

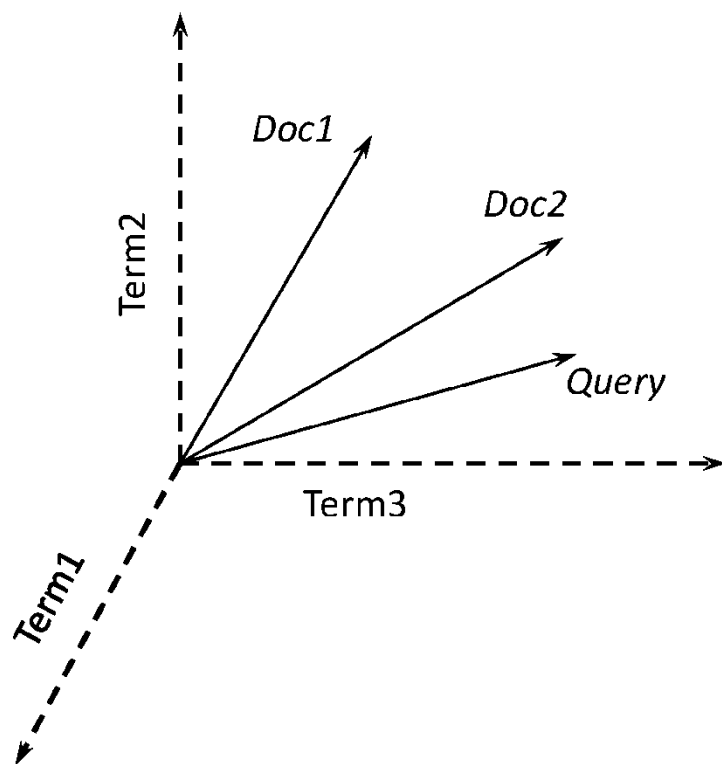
Кластеризация текстов

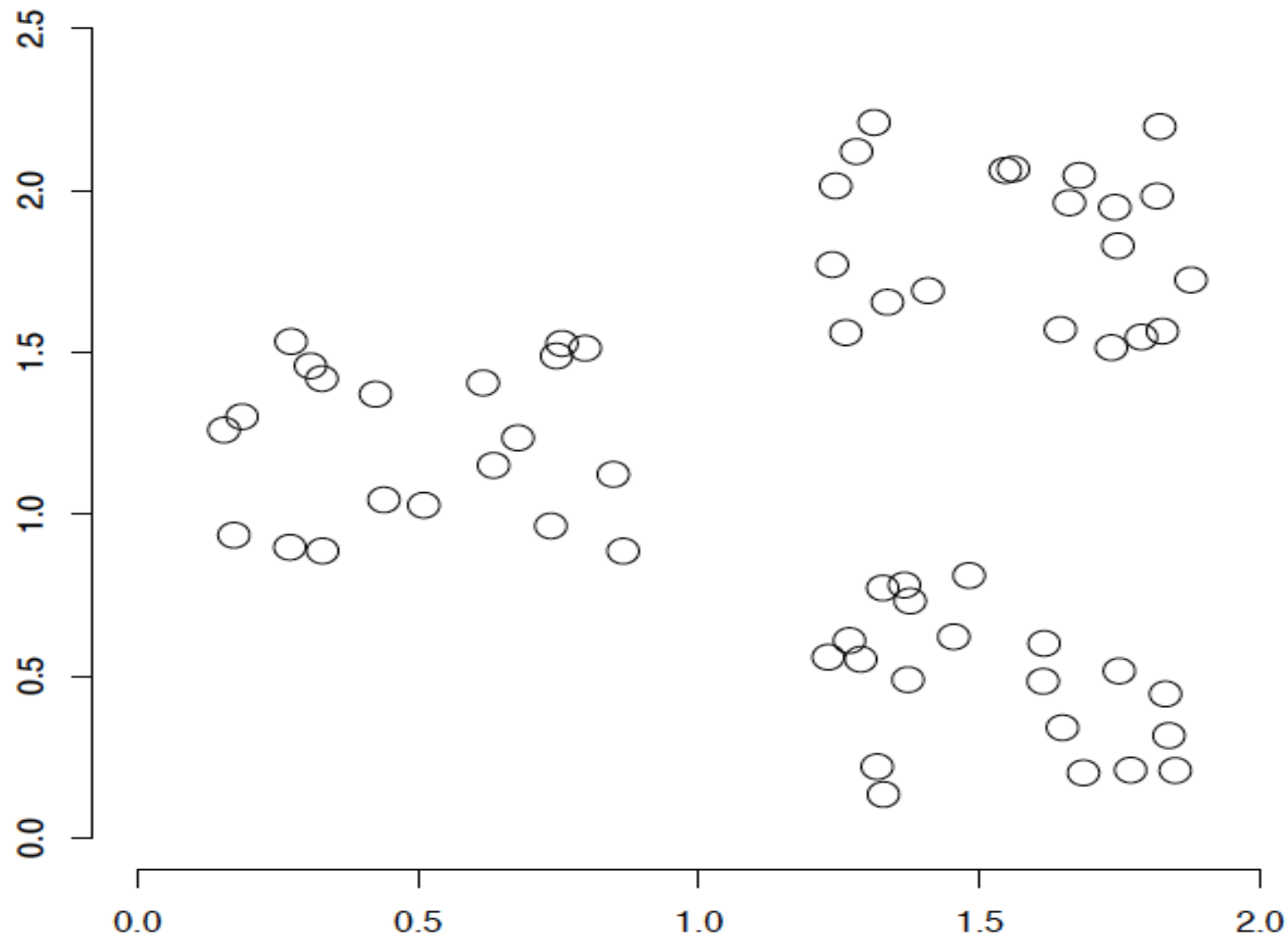
Маннинг и др. Введение в
информационный поиск,
гл.16, 17

Автоматическая кластеризация текстов

- Имеется текстовая коллекция
- Нужно разбить коллекцию на классы близких документов
- Могут быть созданы иерархические классы
- Сейчас: одно из важных средств для визуализации большой выдачи документов при поиске
- Для визуализации важно: хорошее название кластера
- Примеры:
 - Новостные агрегаторы (Яндекс.Новости, Рамблер.Новости, Google.News, Новотека)
 - Кластеризация результатов поиска (Clusty, Нигма)

Прошлые лекции: векторная модель представления документов





Кластеризация выдачи поисковой системы



Фильтр ▼

Как это помогает искать? ▢

- ☒ кластеризация данных
- ☐ кластерный анализ
- ☐ кластеризация документов
- ☒ кластеризация
 - ☐ алгоритм кластеризации
 - ☐ метода кластеризации
 - ☐ алгоритмы кластеризации
 - ☐ методы
 - ☐ интернет магазины
 - ☐ Русскоязычные сайты

Фильтровать

Со всеми: ☐ сбросить

☒ выбрать ☒ исключить

[Интернет](#) [Картинки](#) [Книги](#) [Музыка](#) [Математика](#) [Мини-игры](#)

кластеризация

☐ В найденном

☐ в Москве

Поисковики

Язык

С

193 тыс. результатов.

1. [Кластерный анализ — Википедия](#)

Теперь возникает вопрос устойчивости принятого кластерного решения. По сути, проверка устойчивости **кластеризации** сводится к проверке её достоверности. ...

[Найти слова](#) | <http://ru.wikipedia.org/wiki/%CA%EВ%E0%F1%F2%E5%F0...> 108 Кб

Понравился поиск? Сделай Нигма.рф [поиском для FireFox](#) ! x

2. [Кластеризация](#)

Кластеризация. Материал из MachineLearning. Перейти к: навигация, поиск. ... 2 Формальная постановка задачи **кластеризации**. 3 Ссылки. ...

[Найти слова](#) | www.machinelearning.ru/...dex.php?title=Кластеризация 47 Кб

3. [кластеризация — Викисловарь](#)

1. группировка, разбиение множества объектов на непересекающиеся подмножества, кластеры, состоящие из схожих объектов ♦ Во всех этих случаях может применяться иерархическая **кластеризация**, когда крупные кластеры дробятся на более мелкие... ..

