Воронежский государственный университет
Факультет компьютерных наук
Кафедра информационных систем

Классификация и кластеризация документов

Что такое классификация?

- Объединение документов или их представлений в одну группу, которая в дальнейшем может рассматриваться и использоваться как единая сущность
- Сами группировки могут определяться заранее либо формироваться алгоритмически
- Процесс группировки может выполняться вручную либо автоматически
- Полезность классификации заключается в том, что элементы группировки могут с большой вероятностью оказаться релевантными

Автоматическая классификация

- Два способа:
  - на основе заранее заданной схемы классификации и уже имеющегося множества классифицированных документов
  - полностью автоматизированная *кластеризация*

Использование автоматической классификации при информационном поиске

- Фильтрация входящих документов
- Сжатие информации (аннотирование, реферирование)
- Расширение запросов за счет характеризующих тематику класса терминов
- Реализация обратной связи по релевантности путем предварительной классификации результатов выборки по классам и выбора пользователем релевантных классов
- Снятие омонимии путем отображения слов в обобщенные концепты
- Увеличение эффективности поиска путем сведения к поиску классов
- (Полу-)автоматическое структурирование порталов – автоматическое формирование тематических каталогов.

Классификация вручную. Эксперимент.

- *10* человек, *5* запросов
- Наиболее частые запросы с существенно различающимися терминами
  - Количество документов по каждому запросу: *15*, *16*, *16*, *11*, *10*
  - *6* человек работали только с *заголовками* и *URL*, остальные *4* – с *полными текстами* документов
- Ставилась задача разбить документы на минимально пересекающиеся классы; ограничение по времени отсутствовало.

Классификация вручную. Результаты эксперимента.

- Наблюдается разброс в результатах классификации документов разными людьми
- Степень подобия между результатами разных людей очень маленькая
- Люди склонны создавать очень маленькие классы
- Использование полного текста документа значительно помогает при классификации, при этом получается меньше классов, но они в большей степени пересекаются
- Человек склонен привязывать результат классификации к контексту
- Классификация не зависит от смысла запроса

Кластеризация

- Основой методов кластеризации является *кластерная гипотеза* (C.J. van Rijsbergen), которая гласит:

  Тесно связанные между собой документы оказываются релевантными по отношению к тем же запросам.

- Кластеризация может быть использована для распределения документов в коллекции по классам (классификации), что позволяет *повысить скорость поиска* документов и *точность ответа*.
- Кластеризация включает в себя две процедуры: *генерацию* кластеров и *поиск* кластеров по запросу пользователя.

Кластеризация. Предыстория.

- *Fairthorne* “The Mathematics of Classification” (1961)
- Эксперименты *Марона* (1961), *Борко* и *Берника* (1963)
- Работы по численной таксономии и ее приложениям при информационном поиске – *Джардайна* (Jardine), *Сибсона*, ван *Рийсбергена*, *Сэлтона* (1970). Кластеризация рассматривалась исключительно с точки зрения эффективности поиска нежели в семантическом аспекте.
- Интерес к кластеризации утрачен в 80-х годах, возрождение интереса в 90-х годах в приложениях, связанных с навигацией и обработкой данных.
- Работа *Спарк-Джонс* по кластеризации терминов.

Кластеризация. Методы.

- Кластеризация – удобный инструмент при работе с документальным пространством, имеющим, как правило, *высокую размерность*.
- Среди *автоматических* методов кластеризации выделяют следующие:
  - Методы иерархической кластеризации
  - Методы кластеризации, основанные на разбиении множеств.
  - Гибридные методы.

Методы кластеризации. Критерии адекватности.

1. Обработка вновь поступающих документов не должна существенным образом изменять результат кластеризации
2. Устойчивость: незначительные ошибки в описании объектов могут вызывать также незначительные изменения в результатах кластеризации
3. Независимость результата кластеризации от исходного порядка на множестве объектов

- Выполнение этих критериев не является обязательным во всех приложениях

Методы кластеризации, основанные на разбиении множеств

- *Целью* является разбиение исходного множества из ***N*** документов на ***k*** кластеров.
- Из них можно выделить
  - глобально оптимальные алгоритмы, которые исчерпывающе перечисляют все разбиения
  - эффективные эвристические методы, например метод *К-средних*.

