

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)

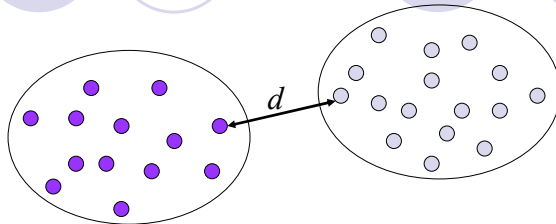
Агломеративные иерархические

- 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....

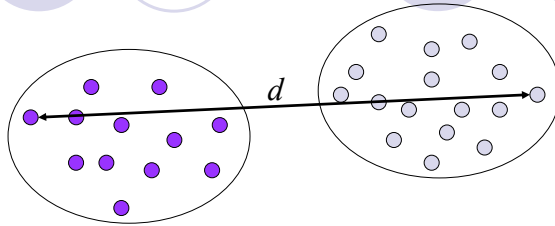
Метод одиночной связи



- Расстояние между классами = расстояние между ближайшими представителями

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

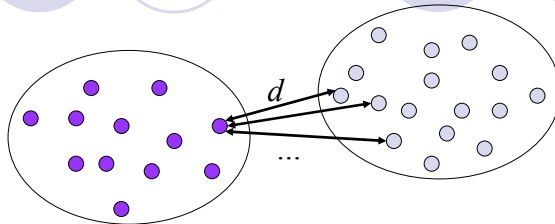
Метод полной связи



- Расстояние между классами = расстояние между наиболее отдаленными представителями

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

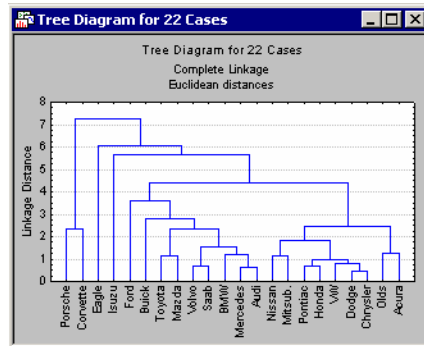
Метод группового среднего



- 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 \neq bag of words (!)
- Учитывается порядок слов
- Используется для организации результатов поиска на основе сниппетов (а не полных документов)
- Результирующие кластеры могут пересекаться

2006

П.И. Браславский - Интеллектуальные ИС

11

clusty.com

Адрес: <http://clusty.com/search?input-form=simple-clusty&query=Pavel+Braslavski>

Web+ News Images Shopping Wikipedia Blogs Jobs Customize!

Pavel Braslavski Cluster

Cluster by: Topics

All Results (43)

- Meta-search engine (9)
- Style_Pavel_Braslavski_Andrey_Tselishev (5)**
- University_Carnegie (6)
- ISI_P_Publications (6)
- UKRAINE (5)
- France_Paris (3)
- User-Centered Comparison Of Web Search Tools (2)
- Braslavski_Dmitri (2)
- Leonid_Braslavski (2)
- Other Topics (6)

Top 43 results retrieved for the query **Pavel Braslavski** (Details)

- NYC 2004 WWW Conference** Domain **Pavel Braslavski**, Institute of Engineering Science UB RAS, Ekaterinburg, Russia Gleb Alshanski, In Physics UB RAS, Ekaterinburg, Russia www.2004.org/posters.htm - [cache] - Wisenut, MSN, Ask, Gigablast
- Programme of Combining Shallow and Deep Processing for NLP (ComShaDeP...)** Pavel Braslavski Document Style Recognition Using Shallow Statistical Analysis 17:45-18:30 13th August 20 Hinrichs and Julia S. Trushkina Morphological Disambiguation and Grammatical ... www.bultreebank.org/ComShaDeP/ComShaDePProgramme.htm - [cache] - Gigablast, MSN, Ask
- ProThes: Thesaurus-based Meta-Search Engine (ResearchIndex)** ...for a Specific Application Domain **Pavel Braslavski** 34 Komsomolskaya St. ... **Braslavski**. Documents on the (<http://www.2004.org/>) ... citeseer.ist.psu.edu/713443.html - [cache] - Gigablast, Ask
- Librairie du Globe - la librairie russe de Paris** ... du documentaire « Le ballet russe sans Russie » suscita l'intérêt du réalisateur **Pavel** ... A Saint-Petersbo Karakoz fit la connaissance de Pierre Benoit-**Braslavski**, petit-fils d ... www.librairie duglobe.com/cms/index.php?p=131 - [cache] - Wisenut, Gigablast, MSN
- Содержание** **Pavel Braslavski**, Andrey Tselishev. Experiment on Style-Dependent Document Ranking. М.В. Киселев, В.С. М.М. Шмулевич. Метод кластеризации ... company.yandex.ru/grantlist.xml - [cache] - MSN, Gigablast