[Найти слова](#) | ru.wiktionary.org/wiki/кластеризация 46 Кб

Справка

[Кластерный](#)
[подробнее](#)

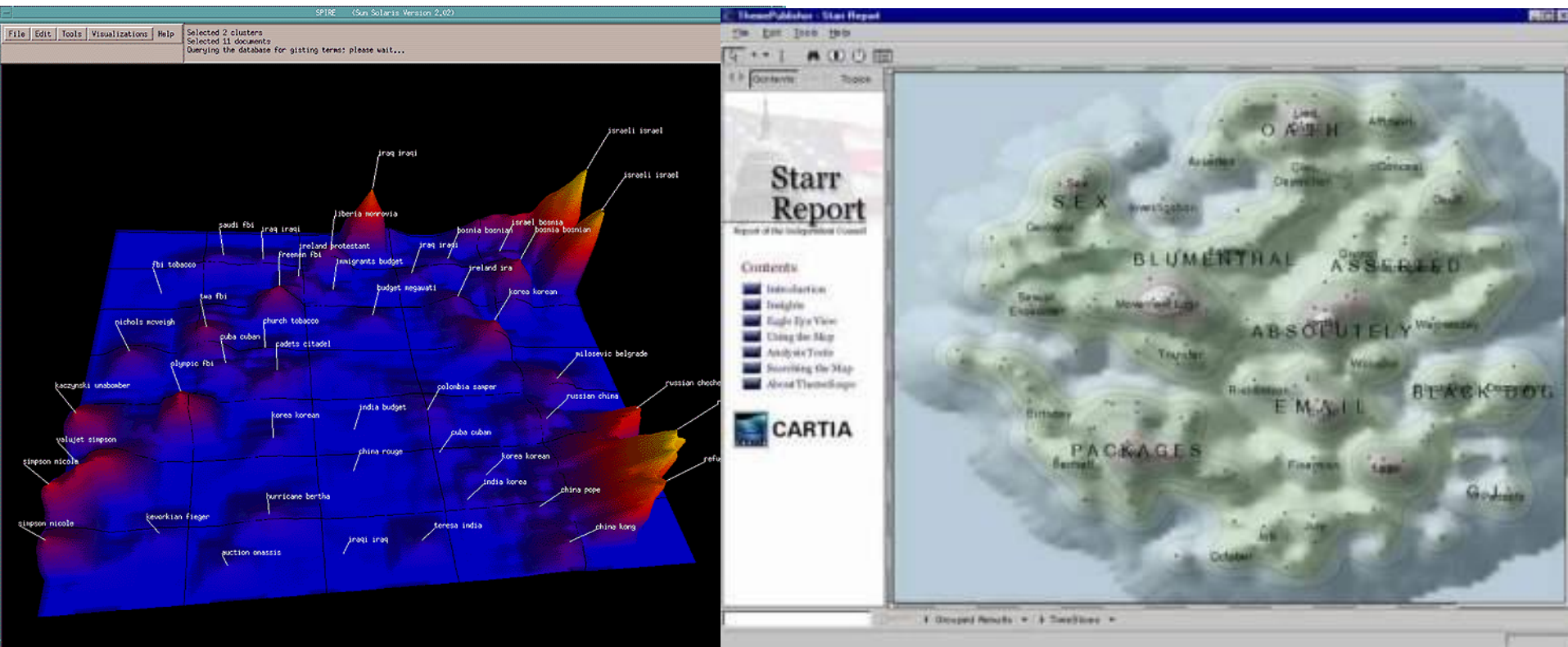
Ищи
на Н

БЕО

Яндекс

Для визуализации тематик в коллекции документов

- Wise et al, “Visualizing the non-visual”
- ThemeScapes, Cartia
 - [Mountain height = cluster size]



Новостные агрегаторы

[illegible]

Google Новости Россия

Персональная версия - **Главные новости**

Поиск

- Главная новости
- В мире
- Россия
- Бизнес
- Наука и техника
- Экология Подмосковья
- Sci/Tech (U.S.)
- Спорт
- Культура
- Здоровье

> Любое содержание
Заголовки
Изображения

Поиск новостей Поиск в Интернете Расширенный поиск новостей Настройки

Обновление 6 мин. назад

Главные новости

В деле об убийстве Анны Политковской появились новые фигуранты

Радиостанция ЭХО МОСКВЫ • 18 мин. назад

Спустя три года после убийства журналистки Анны Политковской, дети считают, что следствию дан новый и последний шанс найти заказчиков и убийц. Время уходит и надежда на раскрытие дела все меньше, сказала Вера Политковская сегодня на пресс-конференции. Напомним, что ранее Верховный суд вернул...

В деле Политковской появились "новые лица" - коллеги журналистки

RIA Новости

"Новая газета": по делу Политковской есть новые подозреваемые

Газета.RU

Интерфакс - Коммерсант - Общая Газета.RU - ФИНАМ FM

Еще похожие статьи: 26 ..

Как студенты СФУ прошли практику на Сааяно-Шушенской ГЭС

Комсомольская правда • 57 мин. назад

Сааяно-Шушенская ГЭС стала не только местом, где произошла страшная авария, но и своеобразным учебным полигоном. Сааяно-Шушенская ГЭС стала не только местом, где произошла страшная авария, но и своеобразным учебным полигоном. студенты Сааяно-Шушенского филиала Сибирского федерального Университета

Братская нагрузка Сааяно-Шушенской ГЭС

Газета RU


Комиссия проверит готовность СУПЭС к строительным работам зимой

RIA Новости

IA REGNUM - Утро.RU - Росбалт.RU - Российская Газета

Еще похожие статьи: 1 292 ..

Суд отклонил жалобу одного из



Жидлопармент Москвы объясняет в связи с делом о мошенничестве

Lenta.ru • 33 мин. назад - все статьи (67) >

Нобелевскую премию по физике получили ученые из США и Англии

Утро.RU • 22 мин. назад - все статьи (479) >

Сообщения, которые планируют освещать РИА Новости 6 октября

RIA Новости • 05.10.2009 - все статьи (9) >

Communication pioneers win 2009 physics Nobel

Reuters • 10 мин. назад - все статьи (820) >

Кадыров против Орлова: заседание продолжается

Радио Свобода • 15 минута назад - все статьи (92) >

Момент «Разбойник и колхозница»: вернут на место 5 декабря

Полит.ру • 2 часа. назад - все статьи (20) >

В России началась промышленная набортка вакцин против свиного гриппа


Интерфакс • 48 Мин. назад - все статьи (260) >

Корейский депутат рассказал, когда Ким Чен Ир откажется от власти

Росбалт RU • 1 час назад - все статьи (339) >

Новости

Дмитрий Медведев	Гус Хиддинк
Владимир Путин	Юрий Лужков
Кристина Орбакайте	Петр Сумин




[Интернет](#)
[Новости](#)
[Картинки](#)
[Видео](#)
[Top100](#)
[Товары](#)
[Визитки](#)
[еще ▾](#)

[Выпуск России](#) | [Украина](#)

[ГЛАВНОЕ](#)
[КАРТИНА ДНЯ](#)
[КАРТИНА НЕДЕЛИ](#)
[РОССИЯ](#)
[МИР](#)
[БИЗНЕС](#)
[РЫНКИ](#)
[ТЕХНОЛОГИИ](#)
[НАУКА](#)
[СПОРТ](#)
[КИНО](#)
[КУЛЬТУРА](#)
[МУЗЫКА](#)
[ЖИЗНЬ](#)
[АВТО](#)
[ЖЕНСКИЙ КЛУБ](#)
[ЗДОРОВЬЕ](#)
[ЗВЕРИНА И СЕКС](#)

ГЛАВНОЕ




Инфляция в РФ остается нулевой второй месяц подряд

Инфляция в России в сентябре была нулевой. Об этом говорится в сообщении Росстата. За период с января по сентябрь 2009 года инфляция составила 0,1 процента.

Вы заглянули рост цен на продукты в последние месяцы? Проголосуйте

- Верховный суд сократил срок убийце Анны Бельской
- В деле Анны Политковской появились «новые лица»
- Дума поддержит выделение бюджетных средств АвтоВАЗу
- Японцы создали работавший на метаноле спиртелефон
- Хиндики раскрыл преималиные сберной
- Суд подтвердил отказ в компенсации пострадавшему от Енисеюга
- Правительство сократит армию киноинкоки
- Старукини МВД провели обыски в департаменте химической политики Москвы
- Монголии отказ Канаде обещанное России нсторонение

РОССИЯ



Бензин подешевел


Суд принял закончик отказать зарегистрировать брак москвичек

В Санкт-Петербурге прошел ночной истор

Виктор Ерофеев ответил на «форменный доносок» Филологос

«Рабочего и колонизинки» вернут на место 5 декабря

МИР



Власти Китая решили создать крупнейше мадинокосини


Французские субмарини оснастат протитоперленными системами

Сторонники Януковича разблорковали Веруюую Раду

Хункороксий депутат назвал раск уоода Ким Чен Ира

Индонезия постанит на вооружение российские ракеты

БИЗНЕС




«Лукойл» и Казакостан позвали французос на Каспий

В США «обанкротились» сразу три банка

Societe Generale увеличилнат капитал на 4,8 миллиарда евро

ВБ приступит к созданию кодовой госиперини

РЫНКИ



9 октября закончилось прини есчетов на XII Конкурс годовойс отчетов

Российские Фондовые торги открьлись ростом

Доллар упал ниже 30 рублей

Торги в Азии проодат разнонаправленно

ПОГОДА: МОСКВА

+7°
 +2°
 +8°

[Днев](#)
[Ночью](#)
[Утром](#)

КУРСЫ ВАЛЮТ

ЦБ (66.00)	Почта	Продажа
\$ 30,0785	29,70	30,05
€ 40,0259	43,70	44,10

[Наши новостные каналы](#)

[illegible]

Понятие сходства/расстояния

- Идеал: семантическое сходство
- На практике: статистическое сходство
 - Косинусное сходство
 - Документы как вектора

Алгоритмы кластеризации

- Плоские алгоритмы
 - Обычно начинаются со случайного разбиения
 - Итеративное уточнение
 - *K средних (K means)*
- Иерархические алгоритмы
 - Снизу-вверх, аггломеративный
 - (Сверху-вниз, разбиение)

Жесткая и мягкая кластеризация

- Жесткая кластеризация: Каждый документ принадлежит только к одному кластеру
 - Легче выполнить
- Мягкая кластеризация: Документ может принадлежать более, чем к одному кластеру
 - Полезно, но сложнее разбивать и использовать
- Далее только жесткая кластеризация

K-Means: основные идеи

- Рассматривает документы как вектора с вещественными значениями
- Кластеры базируются на понятии центроида точек в кластере, c :

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Присваивание документов к кластеру базируется на сходстве с текущими центроидами кластеров

Алгоритм K-Means

- Выберем K случайных документов $\{s_1, s_2, \dots, s_K\}$ как исходное множество (seeds) – это как бы центроиды будущих кластеров.