Метод К-средних

- Документы описываются векторами с вещественными компонентами
- Каждый кластер идентифицируется с помощью *центроида*, который вычисляется как усредненный вектор от всех его элементов:

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

- Далее происходит перераспределение документов по кластерам в зависимости от их расстояния до центроидов кластеров

Алгоритм К-средних

- Задается метрика d для вычисления расстояния между элементами множества.
- Случайным образом выбираются k. зерновых элементов $\{s_1, s_2, \ldots, s_k\}$
- До тех пор пока не выполнится условие остановки:
  - Для каждого элемента x_i:
    - поместить x_i в кластестер c_j такой, что $d(x_i, s_j)$ минимально
  - Сделать центроиды классов зерновыми элементами, т.е. для каждого кластера c_j

$$s_j = \mu(c_j)$$

Алгоритм К-средних

- Возможное условие остановки цикла:
  - Количество итераций
  - Группировка документов по кластерам не изменяется
  - Позиции центроидов не изменяются
- Временная сложность алгоритма $O(n*\log n)$

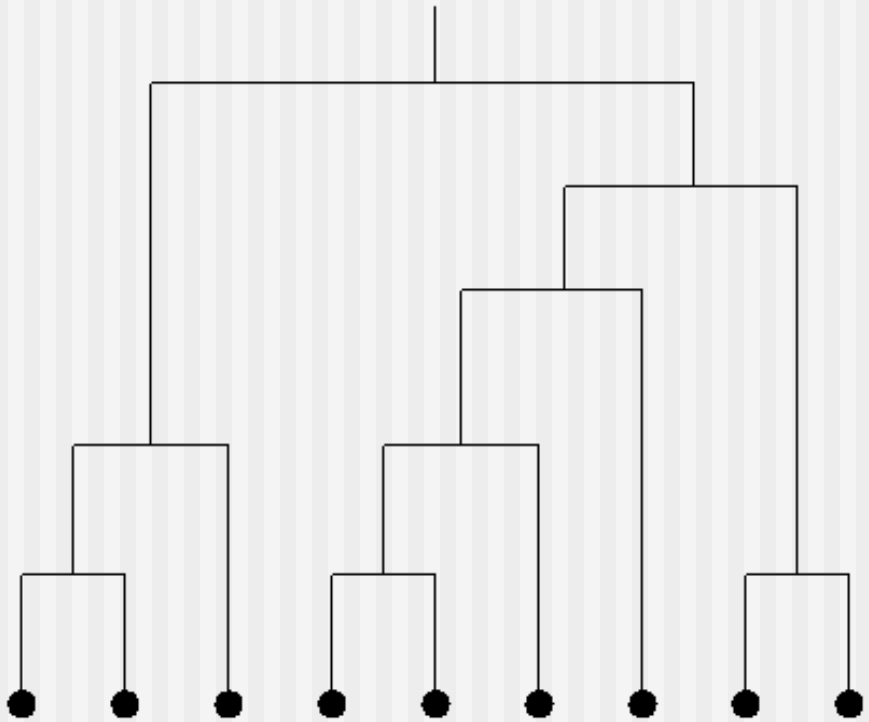
Метод К-средних

- Недостатки:
  - Результат кластеризации зависит от выбора стартовых элементов.
  - Значение ***k*** выбирается вне алгоритма.
  - Кластеры не пересекаются (жесткая кластеризация), т.е. для документа исключена возможность принадлежности к нескольким кластерам одновременно.
- Модификация: “мягкая” кластеризация.

Иерархическая агломеративная кластеризация

- Используется матрица сопряженности типа “документ-документ” (матрица подобия).
- Начальное количество кластеров совпадает с исходным количеством документов, затем итерационно происходит объединение кластеров в суперкластеры по степени подобия.
- В итоге получается один общий кластер, являющийся корнем древовидной структуры – дендограммы.

Иерархическая агломеративная кластеризация



Сечение дендограммы на любом уровне дает набор кластеров из *связных* между собой элементов, фактически получается дерево кластеров

Вычислительная сложность $O(n^2)$

Иерархическая аггломеративная кластеризация

- Порог принятия решения о подобии кластеров задается вне алгоритма.

Поиск кластера

- Входной *запрос* представляется в виде *t-мерного вектора* и сравнивается с *центроидами* кластеров.
- Поиск продолжается в кластерах, степень подобия для которых превысила заданный порог.
- Для вычисления степени подобия часто используется косинусная метрика.

Кластеризация в распределенных системах

- Существует необходимость распределенного хранения документов в кластерной системе
- По какому принципу распределять?
  - административным и т.п. способами;
  - по тематике.
- Как определять тематику, если она заранее не задана?

  Решение – кластеризация.

