}, \mathbf{W_p})} } \right)^{\frac{m}{(m-1)}} \\ 1, \text{если } dist(\mathbf{X_j}, \mathbf{W_i}) = 0 \\ 0, \text{если } \exists i: dist(\mathbf{X_j}, \mathbf{W_i}) = 0 \end{cases}$$

$$\boldsymbol{W_i}(t+1) = \boldsymbol{W_i}(t) + \frac{\sum_{j=1}^{N} \mu_{ij}(t)(\boldsymbol{X_j} - \boldsymbol{W_i}(t))}{\sum_{j=1}^{N} \mu_{ij}(t)}$$

$$m_t = m_0 - \frac{t(m_0 - 1)}{t_{max}}$$

Определение количества кластеров

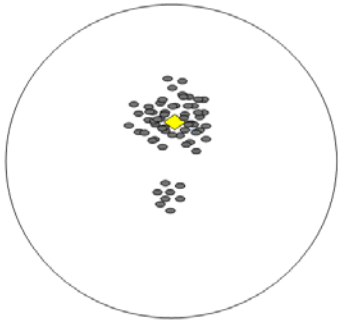




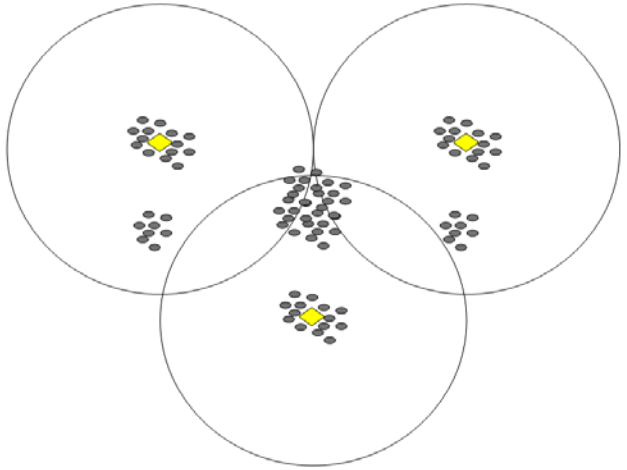
$$f(k) = \frac{\sum_{p=1}^{k} AVGu_{ij}}{k} -> MAX$$

Сокращение времени кластеризации

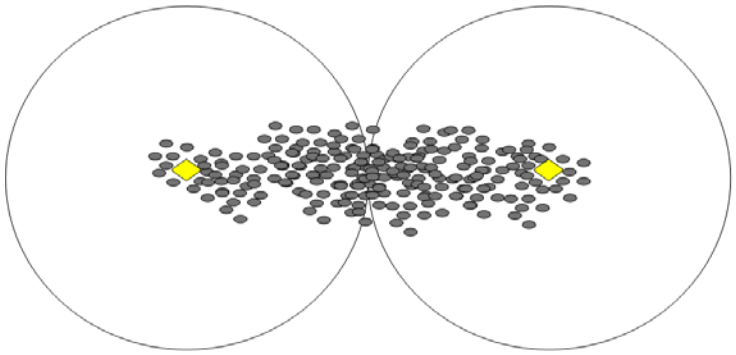
Случай 1 - Укрупненный кластер не является плотным скоплением векторов



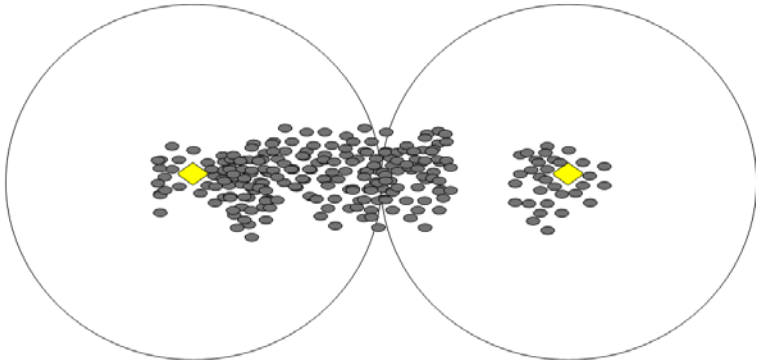
Случай 2 - Укрупненные кластеры разделяют между собой один или несколько кластеров



