Интеллектуальные информационные системы

Тема 8

Тематическая кластеризация

Павел Исаакович Браславский
pb@imach.uran.ru
весенний семестр 2006

Классификация без учителя

Неформальная постановка задачи:

Дано множество документов, необходимо разбить его на непересекающиеся подмножества (кластеры) *семантически близких* документов.

На практике: *близкие* в терминах метрики в векторном пространстве

Кластеризация: применение

- Организация результатов поиска
- Улучшение интерфейса при браузинге
- Обработка потока новостей
- Повышение
  - качества поиска
  - скорости поиска

Кластеризация: исходные данные

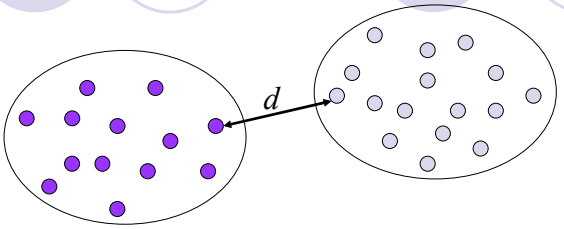
- Данные для вычисления сходства
  - Векторы + метрика
  - Готовая матрица сходства
- Алгоритм
- (Целевое количество кластеров)

Агломеративные иерархические

- Одиночная связь - Single Link
- Полная связь – Complete Link
- Групповое среднее – Group Average

0. Все объекты – одиночные кластеры
1. Находим два ближайших кластера, объединяем в один и т.д.….

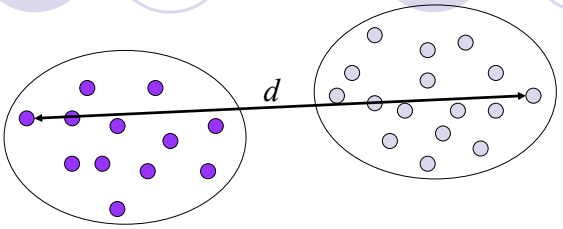
Метод одиночной связи



- Расстояние между классами = расстояние между ближайшими представителями

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x, y)$$

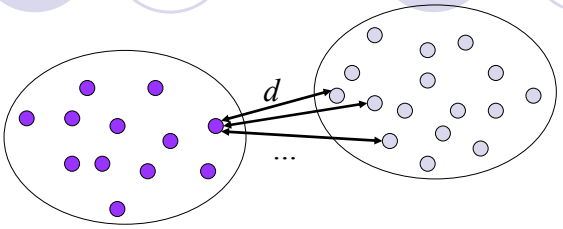
Метод полной связи



- Расстояние между классами = расстояние между наиболее отдаленными представителями

$$d(c_i, c_j) = \max_{x \in c_i, y \in c_j} d(x, y)$$

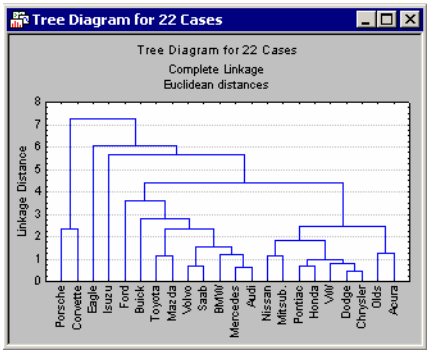
Метод группового среднего



- Расстояние между классами = среднее всех попарных расстояний между представителями классов