Подходы к кластеризации в распределенных системах

- Вырожденный случай – *единая система*
- Гетерогенные коллекции
  - кластеры построены заранее
- Глобальная кластеризация
  - все помещается в общее хранилище и выполняется кластеризация
  - для каждого кластера строится тематическая модель
- Локальная кластеризация
  - кластеризуется каждая из гетерогенных коллекций
  - внутри каждой коллекции формируются тематические модели для каждого полученного кластера (т.е. модель указывает на кластер внутри локальной коллекции)
- Политематическое представление
  - кластеризуется каждая из гетерогенных коллекций
  - тематические модели для каждой из коллекций собираются вместе (т.е. модель указывает на коллекцию)

Выполнение запроса в распределенной системе

- Выполнение запроса:
  - Ранжирование коллекций относительно запроса
  - Выбор *n* лучших коллекций
  - Выборка *N* лучших документов из каждой коллекции
  - Слияние списков из коллекций в общий список на основе показателя доверия

Кластеризация в распределенных системах: выводы

- Тематическая кластеризация эффективна для распределенного информационного поиска
- Лучшие результаты получаются при *глобальной* тематической кластеризации, поскольку подобные документы оказываются вместе
- При невозможности *глобальной* кластеризации относительно хорошим решением является *локальный* вариант, при этом
  - сохраняется административное распределение коллекций;
  - довольно большая нагрузка приходится на локальные сайты
- Хорошим решением является *множество тематик*, указывающих на единую коллекцию:
  - это лучше чем модель единой системы
  - кластеризация и формирование моделей затрагивает только локальный сайт
  - в дальнейшем исключается нагрузка на сайты с данными коллекции

Классификация документов на основе гиперссылок

Соседние в гиперссылочном графе документы могут содержать информацию, полезную при классификации:

- На текущий документ D_i может ссылаться другой документ D_j , содержащий высокоспецифичные термины, отсутствующие в самом документе D_i
- На текущий документ D_i может ссылаться другой документ D_j , содержащийся в релевантном разделе каталога, созданного вручную
- На текущий документ D_i может ссылаться другой документ D_j , содержащий ссылки также на многие другие документы, которые были использованы для настройки классификатора на релевантную тему (раздел каталога)

Использование свойств документов-соседей в графе гиперссылок

- *Идея*: использовать при классификации термины и метки классов документов-соседей по графу
- *Подход 1*:
  - описание документа расширяется за счет *терминов* документов-соседей, находящихся в графе в пределах радиуса $r \leq R$ (возможен учет расстояния r в виде весовой функции $1/r$)
  - недостаток: чувствительность к смещению темы
- *Подход 2*:
  - учет *меток классов* документов-соседей по графу в процессе классификации

Модификация запросов

- Переформулировка запроса
  - Расширение запроса
    - Добавление терминов в запрос для уточнения информационной потребности
  - Обратная связь
    - Оценка документов в выдаче как релевантных/нерелевантных
- Представление результатов
  - Кластеризация выданных результатов

Обратная связь по релевантности

- Метод *Rocchio*:

$$Q_1 = \alpha \cdot Q_0 + \frac{\beta}{n_1} \cdot \sum_{i=1}^{n_1} R_i - \frac{\gamma}{n_2} \cdot \sum_{i=1}^{n_2} S_i$$

где

Q_0 – вектор начального запроса

R_i – вектор i-го релевантного документа

S_i - вектор i-го нерелевантного документа

n_1 – число выбранных пользователем релевантных документов

n_2 - число выбранных пользователем нерелевантных документов

α, β, γ – параметры настройки

Латентно-семантическое индексирование как кластеризация

- LSI может рассматриваться как метод кластеризации (спектральная кластеризация)
- Вариант реализации для *k* кластеров:
  - Каждый элемент представляется точкой в *k*-мерном пространстве
  - Каждая точка проецируется на ось, соответствующую наибольшей по величине координате этой точки

Литература

- R. Larson “Principles of Information Retrieval”. Слайды (http://www.sims.berkeley.edu/academics/courses/is240/s06/)
- G. Weikum “Information Retrieval and Data Mining”. Слайды (http://www.mpi-sb.mpg.de/departments/d5/teaching/ws05_06/irdm/material.html)
- J. Allan “Information Retrieval”. Слайды. (http://ciir.cs.umass.edu/cmpsci646/)
- S.A.Macskassy, A.Banerjee, B.D.Davison, H.Hirsh “***Human Performance on Clustering Web Pages”***. Technical Report DCS-TR-355, Department of Computer Science, Rutgers University, 1998.

Литература

- J. Xu, W.B.Croft "***Cluster-based language models for distributed retrieval***." In Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99), 1999.
- S. Chakrabarti"***Mining the Web. Discovering Knowledge from Hypertext Data"***. Morgan Kaufmann Publishers, 2003.
- S.Deerwester,S.T.Dumais,G.W.Furnas,T.K.Landauer,R.Harshman "***Indexing by Latent Semantic Indexing***" JASIS, 1990.