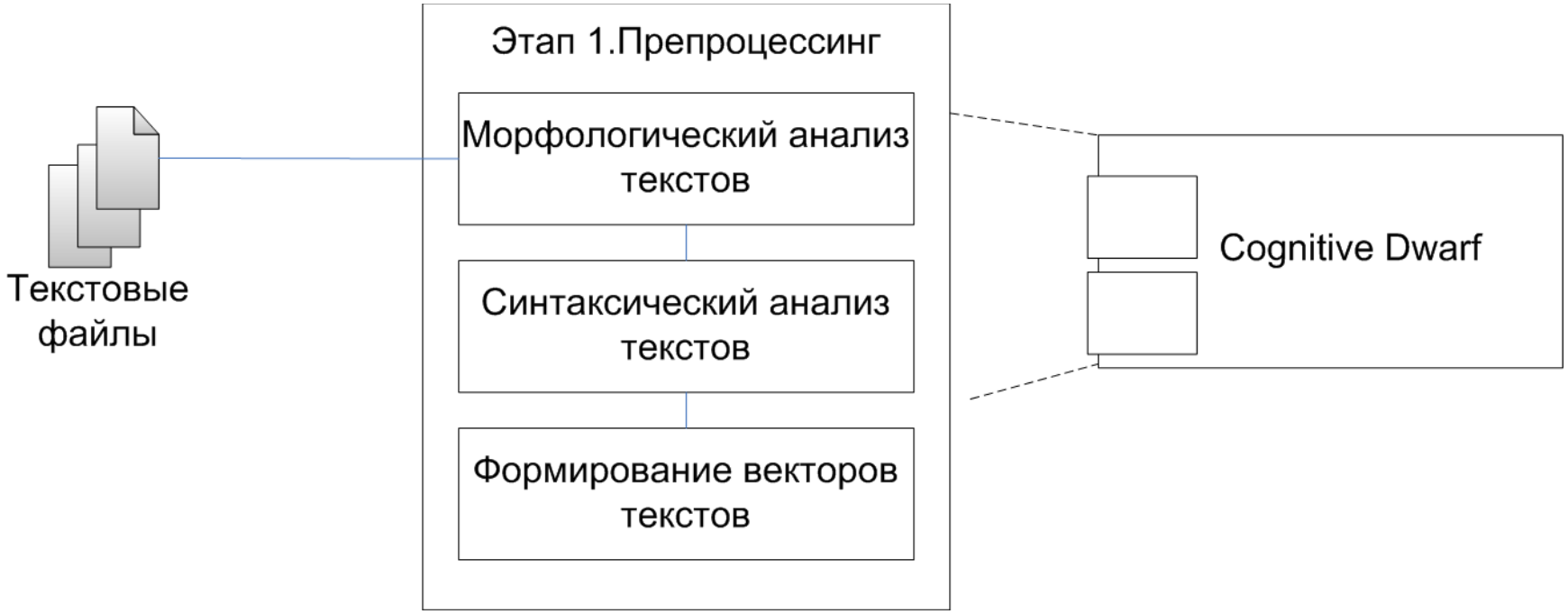
Случай 3 - укрупненные кластеры являются одним кластером