$$d(c_i, c_j) = \frac{1}{|c_i||c_j|} \sum_{x \in c_i, y \in c_j} d(x, y)$$

Процесс = дендрограмма



Это только пример из Statistica Help, не относится к кластеризации *текстов*

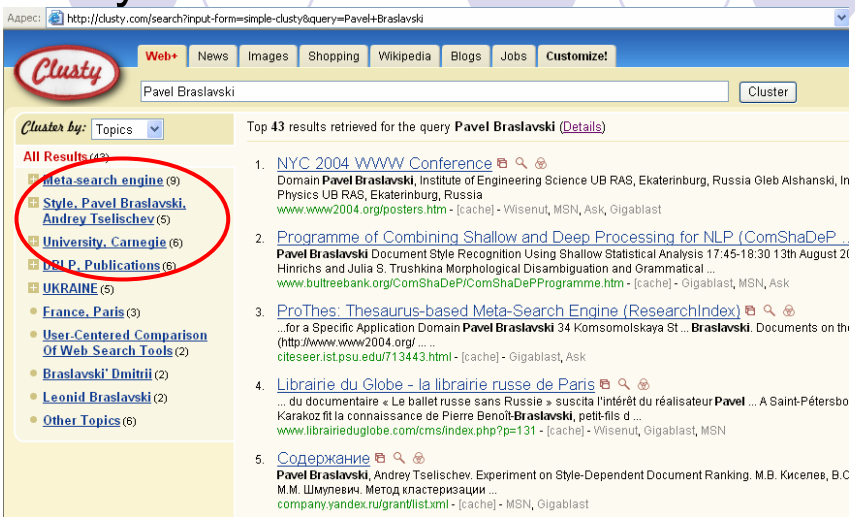
К средних: K-means

- Случайным образом назначить центры классов
- Провести разбиение
- Пересчитать центры
- Провести разбиение
- Пересчитать центры
- …
- Критерий остановки: кол-во итераций, центроиды не сдвигаются, разбиение не изменяется

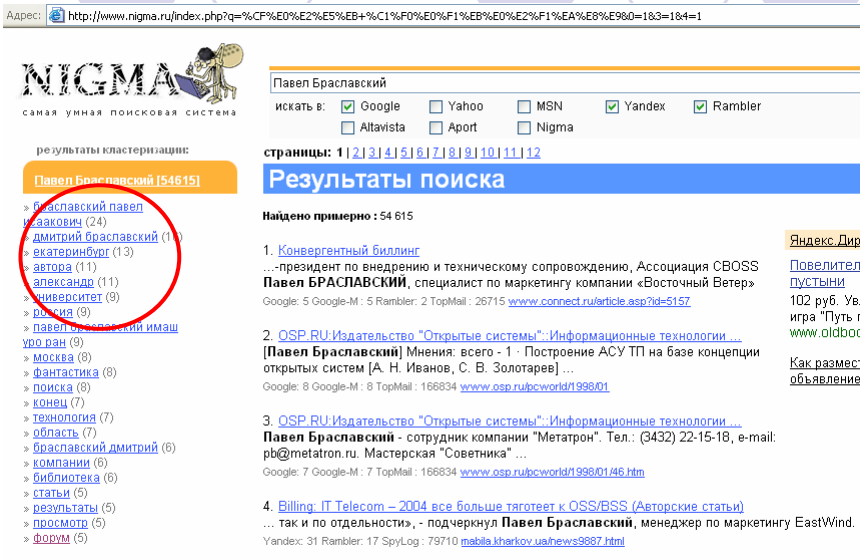
Suffix Tree Clustering (STC)

- Классификация на основе суффиксных деревьев
- Представление документов ≠ bag of words (!)
- Учитывается порядок слов
- Используется для организации результатов поиска на основе сниппетов (а не полных документов)
- Результирующие кластеры могут пересекаться

clusty.com



nigma.ru

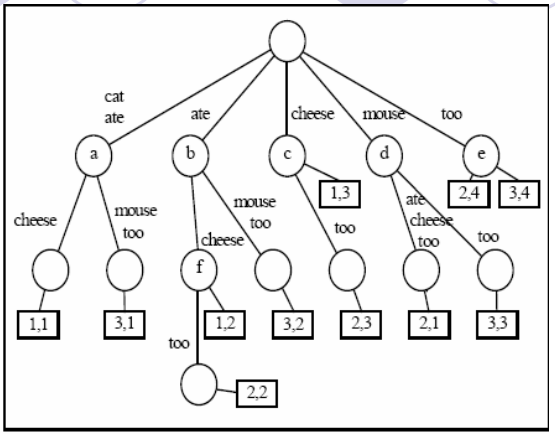


Этапы

1. Очистка
   простейший стемминг, границы предложений, удаление пунктуации и специальных символов
2. Определение базовых кластеров
   - построение суффиксного дерева по набору документов
   - вес кластера B: $s(B)=|B|f(|P|)$, P – длина строки – основы объединения в кластер
3. Объединение базовых кластеров по критерию:

$$|B_m \cap B_n| / |B_m| > 0.5$$
$$|B_m \cap B_n| / |B_n| > 0.5$$

Суффиксное дерево



cat ate cheese, mouse ate cheese too, cat ate mouse too

Граф связности кластеров