УДК 004.912

КЛАСТЕРНЫЙ АНАЛИЗ СОДЕРЖИМОГО САЙТА ВОПРОСОВ-ОТВЕТОВ STACKEXCHANGE.COM

К.Ю. СЛИСЕНКО[1], С.И. СИРОТКО[2]

Белорусский государственный университет информатики и радиоэлектроники
ул. П. Бровки, 6, г. Минск, 220013, Республика Беларусь
[1]kslisenko@gmail.com; [2]sergeyis@tut.by

Рассматривается процесс кластеризации данных сайта вопросов и ответов, имеющего большое количество неструктурированного содержимого. В виду предполагаемого большого объёма обрабатываемых данных рассматривается и обосновывается применение специализированных технических средств анализа данных – фреймворков Apache Hadoop и Mahout [1]. Рассматривается применение алгоритма кластеризации К-средних для больших объёмов данных. Даётся интерпретация и пояснение полученных результатов.

Ключевые слова: машинное обучение, анализ данных, кластеризация, векторизация, Apache Hadoop, Apache Mahout.

Многочисленные форумы, сайты с вопросами и ответами и иные интернет-ресурсы могут содержать актуальные, свежие и новые знания. Например, популярный ресурс stackexchange.com содержит 8 миллионов вопросов и 14,6 миллионов ответов [2]. Однако проблема заключается в неструктурированности и большой зашумлённости имеющейся информации. Поэтому актуальной задачей является исследование данной информации с целью выявления закономерностей и групп схожих вопросов и ответов, а так же структурирование и последующее визуальное представление общей картины.

Исследуемые данные имеют объём около 15 Гб, поэтому необходимо использовать специализированные технологии и алгоритмы для обработки больших объёмов данных [1]. Для работы был выбран стек технологий на основе фреймворка с открытым исходным кодом Apache Hadoop [3]. Выбор обоснован возможностью масштабирования обработки данных на кластере из множества машин, а так же автоматическим обеспечением отказоустойчивости вычислений [1]. Для осуществления поставленной задачи необходимо прибегнуть к алгоритмам машинного обучения. Для анализа был выбран алгоритм кластеризации К-средних, поскольку он не требует задания ожидаемых групп вопросов, и может быть масштабирован на кластере из множества машин [4]. Данный алгоритм реализован в составе библиотеки Apache Mahout, работающей поверх Hadoop [5].

Первым этапом анализа является подготовка исходных данных в формате XML для последующего применения алгоритма кластеризации. Алгоритм подготовки данных выглядит следующим образом:

1. Разбор XML, извлечение текста вопросов;
2. Обработка текста: разбивка на слова, удаление нетекстовых слов, приведение к нижнему регистру, отсечение слов по минимальной и максимальной длине, удаление частоупотребимых слов, обработка алгоритмом Портера для приведения слов в начальную форму и удаления окончаний и суффиксов, формирование словаря;
3. Векторизация с учётом частоты слов относительно всех текстовых документов (применение алгоритма TF-IDF).

На вход алгоритму кластеризации приходят сформированные векторы вопросов и ответов. Алгоритм К-средних требует установки количества кластеров, которое зада-

ётся во входном параметре K. Данный алгоритм при выполнении на множестве машин выглядит следующим образом:

1. Выбирается k случайных точек как центры кластеров;
2. На фазе Map каждая точка ассоциируется с ближайшим центром кластера;
3. На фазе Reduce пересчитываются центры кластеров;
4. Производится расчёт величины смещения кластеров.

Для измерения близости между точками и кластерами была использована метрика косинусного расстояния, поскольку она наиболее оптимальна для текстовых данных. Идентификаторы вопросов и ответов сортируются и объединяются с изначальным текстом. На рис. 1 изображены результаты визуализации найденных кластеров.

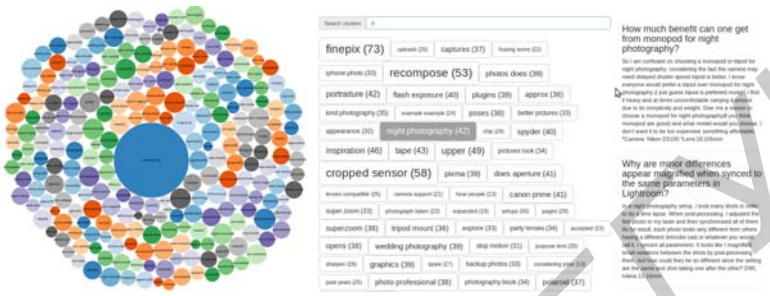

Рис. 1. Результаты визуализации найденных кластеров

На изображении слева каждая окружность соответствует кластеру, величина окружности пропорциональна количеству вопросов. На изображении справа виден список кластеров с возможностью просмотра вопросов, входящих в каждый кластер. Выполненная визуализация помогает понять, насколько широко обсуждаема та или иная тематика посетителями сайта.

В рамках исследования был успешно проведен кластерный анализ содержимого web-сайта. При этом было выявлено, что наибольшим образом на качество кластеризации влияет предобработка текстовых данных. В дальнейшем планируется разработать критерии оценки качества кластеризации и на основе них сравнить различные алгоритмов и метрики. Также планируется применить алгоритм для оценки первоначального числа кластеров, чтобы упростить в дальнейшем автоматизацию процесса обработки.

Список литературы

1. *Слисенко, К. Ю., Сиротко, С.И., Кириченко А.Ю.* // Сб. тез. докл. междунар. на-уч. конф. “Информационные технологии и системы 2013”. Минск, 23 октября 2013 г. С. 326–327.
2. Stackexchange: сайт вопросов и ответов. [Электронный ресурс]. – Режим досту-па: http://stackexchange.com. – Дата доступа: 14.01.2014.
3. *White, T.* Hadoop the definitive guide, Third edition. O’Reilly Media, 2012.
4. *Gillick, D.* // MapReduce: Distributed Computing for Machine Learning // Berkeley University, CS262A. 2006.
5. *Oven S., Anil R.* Mahout in Action. Manning Publications, 2012.