2006

П.И. Браславский - Интеллектуальные ИС

12



результаты кластеризации:

Павел Браславский [54615]

- > [Браславский павел](#)
- > [Исаакович \(24\)](#)
- > [дмитрий браславский \(1\)](#)
- > [екатеринбург \(13\)](#)
- > [автова \(11\)](#)
- > [александр \(11\)](#)
- > [иниверситет \(9\)](#)
- > [васильев \(9\)](#)
- > [павел браславский имаш](#)
- > [уор ван \(9\)](#)
- > [москва \(8\)](#)
- > [фантастика \(8\)](#)
- > [поиска \(8\)](#)
- > [конец \(7\)](#)
- > [технология \(7\)](#)
- > [область \(7\)](#)
- > [браславский дмитрий \(6\)](#)
- > [компани \(6\)](#)
- > [библиотека \(6\)](#)
- > [статьи \(5\)](#)
- > [результаты \(5\)](#)
- > [просмотры \(5\)](#)
- > [форум \(5\)](#)

Павел Браславский

искать в: Google Yahoo MSN Yandex Rambler
 Altavista Aport Nigma

страницы: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#)

Результаты поиска

Найдено примерно : 54 615

1. Конвергентный биллинг

...президент по внедрению и техническому сопровождению, Ассоциация CBOSS
Павел БРАСЛАВСКИЙ, специалист по маркетингу компании «Восточный Ветер»
 Google: 5 Google-M : 5 Rambler: 2 TopMail : 26715 www.connect.ru/article.asp?id=5157

2. OSP.RU Издательство "Открытые системы": Информационные технологии ...
[Павел Браславский] Мнения: всего - 1 · Построение АСУ ТП на базе концепции открытых систем [А. Н. Иванов, С. В. Золотарев] ...
 Google: 8 Google-M : 8 TopMail : 166834 www.osp.ru/ocworld/1998/01

3. OSP.RU Издательство "Открытые системы": Информационные технологии ...
Павел Браславский - сотрудник компании "Метатрон". Тел.: (3432) 22-15-18, e-mail: pb@metatron.ru. Мастерская "Советника" ...
 Google: 7 Google-M : 7 TopMail : 166834 www.osp.ru/ocworld/1998/01/46.htm

4. Billing: IT Telecom – 2004 все больше тяготеет к OSS/BSS (Авторские статьи)
 ... так и по отдельности», - подчеркнул **Павел Браславский**, менеджер по маркетингу EastWind.
 Yandex: 31 Rambler: 17 SpyLog: 79710 mabla.kharkov.ua/news/887.html

[Яндекс.Дир](#)

[Повелитель](#)

[пустыни](#)

102 руб. Ув.

игра "Путь г

www.oldboos

[Как размес:](#)

[объявление](#)

Этапы

1. Очистка

простейший стемминг, границы предложений, удаление пунктуации и специальных символов

2. Определение базовых кластеров

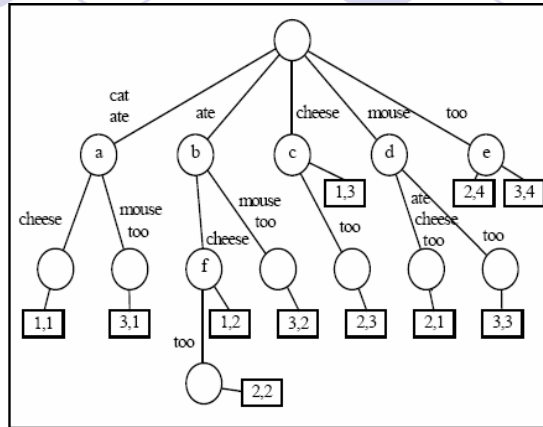
- построение суффиксного дерева по набору документов
- вес кластера $V: s(B)=|B|f(|P|)$, P – длина строки – основы объединения в кластер

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

