Рецензия

магистерской диссертации Ромашова Дмитрия Сергеевича
«Поиск похожих объектов в мультимедийных данных»

Диссертация Ромашова Д.С. посвящена проблеме поиска похожих кортежей данных. Элементы кортежей произвольны; в качестве примеров рассматриваются числовые, строковые и мультимедийные (графические) данные. По условиям задачи, критерии похожести определены нечётко. Доступна выборка, содержащая примеры похожих и непохожих данных и созданная экспертом вручную. Целью диссертационной работы было исследование применимости различных методов классификации к решению задачи поиска похожих объектов. Поставленная задача относится к актуальному направлению Big Data, и сама по себе также является актуальной.

Как указано выше, сравнение составных данных является нетривиальной задачей. Первая глава диссертации посвящена обзору методов, позволяющих эту задачу упростить. От сравнения текстовых данных автор предлагает перейти к сравнению словарей текстов, причём в словарь предлагается добавлять основу слова, которую необходимо предварительно выделить с помощью стемминга или лемматизации. Графические данные тоже невыгодно сравнивать напрямую; вместо этого предлагается сравнивать их хеши или гистограммы. Завершается глава рассмотрением методов классификации составных данных и критериев, по которым эти методы можно сравнить между собой.

Во второй главе рассказывается о данных, подлежащих сравнению. Приводится структура базы данных. Производится выбор признаков для сравнения кортежей. Числовые данные сравниваются традиционным способом; текстовые – так, как предложено в главе 1. Для сравнения графических данных применяются гистограммы, т.к. сравнение хешей не позволяет выявлять пары изображений, различающихся только углом съёмки или расположением объекта. Наконец, строится вектор признаков кортежа фиксированной длины, элементы которого могут быть просуммированы для получения скалярной метрики, удобной для сравнения.

Третья глава описывает практическое исследование. Для сравнения автор выбрал метод k ближайших соседей, метод опорных векторов, метод решающих деревьев и метод случайного леса. Параметры каждого метода подбирались методом кросс-валидации. Все методы были реализованы программно на языке Python. Были измерены показатели точности, полноты и F_1-мера каждого метода на тестовой выборке.

В работе приведены экспериментальные результаты, доказывающие достаточно хорошую работу методов поиска. Показано, что довольно значительные изменения текстовых полей и изменение угла съёмки не препятствуют правильному распознаванию кортежей как дубликатов. Приводится и пример кортежей, корректно определённых как различные, несмотря на общую тему текста и сходство изображений.

Предложенные в работе алгоритмы хорошо поддаются распараллеливанию. Учитывая выбор языка Python для программной реализации, перспективным представляется интеграция с одной из платформ Big Data, например, Hadoop.

Оценки производительности разных методов, приведённые в работе, близки между собой. К сожалению, не приводятся значения точности и доверительной вероятности измеренных величин, что не позволяет понять, превышает ли разница между показателями разных методов статистическую погрешность. Другим недостатком работы является непоследовательность в

использовании русских и английских терминов («случайный лес» и “random forest” и пр.)

Несмотря на отмеченные недостатки, считаю, что магистерская диссертация Ромашова Д.С. заслуживает оценки «**Отлично**»

Рецензент,
к.т.н., ст.н.с. ООО «НСН» Епифанов/ Епифанов Н.А.