- До тех пор пока кластеризация не сойдется (или другой критерий остановки):

Для каждого документа d_i :

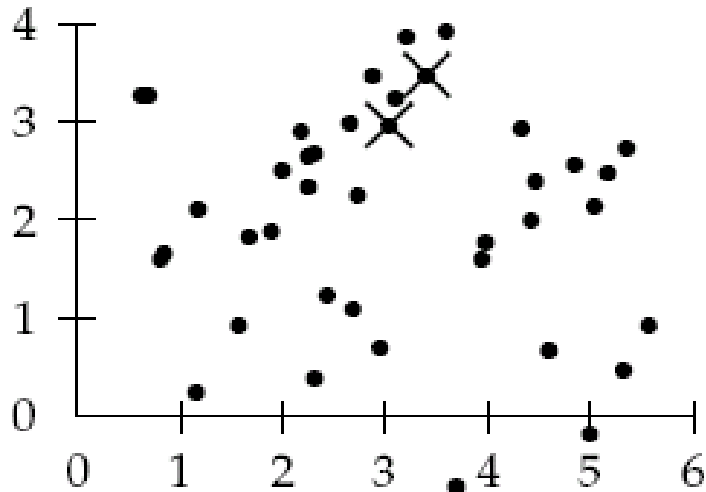
Присваиваем d_i к кластеру c_j такому, что $\text{similarity}(x_i, s_j)$ - максимально.

Затем обновляем множество $\{s_1, s_2, \dots, s_K\}$, заменяем на центроиды текущих кластеров