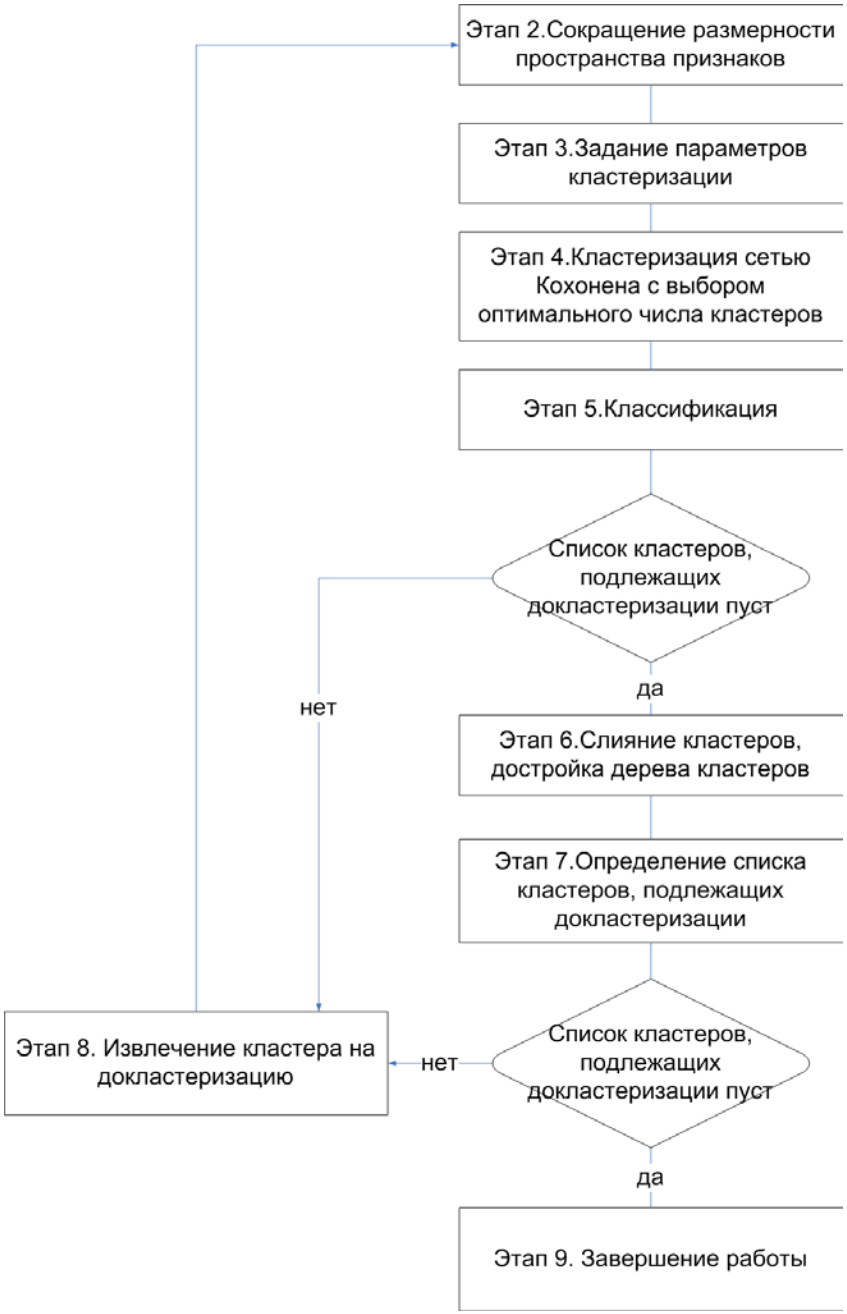


Случай 4 - укрупненный кластер содержит часть другого кластера



Алгоритм иерархической кластеризации - этап 1





Этапы 2-9

Расстояние между текстами:

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\sum_k (x_{ik})^2} * \sqrt{\sum_k (x_{jk})^2} - \sum_k (x`_{ik} * x`_{jk})$$

Сокращение размерности:

