МЕТОД КЛАСТЕРНОГО АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ ДЛЯ ИНФОРМАЦИОННОГО ПОИСКА

Егорова И.Н., Егоров С.В.
Харківський національний університет радіоелектроніки

Необходимость анализа и исследования больших объемов текстовой информации существует в различных областях, таких как data mining, базы данных, – информационный поиск. Современным эффективным инструментом такого исследования является кластерный анализ.

В связи с этим, работа, направленная на разработку метода кластерного анализа, позволяющего автоматически осуществлять кластеризацию текстовых документов, представляется актуальной.

Целью работы является создание метода кластерного анализа текстовых данных для повышения качества и скорости формирования результатов информационного поиска.

Проведенное в работе сравнительное исследование алгоритмов кластерного анализа, позволило определить наиболее эффективные из них [1]. Таковыми оказались алгоритмы K-means и bisecting K-means, реализующие методы разделения (partitioning methods), и алгоритмы DBSCAN и OPTICS, реализующие методы, основанные на плотности (density-based methods).

Установлено, что алгоритмы K-means и bisecting K-means эффективны для кластеризации баз данных малых и средних размеров. Ограничения в виде необходимости предварительного задания количества кластеров K, которое может быть определено только эмпирическим путем, делает их малоэффективными для кластеризации больших объемов данных, а также не позволяет осуществлять автоматическую кластеризацию текстов.

Алгоритмы DBSCAN и OPTICS предназначены для нахождения кластеров произвольной формы [2]. Эти алгоритмы рассматривают кластеры как плотно сжатые области объектов в пространстве данных, разделенных областями низкой плотности (шумами).

Проведенное в работе исследование группы алгоритмов, основанных на плотности, позволило установить, что данным алгоритмам не свойственны ограничения, аналогичные выявленным для группы алгоритмов разделения.

Следует, однако, заметить, что алгоритм DBSCAN требует задания пользователем минимального количества членов кластера *MinPts*, а также определения радиуса поиска объектов ε в общем множестве. Такие установки параметра обычно задаются эмпирически и являются трудноопределимыми, особенно в условиях реального мира для больших объемов данных.

Алгоритм OPTICS существенно расширяет работу алгоритма DBSCAN, в котором для постоянного значения *MinPts* кластеры высшей плотности полностью содержатся в наборах, полученных относительно низшей плотности. С целью создания набора отсортированных кластеров предусмотрена возможность одновременной обработки разных наборов значений расстояний.

Основным преимуществом алгоритма OPTICS по сравнению с другими алгоритмами группы методов, основанных на плотности, является возможность более точного формирования кластеров, начиная с кластеров высокой плотности

и заканчивая более разреженными. Качество кластеризации напрямую зависит от точности разбиения объектов на кластеры.

Ограничением алгоритма OPTICS является неспособность самостоятельно осуществлять кластеризацию упорядоченных объектов. Для этой цели необходимо использовать другие алгоритмы, например ExtractDBSCAN-Clustering. Еще одним ограничением алгоритма OPTICS является необходимость задания пользователем значений *MinPts* и ε.

Таким образом, рассмотренные алгоритмы обладают рядом существенных ограничений, непосредственно влияющих на точность кластеризации, и не позволяют осуществлять сам процесс кластеризации в автоматическом режиме.

Предложенный в работе метод позволяет усовершенствовать метод кластеризации, основанный на плотности. Прежде всего, предложено осуществлять поиск документов, соответствующих запросу пользователя, не в базе данных документов, а в базе данных аннотаций [3].

Следующим основополагающим понятием предлагаемого метода является формирование из документов, а в нашем случае, – аннотаций, максимально плотных кластеров.

Предлагается для поиска документа, максимально соответствующего запросу, учитывать среднестатистическое количество слов запроса, а в качестве значения радиуса соседства использовать косинусную меру подобия.

Предложенный алгоритм позволяет максимально быстро сформировать перечень документов, наиболее полно соответствующих запросу пользователя, а также повысить скорость его работы за счет распараллеливания потоков данных на этапах формирования кластеров.

Таким образом, предложенный в работе метод реализует возможность автоматически осуществлять кластеризацию документов в БД больших и сверхбольших объемов и позволяет более точно формировать кластеры, начиная с кластеров с максимально возможной плотностью. Существенный выигрыш в скорости обусловлен реализованной в методе возможностью многопоточной обработки данных при формировании кластеров. Метод может быть использован для совершенствования информационного поиска.

1. Jiawei H. Data Mining: Concepts and Techniques. / H. Jiawei, Kamber M. ; Second edition. – Morgan Kaufmann Publishers. – 2006. – 772 p.

2. Егорова И.Н. Разработка программного обеспечения для решения задач распознавания образов / И.Н. Егорова, С.В. Егоров, Восточно-Европейский журнал передовых технологий. – Харьков.-2010. – №1/5(43). – с. 67-68.

3. Егоров С.В. Семантическое аннотирование в информационном поиске / С.В. Егоров, Инновации молодежной науки: тез. докл. Всерос. науч. конф. молодых ученых / С. Петербургск. гос. ун-т. технологий и дизайна.- СПб.: ФГБОУВПО «СПГУТД», 2013.- с. 113.