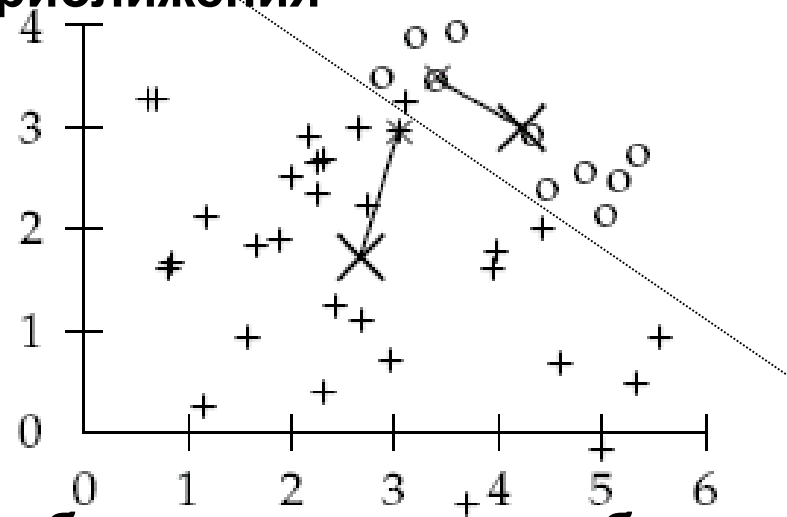
Для каждого кластера c_j

$$s_j = \mu(c_j)$$

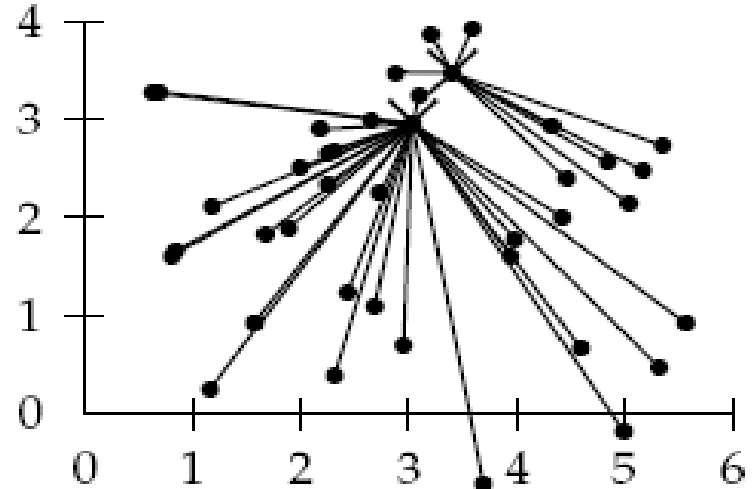
Метод k-means



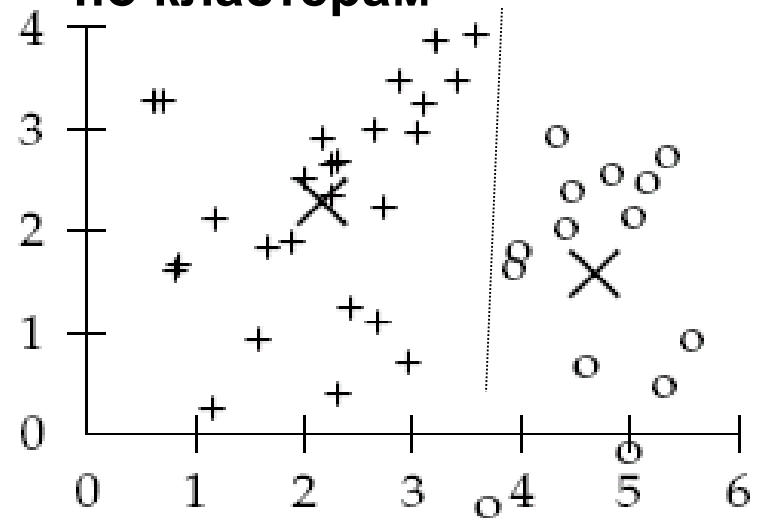
Выбор начального приближения



Выбор следующего приближения для центров

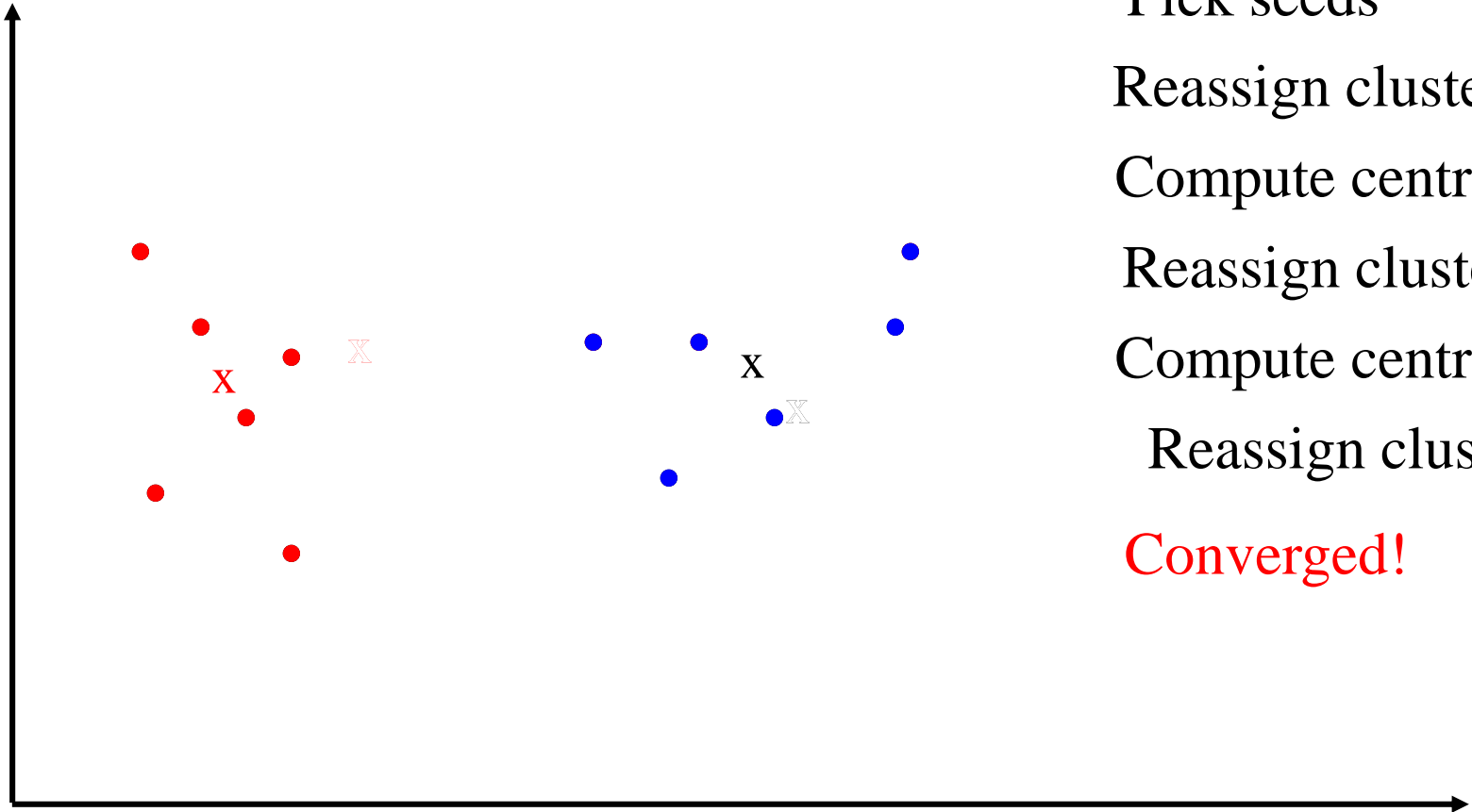


Распределение документов по кластерам



Перераспределение документов

K Means (Пример)($K=2$)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

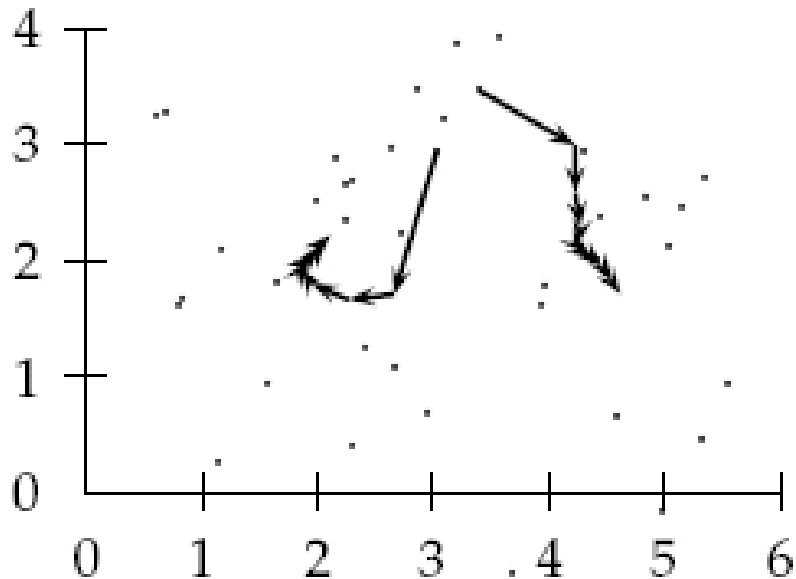
Converged!

Условия остановки

- Несколько возможностей
 - Фиксированное число итераций
 - Не меняется разделение по документов
 - Не меняется позиция центроидов

Сколько кластеров?

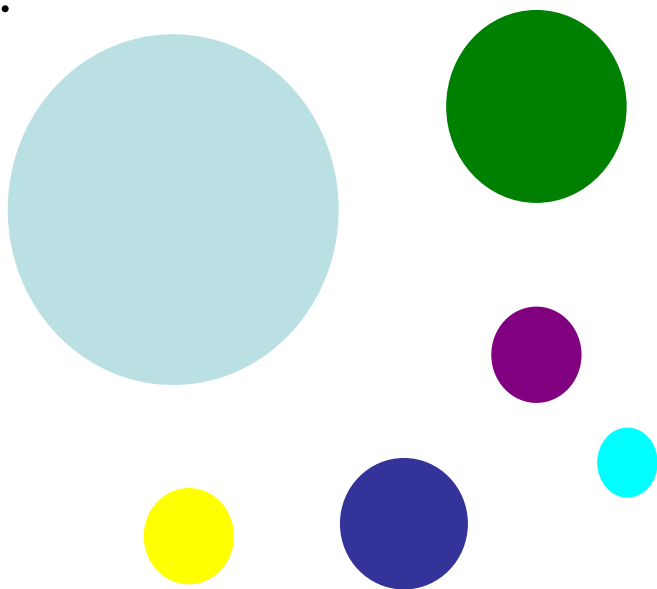
- Число кластеров задается – K
 - Разделяет документы на определенное число кластеров
- Нахождение правильного числа кластеров – это часть проблемы
- Могут использоваться специальные методы подбора k



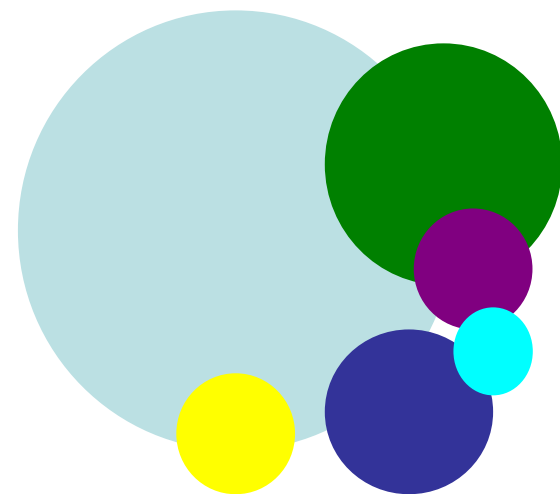
Особенности:

- классический метод — задание фиксированного множества кластеров
- кластеры стремятся быть одинаковыми и «круглыми»

Любой метод кластеризации хорош, если:

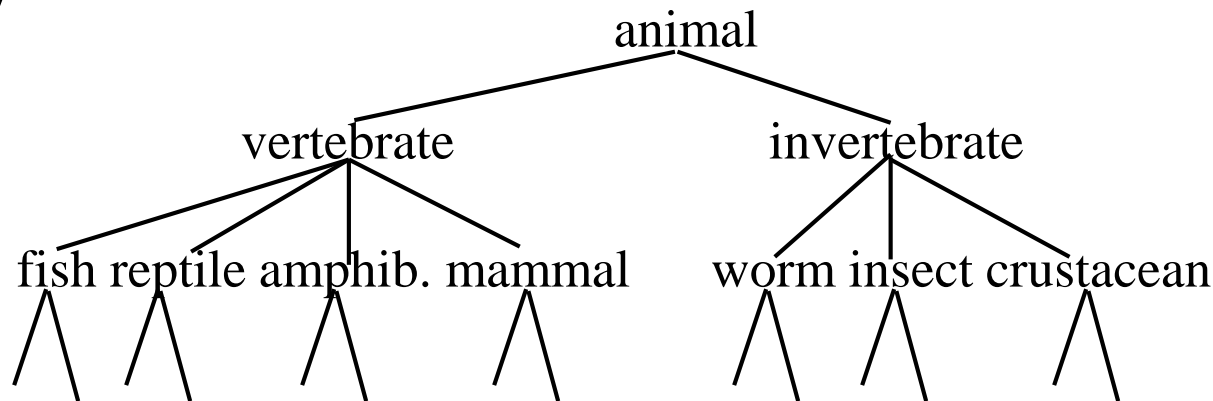


Но на практике часто (по смыслу):



Иерархическая кластеризация

- Строит древообразную иерархическую таксономию (дендрограмму) на основе множества документов

