рия Фишера); рассмотрены конкретные обучающие примеры применения программ, реализующих статистические критерии. Результатом работы является также адаптация методов непараметрической статистики к применению в управлении качеством конкретных производственных процессов.

Заключение. Реализованные в виде пакета программ для ЭВМ рассмотренные статистические методы предназначены для практического применения инженерами предприятий по качеству, в управлении процессами (в том числе в образовании), для информационного обеспечения оптимальных управленческих решений.

Список литературы

1. Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. – М.: «Наука», 1970. – 168 с.

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА В ПОСТРОЕНИИ ПОИСКОВЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

С.А. Ермоченко
Витебск, ВГУ имени П.М. Машерова

В настоящее время методы кластерного анализа в поисковых системах нашли очень широкое применение. Основной задачей поисковой системы является поиск тех документов среди достаточно большого массива имеющихся документов, которые более всего удовлетворяли бы вводимым пользователем ключевым словам. При этом кластерный анализ применяется для разбиения множества документов на такие подмножества, чтобы документы, наиболее удовлетворяющие критериям поиска, попадали в одно подмножество, или кластер [1].

Основным понятием, применяемым в кластерном анализе, является понятие метрики $\rho(d_n, d_m)$ как некоторого расстояния между двумя документами d_n и d_m. При этом такое расстояние должно быть тем меньше, чем более близки показатели релевантности документов к критерию поиска. Наиболее частым способом определения такого расстояния является сравнение весовых коэффициентов каждого документа по отношению к поисковым термам (ключевым словам). Конкретных способов вычисления таких коэффициентов существует несколько: весовой, вероятностный, семантический и т. д.

Однако для всех этих методик характерна одна общая черта: они предполагают однозначность поискового запроса (набора ключевых слов). Под однозначностью поискового запроса предполагается, что пользователю, составившему поисковый запрос, два документа, одинаково релевантных этому запросу, равно интересны. Но это не всегда так.

Целью данной работы является адаптация методов кластерного анализа для кластеризации результатов поиска, равнорелевантных поисковому запросу, по принадлежности документов к различным тематическим категориям. Такая кластеризация позволит представить результаты поиска не только в виде линейного списка, но и в виде иерархической структуры, в узлах которой располагаются различные категории и подкатегории.

Материал и методы. В качестве основного метода исследования используется сравнительно-сопоставительный анализ имеющихся методов кластерного анализа и их применяемости к результатам поисковой выдачи.

Результаты и их обсуждение. Поскольку основной целью работы является кластеризация равнорелевантных результатов поиска, то способ определения релевантности документа набору терм в данном случае не играет определяющей роли, поэтому рассмотрим самый простой. Рассмотрим набор терм как множество, состоящее из k различных элементов. Каждую терму в данном наборе будем рассматривать как равнозначную другим термам. Тогда каждому i-му документу можно поставить в соответствие вектор $\mathbf{x}_i$ из k элементов такой, что элемент $x_{i,j}$ ($j = 1, 2, \ldots k$) равен частоте упоминания j-ой термы в i-ом документе (вне зависимости от семантического значения и места упоминания термы в документе). В качестве релевантности i-ого документа набору терм можно рассматривать длину соответствующего вектора $\mathbf{x}_i$, например $x_i = \sqrt{\sum_{j=1}^{k} x_{i,j}^2}$.

Таким образом, если поисковый запрос состоит из одной термы «Java», то документы, в которых это слово упоминается одинаковое количество раз, будут иметь одинаковую релевантность. Но при этом в одном из этих документов может говориться об острове Java, в другом – о сорте кофе Java, а в третьем – о языке программирования Java.

Если применить к рассмотренному способу определения релевантности методы кластерного анализа согласно общепринятой методике, а расстояние $\rho(d_n, d_m)$ между двумя документами d_n и d_m определить как скалярное произведение векторов $\mathbf{x}_n \cdot \mathbf{x}_m$ [1], то все три приведённых в примере выше документа попадут в один кластер.

На практике такая ситуация приводит к тому, что пользователю приходится уточнять поисковые запросы, вводя в него новые термы, уменьшая так называемый информационный шум. При этом в различных рекомендациях по эффективному поиску в Интернет пользователям советуют применять последовательное уточнение поискового запроса. Однако, при разработке информационно-поисковых систем, ориентированных на определённую предметную область, важным критерием для пользователей таких систем является время, затраченное на поиск необходимой информации. В таком случае многократное последовательное уточнение поискового запроса приводит к неэффективности поиска нужных документов.

