МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ ПОЛІТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІНСТИТУТ КОМП'ЮТЕРНИХ СИСТЕМ

МАТЕРІАЛИ ДЕВ'ЯТОЇ МІЖНАРОДНОЇ НАУКОВОЇ КОНФЕРЕНЦІЇ СТУДЕНТІВ ТА МОЛОДИХ ВЧЕНИХ

присвячена 55-річчю
ІНСТИТУТУ КОМП'ЮТЕРНИХ СИСТЕМ

" Сучасні інформаційні технології 2019 "

" Modern Information Technology 2019 "

23-24 травня

Одеса
«Екологія»
2019

УДК 004.58

СИСТЕМА АНАЛИЗА НОВОСТНОГО ПОТОКА НА ОСНОВЕ АЛГОРИТМА КЛАСТЕРИЗАЦИИ

Балан А.С.
доц. каф. ИС, к.т.н. Бабилунга О.Ю.
Одесский Национальный Политехнический Университет, УКРАИНА

АННОТАЦИЯ. В данной работе разработана система, осуществляющая сканирование и кластеризацию статей из нескольких интернет-изданий. Система состоит из двух программ, работающих независимо друг от друга – сканер новостей и *web*-сервер. Разбиение потока новостей по сюжетам реализовано на основе алгоритма двухпроходной кластеризации.

Введение. В связи с развитием глобальной сети Интернет и онлайн-ресурсов средств массовой информации (СМИ) значительно возросли объемы информации, с которыми приходится работать пользователю, что осложняет задачу поиска актуальной информации среди новостных статей. Следовательно, задача автоматической обработки новостных лент является востребованной и актуальной. Разработка методов для автоматической обработки и агрегации новостных потоков позволяет существенно сократить объем материалов, необходимых для просмотра и анализа человеком.

Целью работы является разработка системы анализа новостного потока для разбиения статей на кластеры, представляющие собой отдельные темы, категории или события и выбор алгоритма кластеризации, позволяющего реализовать этот процесс автоматически.

Основная часть. Кластеризация данных используется в большинстве современных информационно-поисковых систем, обрабатывающих потоки информации [1]. Объединение схожих статей в кластеры делает интерфейс системы более понятным и повышает эффективность работы пользователя с ней, при этом содержание кластеров определяется только распределением и структурой данных. Существуют статическая, инкрементальная и онлайн-кластеризации [2].

На алгоритмы статической кластеризации не накладываются ограничения по использованию памяти или количеству проходов по множеству документов. Такие алгоритмы требуют возможности произвольного доступа к документам и их содержимому, обработка больших объемов данных может привести к большим потерям в производительности приложений и времени обработки коллекций. В алгоритмах инкрементальной кластеризации наборы данных представляются в виде потоковой модели. Однако, в таких задачах, как автоматическое разбиение статей из новостного потока, общий набор документов, подлежащих кластеризации, не может быть заранее определен, так как на вход системы непрерывно поступают новые статьи. Для решения этой проблемы требуется либо адаптация существующих алгоритмов статической и инкрементальной кластеризации, либо разработка новых алгоритмов с учетом специфики онлайн-кластеризации. Из научных источников известны наивный однопроходной – базовый алгоритм кластеризации и алгоритм двухпроходной кластеризации (*Doubling*-алгоритм) [3]. В данной работе предложено реализовать двухпроходную кластеризацию: алгоритм решает проблему онлайн *k*-кластеризации, т.е. задачу разбиения потоковых данных на кластеры, число которых точно задано и равно *k*. Алгоритм использует два параметра *a* и *b* такие, что $\frac{a}{a-1} \le b$. Рассмотрим *i*-ю итерацию алгоритма. Пусть сформирована коллекция из k кластеров $(C_1,...,C_k)$, d_i – минимальное значение их диаметров. Каждый кластер C_i имеет центроид c_i, которым является один из принадлежащих ему документов. Каждая итерация состоит из двух фаз: слияния и обновления. На этапе слияния мы устанавливаем $d_{i+1} = b \times d_i$, и на основе этого значения из

существующих кластеров формируются новые кластеры по следующему принципу: C_p и C_s объединяются в один кластер, если расстояние между их центроидами c_p и c_s меньше или равно d_{i+1}. В результате работы фазы слияния мы имеем $m \le k$ кластеров. На фазе обновления считываются новые поступающие документы и если расстояние от нового объекта до ближайшего центроида не превышает величину $a \times d_{i+1}$, он добавляется в соответствующий кластер. Если терм-вектор документа лежит достаточно далеко от всех центроидов – образуется новый кластер. Фаза обновления продолжается до тех пор, пока число кластеров не станет равно *k*.

Предлагаемая система («агрегатор») собирает статьи различных интернет-изданий СМИ и выделяет из них те, что относятся к одним и тем же темам или событиям. Система состоит из двух программ, работающих независимо друг от друга – сканер новостей и *web* -сервер. Сканер периодически опрашивает добавленные в список источников новостные серверы, загружая их *RSS*-ленты и проверяя наличие свежих, еще не обработанных статей. Если такие статьи появились с момента прошлого обхода сканера, происходит загрузка страницы с сайта издания, содержащей полный текст сообщения. На следующем этапе *html*-страница подвергается обработке с целью выделения полного текста статьи, не содержащего лишней информации. Вместе с содержанием статьи анализируются ее метаданные – дата публикации, источник, ссылка на первоисточник и т.д. Формируется образ статьи внутри системы с применением лингвистического анализа, алгоритма Портера [4]. Далее осуществляется индексация текста статьи и некоторых метаданных. Новая статья представляется в виде вектора преобразованных слов (термов) и добавляется в хранилище индексов на файловой системе. Модуль кластеризации принимает на вход образ статьи, содержащий ее вектор термов, и считывает из базы данных и хранилища индексов данные, необходимые для определения кластера, в который будет занесена новая статья.

Параллельно системе сканера новостей работает *web*-сервер, который обрабатывает входящие пользовательские соединения и запросы. При новом запросе он обращается в базу данных кластеров и отсылает пользователю информацию о последних событиях, основанную на образованных кластерах.

Выводы. В ходе исследования рассмотрены особенности применения статических, инкрементальных, он-лайн алгоритмов кластеризации к решению задачи анализа новостного потока. В работе реализована система анализа новостного потока и алгоритм обнаружения групп содержательно близких новостных сообщений в подборке новостей. Алгоритм основан на использовании двухпроходной кластеризации векторного представления текстов статей и позволяет достичь хорошего качества автоматической кластеризации без использования сложных методов анализа данных, таких как синтаксический и семантический.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Казиев Г.З., Курдюков В.В. Модели и методы кластеризации больших данных для их анализа и обработки // Модернизационный вектор развития науки в XXI векек: Традиции, новации, преемственность. – Санкт-Петербург, 30 апреля 2016 г.

2. Кутуков Д.С. Применение методов кластеризации для обработки новостного потока [Текст] // Технические науки: проблемы и перспективы: материалы Междунар. науч. конф. (г. Санкт-Петербург, март 2011 г.). — СПб.: Реноме, 2011. — С. 77-83. — URL https://moluch.ru/conf/tech/archive/2/207/ (дата обращения: 04.05.2019).

3. Кондратьев М.Е. Анализ методов кластеризации новостного потока. – Труды 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2006. – Суздаль, 2006. – С. 108–114.

4. [Алгоритм выделения псевдооснов Мартина Портера] [Электронный ресурс]. – Электрон. дан. – Режим доступа: http://snowball.sourceforge.net, свободный.