УДК 004.78:025.4.036

МЕТОД ШИНГЛИРОВАНИЯ В ЗАДАЧЕ ПОИСКА ДОКУМЕНТОВ, ПОХОЖИХ НА ДАННЫЙ

Прияцелюк Н. С.
Брестский государственный технический университет
Научный руководитель: Крапивин Ю. Б., канд. техн. наук

Несмотря на длительный период существования, проблема поиска релевантных документов по произвольному пользовательскому запросу по-прежнему актуальна. То же можно сказать и о проблеме поиска документов, похожих на данный, при котором в качестве запроса выступает некоторый документ-образец, использующийся для нахождения текстов, сходных с ним по содержанию. Решения указанной задачи имеют различные практические применения, например, поиск схожих по тематике документов, поиск дубликатов документов, установление оригинального источника документа.

Существуют различные методы ее решения, например MD5, TF, TF*IDF, TF*RIDF, Long Sent, Heavy Sent, Megashingles, Lex Rand, Log Shingles, Descr Words, Opt Freq, обзор которых можно найти в статье [1]. Все описанные методы основываются на создании сигнатуры документа (одна или несколько контрольных сумм (хешей) в зависимости от метода), с которой можно сравнивать целевой документ.

Для решения поставленной задачи был выбран алгоритм шинглирования, впервые описанный *A. Broder et al.* [2]. Так как изначальное описание алгоритма имело свои недостатки, с течением времени его улучшали. Для реализации использовалось описание алгоритма, предложенное в статье [3]. Алгоритм может быть представлен следующей последовательностью шагов:

1. Начало.
2. Каждый документ в коллекции проходит предобработку (токенизация, отбрасывание знаков препинания и чисел, лемматизация и т. д.).
3. Для каждого токена в преобразованном тексте вычисляется хеш с помощью хеш-функции CRC32.
4. Окном в N хешей ([1-й, 2-й, … N-ый хеш], [2-й, 3-й, …, N+1-ый хеш] и т. д) собирается массив шинглов. Каждый шингл сливается в один объект и хешируется снова.
5. Конец.

Полученный на шаге 4 массив шинглов и является основой для поиска. Шинглы помещаются в инвертированный индекс, где ключом является отдельный шингл, а значением — идентификаторы всех документов, в которых он встречается. Для поиска совпадений входной текст подвергается описанной выше процедуре хеширования, и, далее, производится поиск его шинглов в инвертированном индексе, с последующим возвращением идентификаторов документов, в которых встретился определённый шингл.

Для тестирования данного подхода было разработано приложение и собрана коллекция из 540 документов со средним размером в 1500 лексем. Для каждого документа была проведена предобработка и применён алгоритм шинглирования с окном в 10 лексем. Тестирование показало, что приложение может производить поиск документов, содержащих совпадающие текстовые фрагменты, в среднем за ~0.0005 секунды.

Несмотря на полученные результаты, можно с уверенностью сказать, что есть большой простор для улучшений, как в плане повышения скорости и качества поиска документов, так и в плане мероприятий, предшествующих ему (предобработка). Так, на получаемый результат оказывает влияние ширина окна шинглирования. Если взять слишком широкое окно, то даже незначительное изменение в пределах этого окна может привести к тому, что в процессе поиска для данного шингла не найдётся совпадения. Если же взять слишком узкое окно, то мы рискуем получить много ложноположительных результатов.

Однако одного изменения размера окна недостаточно, чтобы компенсировать такие изменения, как изменения порядка слов или использование синонимов. Для решения последней проблемы на этапе предобработки текста можно применить лексическую базу данных, чтобы привести все синонимы к одной форме.

Список цитированных источников

1. Зеленков, Ю.Г. Сравнительный анализ методов определения нечётких дубликатов для Web-документов. Ю.Г. Зеленков, И.В. Сегалович [Электронный ресурс]. – Режим доступа: URL: http://rcdl2007.-pereslavl.ru/papers/paper_65_v1.pdf

2. Broder, A. Syntactic clustering of the Web. A. Broder, S. Glassman, M. Manasse and G. Zweig. [Электронный ресурс]. — Режим доступа: URL: https://www.hpl.hp.com/techreports/Compaq-DEC/SRC-TN-1997-015.pdf

3. Ивахненко, А. Так устроен поиск заимствований в Антиплагиате. [Электронный ресурс]. – Режим доступа: URL: https://habr.com/ru/company/antiplagiat/blog/429634/

УДК 004.89

МЕТОДЫ ПОСТРОЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАСПОЗНАВАНИЯ ОБРАЗОВ

Хацкевич М. В.

Брестский государственный технический университет», г. Брест, Беларусь

Научный руководитель: Головко В. А., доктор техн. наук, профессор

Свёрточные нейронные сети предназначены для распознавания визуальных образов непосредственно из пиксельных изображений с минимальной предварительной обработкой. Данные нейронные сети могут распознавать образы с крайней изменчивостью, а также с устойчивостью к искажениям и простым геометрическим преобразованиям.

Сверточная нейронная сеть состоит из разных видов слоев (рисунок 1):

1. Входной слой(Input): входное изображение, включая несколько цветовых каналов.
2. Сверточный слой (Convolution): все нейроны слоя, в отличие от персептрона, связаны только с частью нейронов предыдущего слоя.
3. Слой субдискретизации (подвыборочный) (Pooling, Subsampling): выделение наиболее значимых признаков предыдущего слоя и значительное сокращение размерности последующих слоев сети.
4. Полносвязный слой (Fully-connected): представляет собой скрытый слой искусственной нейронной сети типа персептрон.