{(car), (truck), (flower)} --> {(1.3452 * car + 0.2828 * truck), (flower)}

Формула Эккарта — Янга:

$$X_k = U_k D_k V_k^*$$

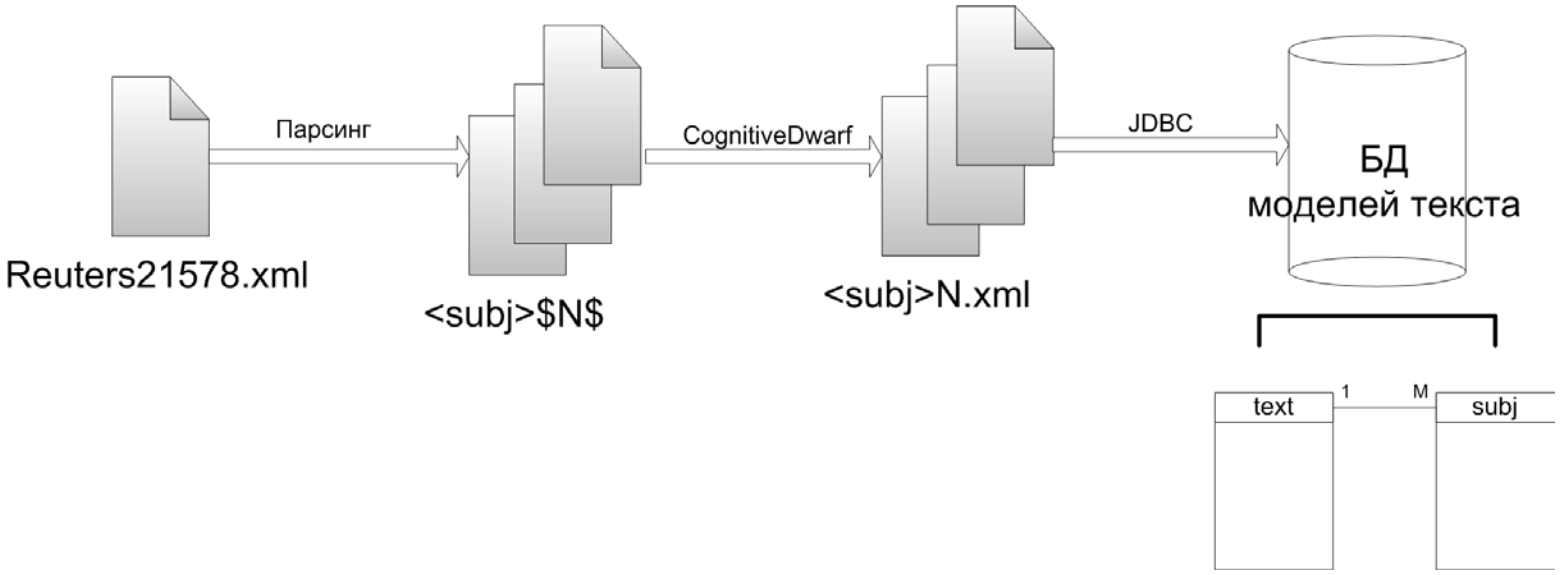
Преимущества методики

- Использование модели текста, ориентированной на семантику;
- Вычислительная сложность на каждом уровне иерархии O(Nmt) в отличие от агломеративных методов;
- Автоматическое определение числа кластеров на каждом уровне иерархии;
- Кластеризация ограниченного количества текстов на каждом уровне иерархии;
- Устранение ошибок классификации с помощью процедуры слияния кластеров
- Построение таксономии, ориентированной априорные представления о ней пользователя:
- Гибкость построения систем поиска за счет использования нечеткой сети Кохонена.