Кроме кластеризации результатов поиска по релевантности можно применить кластеризацию по различным характеристикам, на основе которых вычисляется некоторый приведённый коэффициент λ, позволяющий вычислить метрику для любых двух документов.

Рассмотрим, например, массив документов, каждый из которых имеет три характеристики: x_i, y_i, z_i. При этом эти характеристики могут иметь различные единицы измерения. Это могут быть релевантность, дата создания документа, размер документа. Для вычисления приведённого коэффициента λ_i каждая характеристика рассматривается как некоторая случайная величина с некоторым законом распределения. Любая такая случайная величина может быть преобразована в равномерно-распределённую случайную величину на отрезке [0; 1]. Обозначим преобразованные случайные величины из нашего примера соответственно $\tilde{x}_i$, $\tilde{y}_i$ и $\tilde{z}_i$, тогда приведённый коэффициент может вычисляться как среднее арифметическое преобразованных случайных величин $\lambda_i = \frac{1}{3}(\tilde{x}_i + \tilde{y}_i + \tilde{z}_i)$. Кроме того, приведённый коэффициент может вычисляться как средневзвешенное всех характеристик, что позволяет при кластеризации учесть каждую характеристику в различной степени, например $\lambda_i = 0{,}7\,\tilde{x}_i + 0{,}2\,\tilde{y}_i + 0{,}1\,\tilde{z}_i$.

Тем не менее, такой способ кластеризации также не является универсальным. Очень часто все документы, хранящиеся в информационно-поисковой системе, имеют принадлежность к некоторой тематической категории, и при кластеризации результатов поиска эта одна из наиболее важных характеристик после релевантности документа. Проблема заключается в том, что данная характеристика не имеет измеримой величины. В некоторых случаях, правда, тематические категории могут иметь и числовые характеристики (например, шифры ББК или УДК), однако эти числовые характеристики не всегда отображают степень близости различных категорий между собой.

Одним из способов определения числовых характеристик тематических категорий можно предложить иерархическую структуризацию таких категорий. Если для каждой категории определить подчинённые ей подкатегории и предположить, что все подкатегории одной категории одинаково близки между собой, можно каждой категории назначить свой весовой коэффициент, исходя из её уровня. Фактически, это задача кластеризации самих категорий. Для решения этой задачи можно использовать экспертные системы (что является предпочтительным для специализированных предметных областей), или кластеризовать все документы, но не по релевантности некоторому запросу, а по семантической схожести полного содержания документов между собой. Последняя задача осложняется, во-первых, трудностью определения степени семантической схожести документов, а, во-вторых, значительными вычислительными и временными затратами.

Заключение. В данной работе рассмотрены некоторые аспекты кластеризации результатов поиска информации в информационно-поисковых системах. Продемонстрированы возможности кластерного анализа по решению различных задач, возникающих в процессе построения информационно-поисковых систем. Рассмотрены также и трудности в адаптации методов кластерного анализа к обработке текстовой информации, не имеющих чётких числовых характеристик.

Список литературы

1. Ландэ, Д. В. Интернетика. Навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов. – Москва: Научная и учебная литература, 2009. – 290 с.

ВИЗУАЛЬНОЕ МОДЕЛИРОВАНИЕ ПРИЛОЖЕНИЙ С ПОМОЩЬЮ UML И ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИИ CACHE ПРИ РАЗРАБОТКЕ ПРИЛОЖЕНИЙ

О.Г. Казанцева
Витебск, ВГУ имени П.М. Машерова

Разработка программного продукта – это сложный и многоэтапный процесс, начинающийся с постановки задачи и разработки технического задания. Далее следует разработка архитектуры системы, написание кода, тестирование и частичное внедрение, устранение «багов» (ошибок в программном обеспечении), написание технической документации. Заканчивается процесс разработки передачей программного комплекса в эксплуатацию и дальнейшим сопровождением продукта.

При подготовке студентов специальности «Прикладная математика» в учебном плане предусмотрена дисциплина «Избранные главы информатики». В рамках данной дисциплины студенты изучают современные технологии разработки