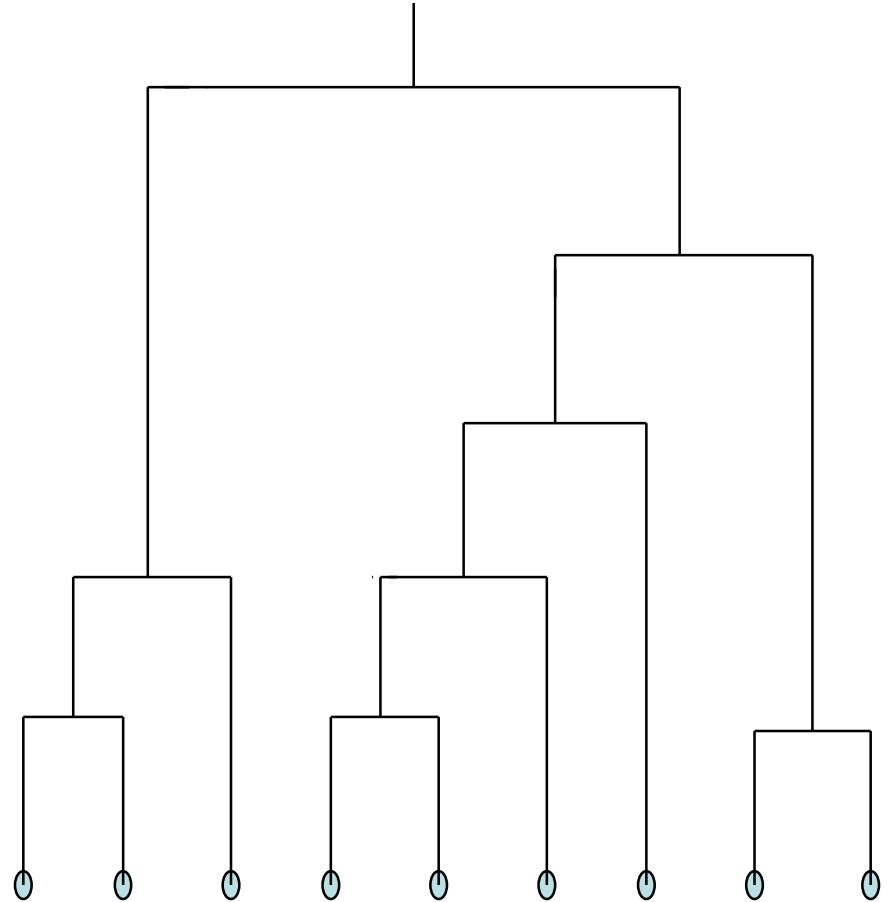


- Метод: рекурсивное применение алгоритма объединения наиболее похожих кластеров

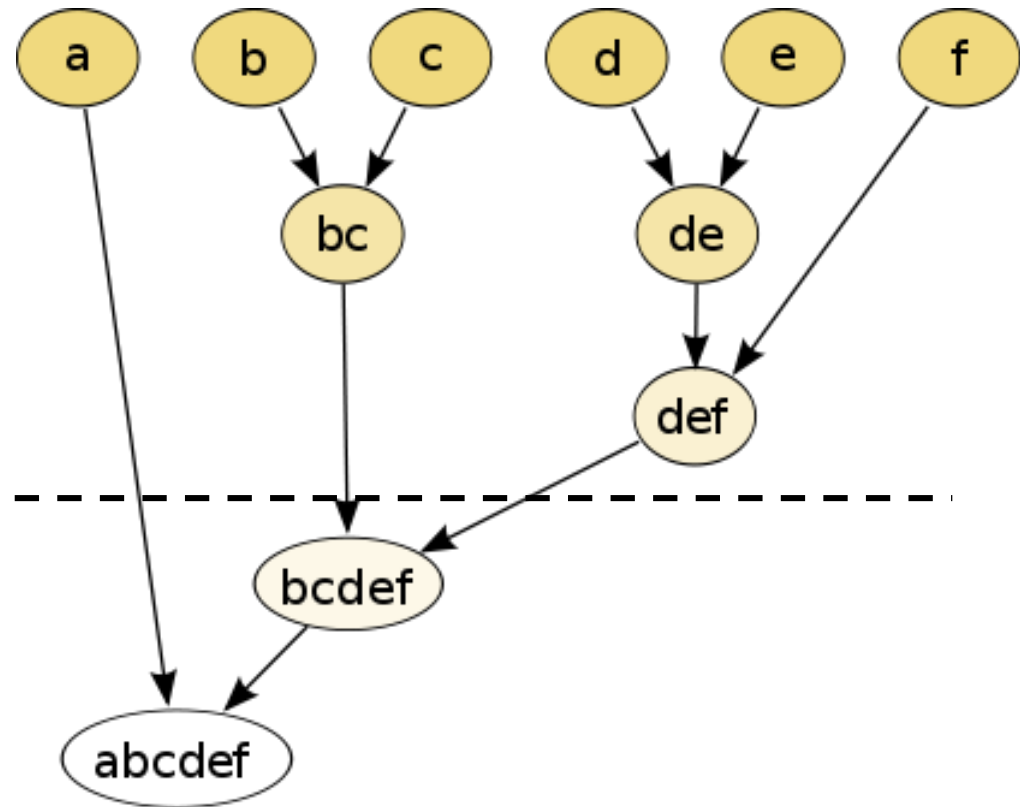
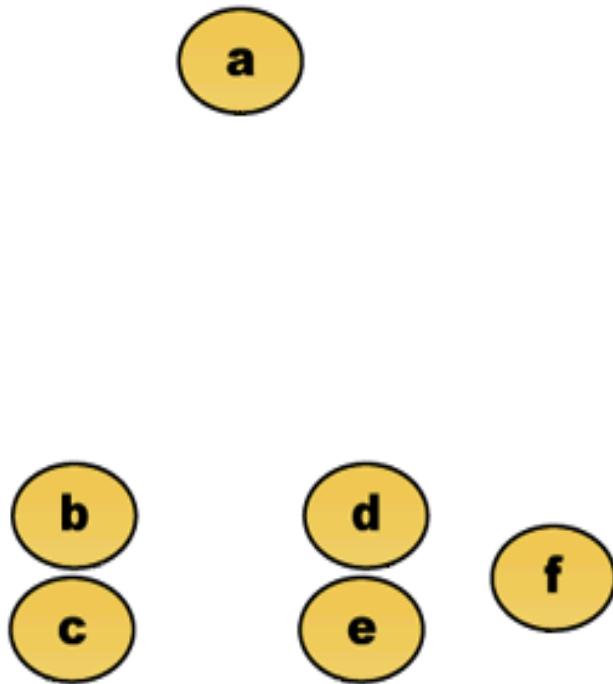
Дендрограмма:

Иерархическая кластеризация

- Кластеризация получается обрезкой дендрограммы на заданном уровне: каждый связный компонент формирует отдельный кластер



Агломеративная кластеризация



Иерархическая аггломеративная кластеризация

- Начинает с рассмотрения документов как отдельных кластеров
 - итеративно объединяет ближайшую пару кластеров, до тех пор пока не останется один кластер.
- История объединения и образует бинарное дерево или иерархию



Ближайшая пара кластеров

- Много способов определения, что такое наиболее сходная пара кластеров
- **Single-link**
 - Сходство по наиболее похожим документам (single-link)
- **Complete-link**
 - Сходство по наиболее непохожим документам
- **Центроид**
 - Сходство по наиболее похожим центроидам
- **Average-link**
 - Средний косинус между парами элементов двух кластеров

Аггломеративная кластеризация: Single Link

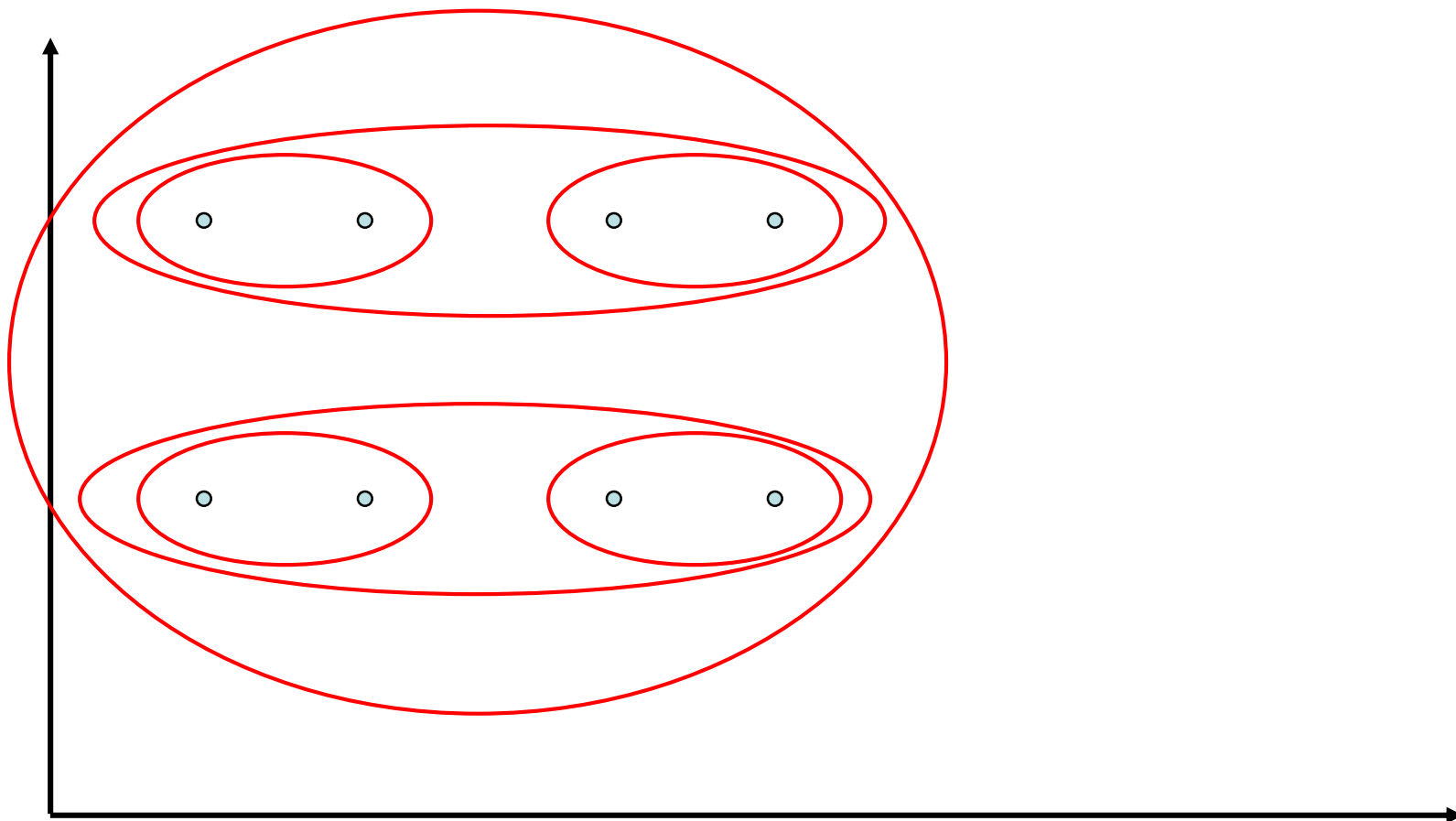
- Использует максимальное сходство пар:

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

- Может породить длинные и тонкие кластеры - цепочки.
- После склеивания c_i и c_j , сходство результирующего кластера к другому кластеру, c_k :

$$\textit{sim}((c_i \cup c_j), c_k) = \max(\textit{sim}(c_i, c_k), \textit{sim}(c_j, c_k))$$