Эксперимент

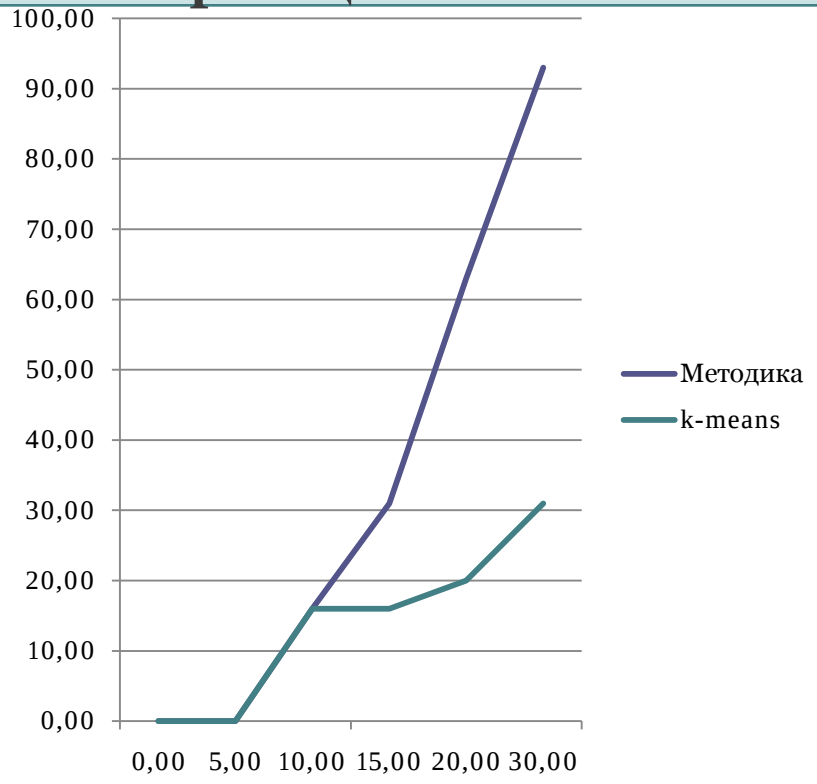
- **Корпус текста:** Reuters-21578
- **Характеристики корпуса:** новости агентства Reuters за 1987 год, 9485 текста, количество кластеров – 57, количество слов от 6 до 1029
- **Опыт:** сравнение с алгоритмом k-means

Схема обработки данных:

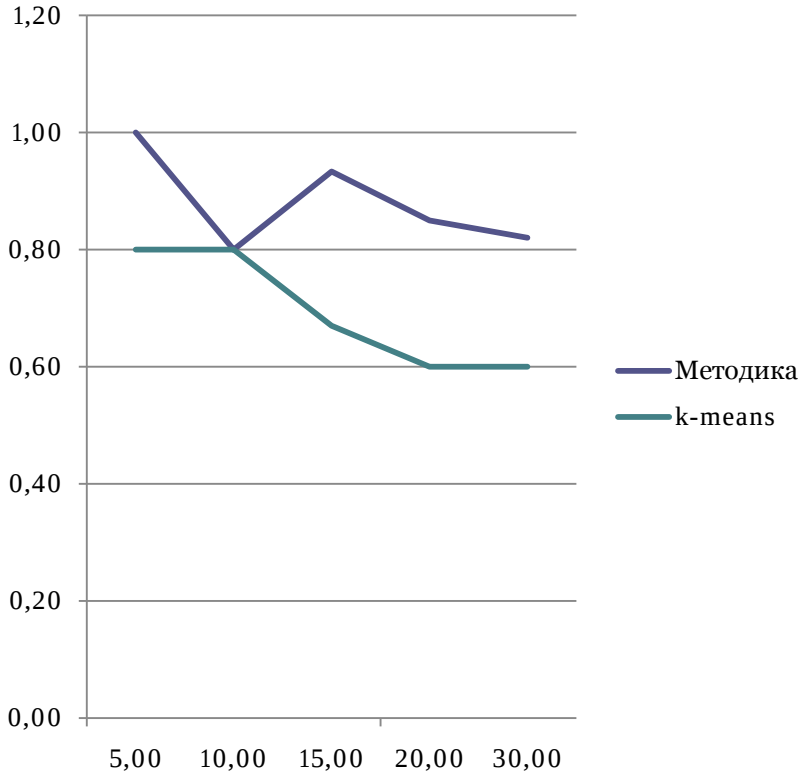


Измеряемые параметры

Время выполнения кластеризации



Качество кластеризации



$$\text{Качество} = \frac{\text{число документов, кластеризованных верно}}{\text{общее число документов}}$$