Пример Single Link



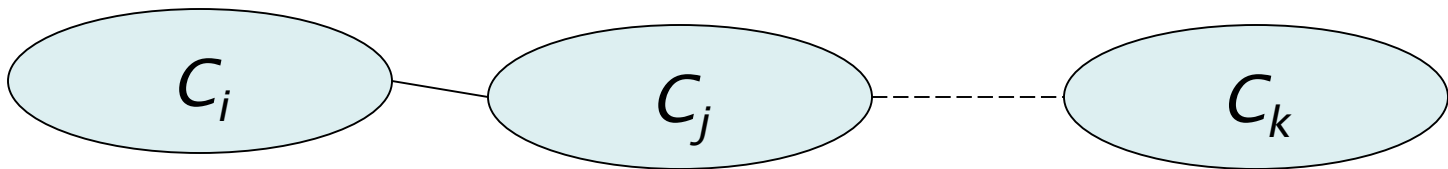
Кластеризация по всем связям (complete link)

- Использует наименее сходные пары:

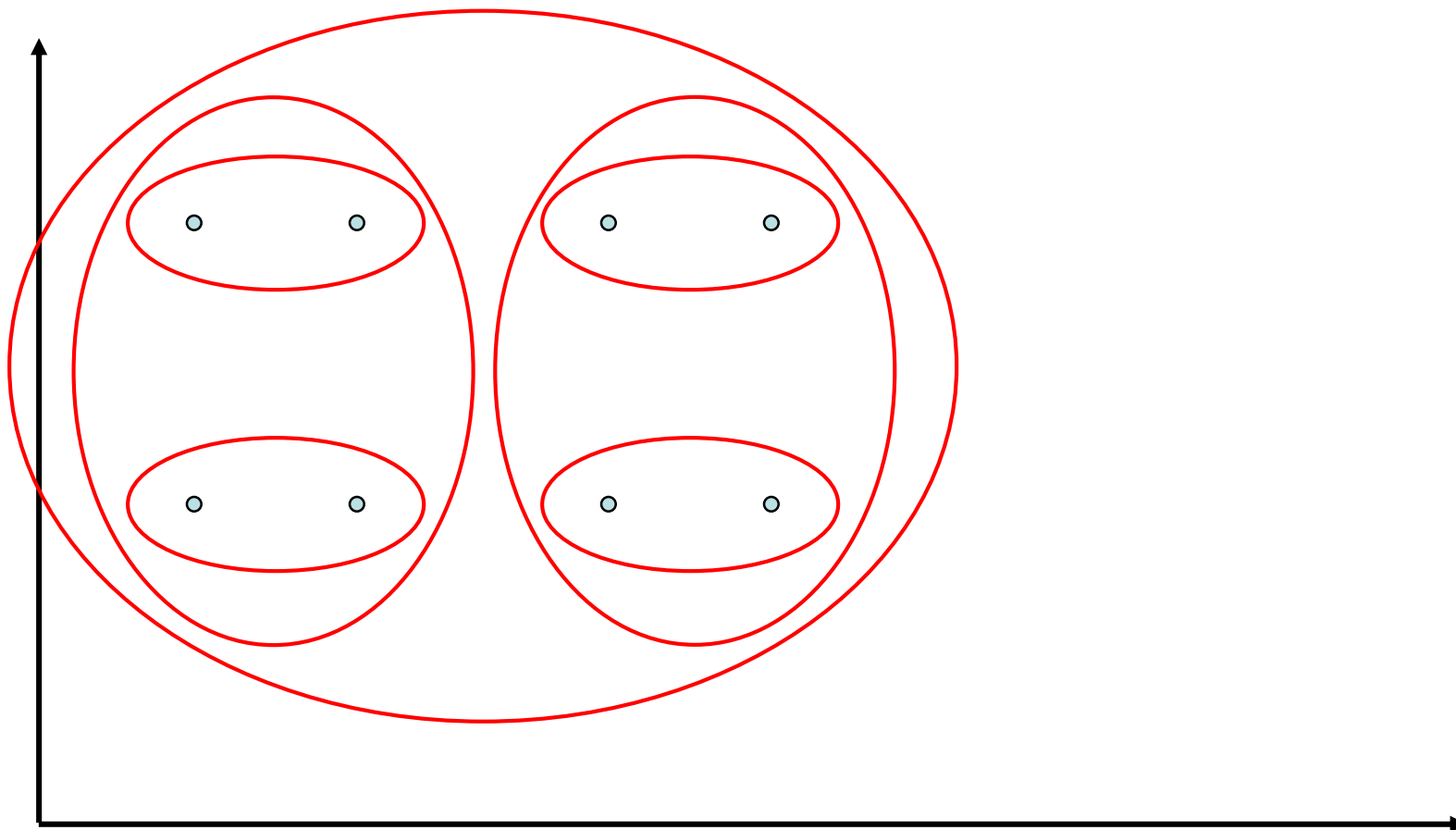
$$\textit{sim}(c_i, c_j) = \min_{x \in c_i, y \in c_j} \textit{sim}(x, y)$$

- Создает более «плотные», сферические кластеры.
- После склеивания c_i и c_j , сходство результирующего кластера с другим кластером, c_k :

$$\textit{sim}((c_i \cup c_j), c_k) = \min(\textit{sim}(c_i, c_k), \textit{sim}(c_j, c_k))$$



Пример Complete Link



Что такое хорошая кластеризация?

- Внутренний критерий: Хорошая кластеризация производит качественные кластеры, в которых:
 - Внутри кластера сходство высокое
 - Между классами – сходство низкое
 - Измеряемое качество кластеризации зависит и от документа, и от меры сходства

Внешние критерии качества кластеризации

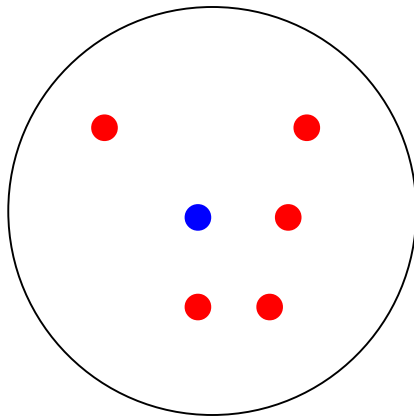
- Качество измеряется способностью кластеризации обнаруживать скрытые классы объектов в эталонных данных (gold standard)
- Оценивает кластеризацию по отношению к «истинным» кластерам (ground truth) ... требует *размеченных данных*
- Предположим, что имеются документы с S правильными кластерами, тогда как наш алгоритм порождает K кластеров, $\omega_1, \omega_2, \dots, \omega_K$ с n_i элементами.

Внешняя оценка качества кластеров: чистота (purity)

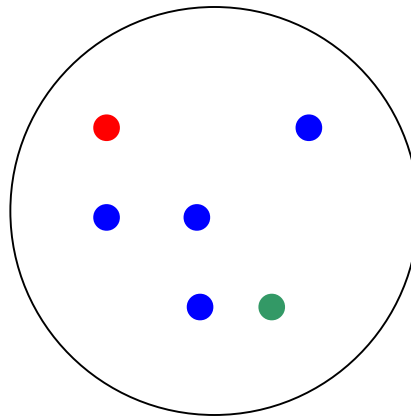
- Простая мера: purity, Отношение между доминантным классом в кластере π_i и размером кластера ω_i

$$Purity(\omega_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C$$

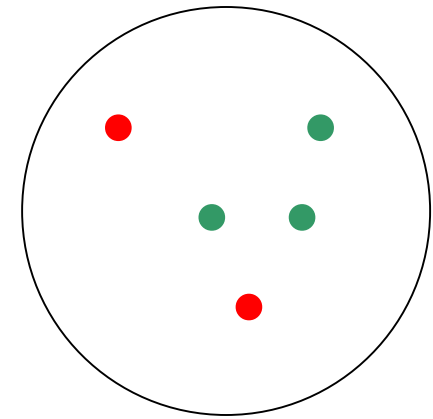
Пример оценки чистоты



Cluster I



Cluster II



Cluster III


Cluster I: Purity = $1/6$ ($\max(5, 1, 0) = 5$) = $5/6$

Cluster II: Purity = $1/6$ ($\max(1, 4, 1) = 4$) = $4/6$

Cluster III: Purity = $1/5$ ($\max(2, 0, 3) = 3$) = $3/5$

Индекс Rand между парами решений,
здесь $RI = 0.68$

Число документов	Тот же кластер в кластериза- ции	Разные кластеры в кластериза- ции
Тот же кластер в эталоне	20	24
Другой кластер в эталоне	20	72



Rand index и F-мера

$$RI = \frac{A + D}{A + B + C + D}$$

Сравним со стандартными полнотой и точностью:

$$P = \frac{A}{A + B}$$

$$R = \frac{A}{A + C}$$

Возможно, лучшей мерой является F-мера

Особенности кластеризации новостей

Особенности обработки новостного потока

- Корпус документов постоянно пополняется
- Временное окно (24-72-120 часов)
- Разные размеры
- Наличие дубликатов, определение первоисточника
- Ошибки при сборе новостных сообщений:
 - ошибки очистки
 - ошибки датировки
- Спамерские технологии источников

Что такое новость?

- Кто?
- Что?
- Где?
- Как?
- Когда?
- Почему?

Традиционное
представление
структуры новости:
перевернутая пирамида



Типичная структура новостного сообщения

2009-10-05 19:41:34 АК&М

А.Чубайс вошел в список лиц, причастных к аварии на СШГЭС

Экс-глава РАО "ЕЭС России" Анатолий Чубайс назван одним из шести человек, которые, по мнению экспертов Ростехнадзора, были причастны к созданию условий аварии на Саяно-Шушенской ГЭС.

Об этом говорится в ... Кроме того, ... Также ...

Напомним, авария на Саяно-Шушенской ГЭС произошла 17 августа.

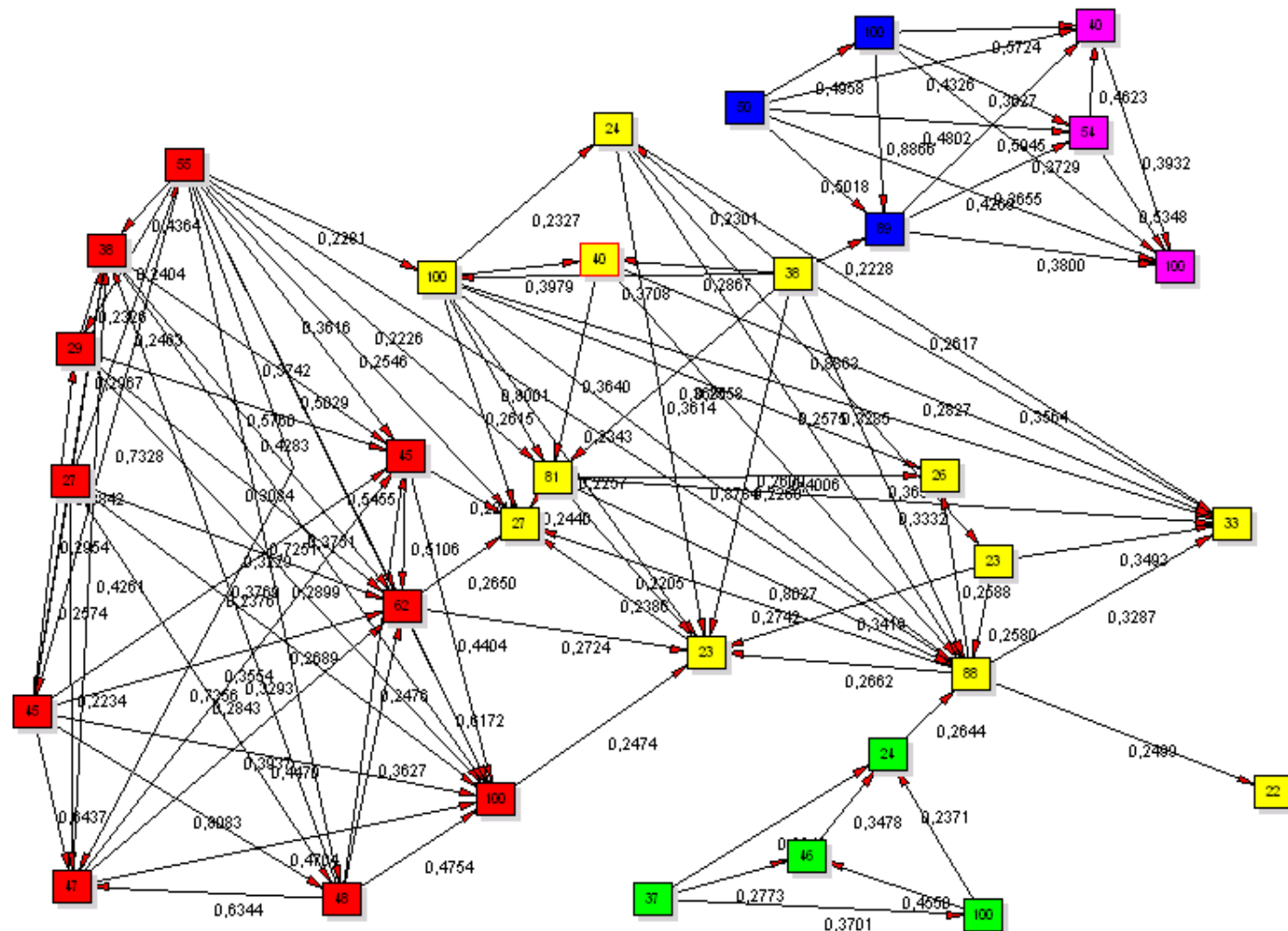
Саяно-Шушенский гидроэнергетический комплекс расположен на реке Енисей на юго-востоке Республики Хакасия в Саянском каньоне - у выхода реки в Минусинскую котловину. Комплекс включает Саяно-Шушенскую ГЭС и расположенный ниже по течению контррегулирующий Майнский гидроузел.

Фрагмент хорошо определенного новостного кластера

0,39	2009.10.05 16:08:54	Игрок сборной Аргентины забил головой с сорока метров	YTP0.ru
0,41	2009.10.05 16:16:00	Футболист забил головой с сорока метров	ИА "Курсор"
0,40	2009.10.05 16:26:00	Аргентинский форвард забил победный гол ударом головой с середины поля (ВИДЕО)	NEWSru.com
0,48	2009.10.05 16:51:11	Аргентинский футболист забил гол ударом головой с 40 метров	Energyland
0,43	2009.10.05 18:29:00	Удар головой с 40 метров завершился голом (видео)	Футбол. Плюс. Хоккей
1,00	2009.10.05 19:57:16	Аргентинский футболист забил гол ударом головой с 40 метров (видео)	Футбол России

Фрагмент сложного новостного кластера

0,24	09.10.05 16:16:00	Шпилька в бок / Износ 98%? - продолжаем работать! :: Общество	chaskor.ru
0,30	09.10.05 19:15:27	Виновников трагедии на ГЭС назовет СКП	YTP0.ru
0,40	09.10.05 19:33:32	Ростехнадзор утвердил методику проверок ГЭС	Ведомости – лента новостей
0,27	09.10.05 19:37:00	Дело об аварии на ГЭС передано Главному управлению СКП РФ	ПРАВО.RU
0,37	09.10.05 19:42:56	Списки лиц, ответственных за аварию на СШГЭС, могут быть расширены.	РБК. Главные новости
0,28	09.10.05 19:52:32	Ростехнадзор утвердил методику проверок ГЭС	Голос России – новости
0,20	09.10.05 19:54:11	Секреты должжителя Чубайса	Svobodanews.ru
0,45	2009.10.05 20:06:14	Ростехнадзор продолжит изучать последствия аварии на СШГЭС	Деловая газета "Взгляд"
0,61	2009.10.05 20:36:04	Ответственных за аварию на Саяно- Шушенской ГЭС станет больше	MI6news.com.ua Украина
1,00	2009.10.05 20:51:33	Ростехнадзор пообещал расширить список виновных в аварии на Саяно- Шушенской ГЭС	РегКорреспонд Украина - Росси



73123415	1
73123416	1
73123418	1
73123419	1
73123420	1
73123425	1
73123427	1
73123429	1
73123430	1
73123432	1
73123433	1
73123436	1
73123437	1
73123438	1
73123440	1
73123441	1
73123442	1
73123443	1
73123445	1
73123446	1
73123447	1
73123448	1
73123449	1
73123450	1
73123451	1
73123452	1
73123453	1

73120528	19
73120543	10
73120336	4
73122751	12
73122754	3

ClusterId = 73120714

Docs = 4

Title = ЕЦ заявляет, что не помогал &quot;Регионам&quot; блокировать Раду

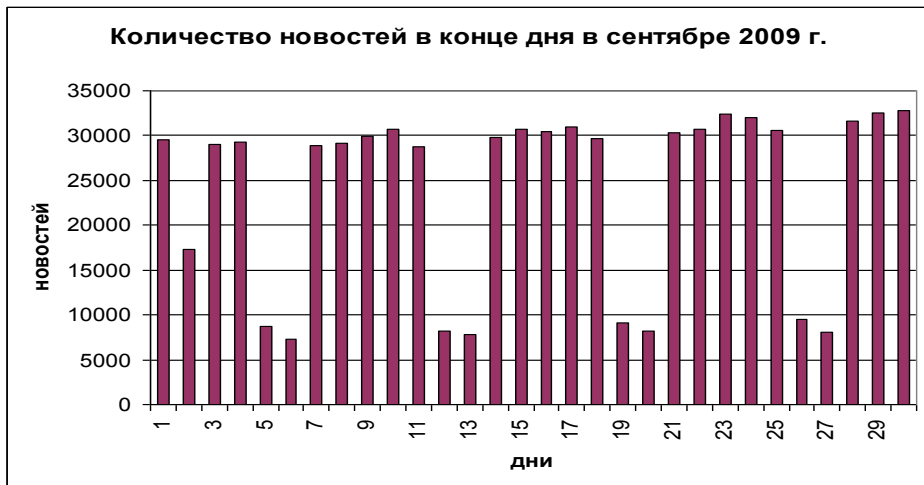
Требования к кластеризации

- Минимизация времени работы
(максимально 15-30 минут)
- Актуальность (главное сообщение)
- Публикация всех кластеров
(не только больших, но и малых)
- Учет перепечаток
- Точность важнее полноты
- Эволюционность кластеризации
- Учет ручного вмешательства:
 - корректировка кластеризации
 - корректировка представления на портале

Данные для исследования

- Данные Romir-2006

Недели	Дни	Размер
Неделя Шеварднадзе	2003.11.20	1752
Обычная неделя	2003.12.03	1715
Неделя выборов	2004.04.02	1809



Эталонное распределение

Возможности редактора:

- Визуализировать, сортировать кластеры по дате, близости к центру и т.п.
- Просматривать близкие кластеры к рассматриваемому (кандидаты на склейку)
- Объединять близкие кластеры
- Разделять существующие кластеры

Внутренние меры

В эталонном
распределении пара
в одном кластере

В эталонном
распределении пара
в разных кластерах

В исследуемом
распределении
пара в одном
кластере

N11	N01
N10	N00

В исследуемом
распределении
пара в разных
кластерах

Точность: $P = N11 / (N11 + N01)$

Полнота: $R = N11 / (N11 + N10)$

F1-мера: $F1 = 2 * P * R / (P + R)$

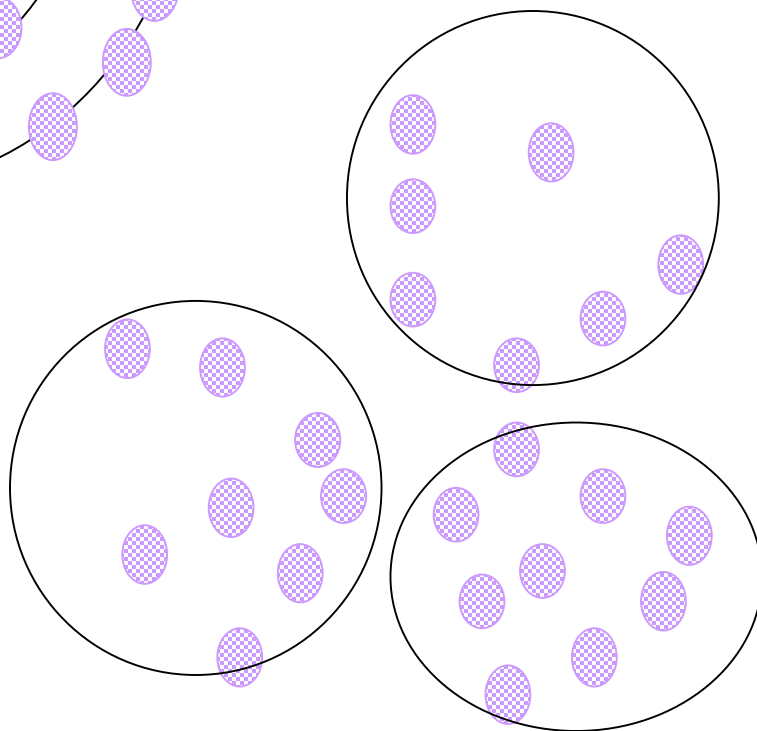
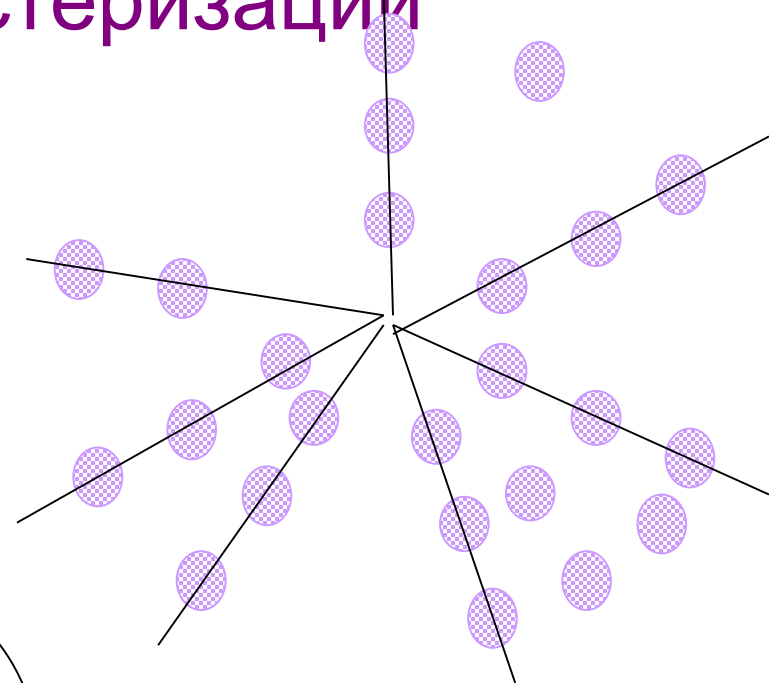
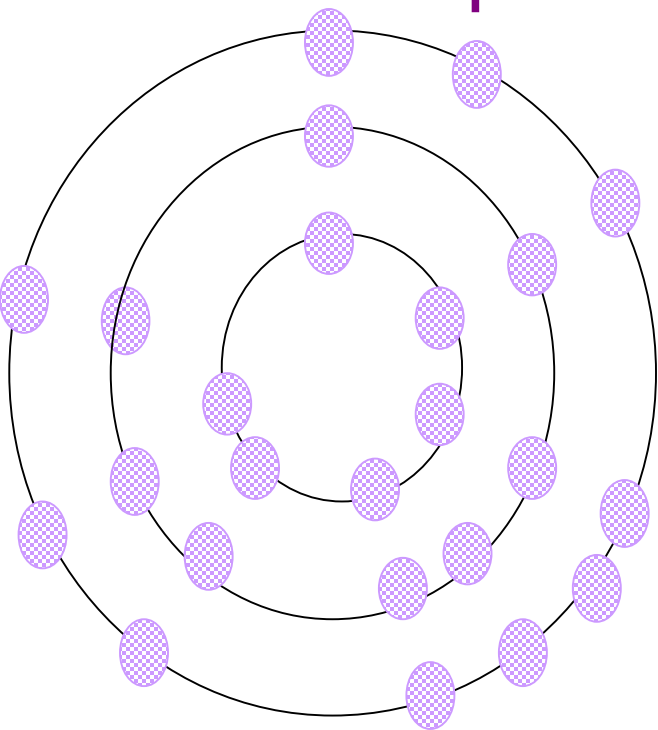
Результаты сравнения F1-меры

Метод	2003-11-21	2003-12-03	2004-04-02	Среднее
FOREL 60:20:20	Result = 0,5282 Ratio = 1,092 Method = center Threshold = 0,32	Result = 0,8383 Ratio = 1,036 Method = center Threshold = 0,34	Result = 0,7364 Ratio = 1,073 Method = average Threshold = 0,24	Result = 0,6890 Ratio = 1,051 Method = center Threshold = 0,34
DBSCAN 60:20:20	Result = 0,5173 Ratio = 1,115 Number = 8 Threshold = 0,28	Result = 0,8648 Ratio = 1,004 Number = 5 Threshold = 0,30	Result = 0,7504 Ratio = 1,053 Number = 3 Threshold = 0,32	Result = 0,6879 Ratio = 1,053 Number = 5 Threshold = 0,30
Modified K-Means 60:15:25	Result = 0,5767 Ratio = 1,000 Iterations = 0,26 Remaining = 0,20 Remaining2 = 0,06 Glue = 0,30	Result = 0,8515 Ratio = 1,020 Iterations = 0,22 Remaining = 0,22 Remaining2 = 0,10 Glue = 0,28	Result = 0,7616 Ratio = 1,038 Iterations = 0,22 Remaining = 0,22 Remaining2 = 0,06 Glue = 0,32	Result = 0,7141 Ratio = 1,014 Iterations = 0,24 Remaining = 0,22 Remaining2 = 0,06 Glue = 0,32
Agglomerative 60:15:25	Result = 0,5470 Ratio = 1,054 Method = center Threshold = 0,26	Result = 0,8250 Ratio = 1,053 Method = average Threshold = 0,18	Result = 0,7549 Ratio = 1,047 Method = min Threshold = 0,30	Result = 0,7003 Ratio = 1,034 Method = center Threshold = 0,26
Agglomerative	Result = 0,5716 Ratio = 1,008 LCH = 40:40:20 Method = average Threshold = 0,22	Result = 0,8685 Ratio = 1,000 LCH = 40:30:30 Method = center Threshold = 0,32	Result = 0,7904 Ratio = 1,000 LCH = 20:50:30 Method = center Threshold = 0,34	Result = 0,7243 Ratio = 1,000 LCH = 40:30:30 Method = center Threshold = 0,30

Различные способы векторизации для метода кластеризации Agglomerative

LCH = (0, 0, 100) (только заголовки)	LCH = (100 ,0, 0) (только леммы)	LCH = (x,0,100-x) (без тезауруса)	LCH = (x,y,100-x-y)
Result =0,4972 Ratio = 1,457 Method = center Threshold = 0,38	Result =0,5767 Ratio = 1,256 Method = min Threshold = 0,38	Result =0,6866 Ratio = 1,055 LCH = 70:00:30 Method = center Threshold = 0,26	Result = 0,7243 Ratio = 1,000 LCH = 40:30:30 Method = center Threshold = 0,30

В качестве заключения: Вариативность кластеризации



Задача: данные

- Есть три предложения:
- Компаниям «Русгидро», «Транснефть» и «Росгеология» предложили подумать о переезде на Дальний Восток.
- "Русгидро", "Росгеология", "Транснефть" получают предложения перенести свои главные офисы на Дальний Восток
- По словам вице-премьера по Дальнему Востоку переезд может как-то затронуть "РусГидро", "Транснефть" и "Росгеологию"

Задача: вопрос

- Найти сходство между предложениями
- - используя булевские вектора $\{0,1\}$ по неслужебным словам
- - как будет происходить объединение предложений в кластер по агломеративному методу центроидов в обоих случаях