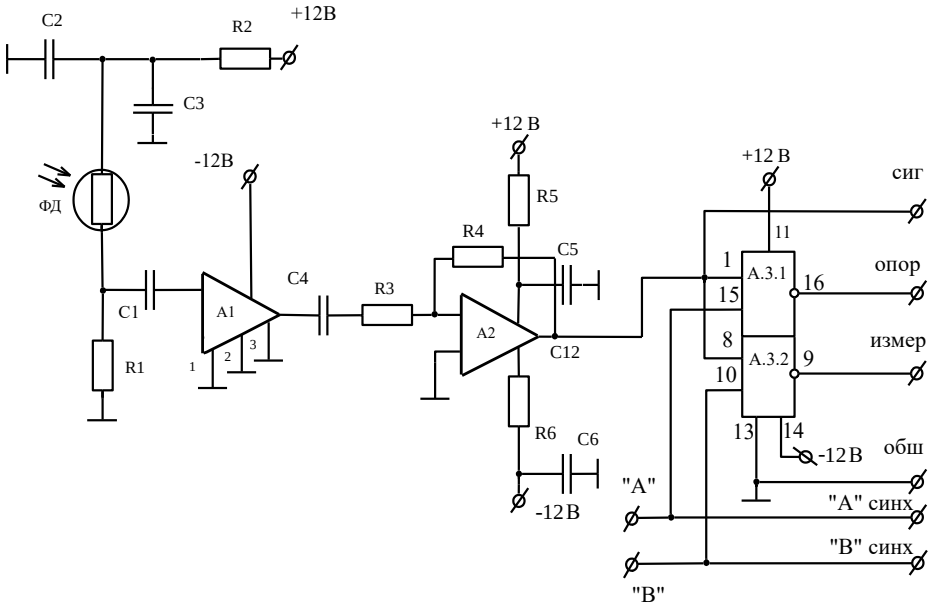

Рис.2. Принципиальная схема приемной части устройства

Общий алгоритм работы прибора следующий. В датчике между светодиодным излучателем и фотоприемником располагается исследуемый образец. Светодиоды просвечивают образец двумя длинами волн, а фотоэлемент, реагируя на свет, генерирует сигнал. Далее, сигнал преобразуется микроконтроллером в цифровую форму и преобразуется в пакет из двух значений, который передается через контроллер порта на ПК, где посредством специального программного обеспечения происходит обработка информации.

В качестве устройства сопряжения используется схема, построенная на основе микроконтроллера Atmega 48 и контроллера USB-порта FTDI.

Литература

1. М. Мухитдинов, Г. Цой. Приборы для измерения и регистрации электрических сигналов на основе персонального компьютера. Т.: - «Алокачи», 2012, с. 41-44.

2. Фукс-Рабинович Л.И., Епифанов М.В. «Оптикоэлектронные приборы» Л.: - «Машиностроение», 1979.,с.120-123.

А.Н. Назарова, И.А. Сюсин

ИССЛЕДОВАНИЕ ТЕХНОЛОГИЙ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ ДАННЫХ О ГЕОГРАФИЧЕСКОМ ПОЛОЖЕНИИ МОБИЛЬНЫХ УСТРОЙСТВ

(Самарский университет)

В современном мире широко используются новые технологии, цифровые устройства, повсеместная генерация цифровой информации, что делает доступной в реальном времени информацию из различных источников, таких

как GPS навигаторы, камеры видеонаблюдения, спутниковые данные, мобильные телефоны, электронная торговля, банковские карты, социальные сети, интернет запросы в поисковиках, электронные сообщения, карты местности пользователя и др. Если все подобные данные накапливать для дальнейшей обработки, то их суммарный объем будет измеряться десятками и сотнями петабайт (10^{15} байт). Совокупность объемных и неструктурированных данных из всех таких источников принято называть «большие данные» [1].

Большие данные – это термин, обозначающий множество наборов данных столь объемных и сложных, что делает невозможным применение имеющихся традиционных инструментов управления базами данных и приложений для их обработки. Изначально проблема состояла в том, что объем информации настолько вырос, что рассматриваемое количество уже фактически не помещалось в памяти компьютера, используемой для обработки, поэтому инженерам потребовалось модернизировать инструменты для анализа всех данных. Так появились новые технологии обработки, например, модель MapReduce (рис.1) компании Google и ее аналог с открытым исходным кодом – Hadoop от компании Yahoo. Они дали возможность управлять намного большим количеством данных, чем прежде. При этом важно, что их не нужно было выстраивать в аккуратные ряды или классические таблицы баз данных.

На данный момент основную проблему представляют сбор, очистка, хранение, поиск, доступ, передача, анализ и визуализация таких наборов как целостной сущности, а не локальных фрагментов. Определяющими характеристиками больших данных принято считать: объём, скорость прироста и необходимость высокоскоростной обработки и получения результатов, многообразие способов обработки различных типов, структурированных и полуструктурированных данных [2].

Огромные объёмы данных обрабатываются для того, чтобы человек мог получить конкретные и нужные ему результаты для их дальнейшего эффективного применения.

Наиболее часто применяемыми функциональными операциями над данными и методами их хранения и обработки являются:

- извлечение данных;
- краудсорсинг – сбор данных от большого числа источников;
- машинное обучение (с учителем и без учителя);
- нейронные сети;
- анализ сетей;
- оптимизация;
- классификация;
- кластерный анализ.

Подход больших данных призван существенно увеличить использование имеющейся информации и позволить представить ее в подходящем для практического применения виде: принятия решений человеком или автоматического управления системами.

Целью данной работы является исследование проблемы обработки

больших объёмов данных, содержащих географическое положение, и их кластеризация.

Вся работа с Big data была разбита на несколько этапов:

– сбор данных;

– структурирование и сортировка полученной информации;

– конечное использование и применение данных для определения геолокации.

В качестве входных и обрабатываемых данных была использована информация, полученная с таких устройств, как мобильные телефоны, GPS-навигаторы, планшеты и т.п.

Для сортировки полученных данных была использована наиболее популярная программная реализация модели параллельной обработки больших объемов данных путем разделения на независимые задачи, решаемые функциями Map и Reduce – Hadoop MapReduce [3]. Работа алгоритма MapReduce состоит из трех основных этапов: Map, Group и Reduce.

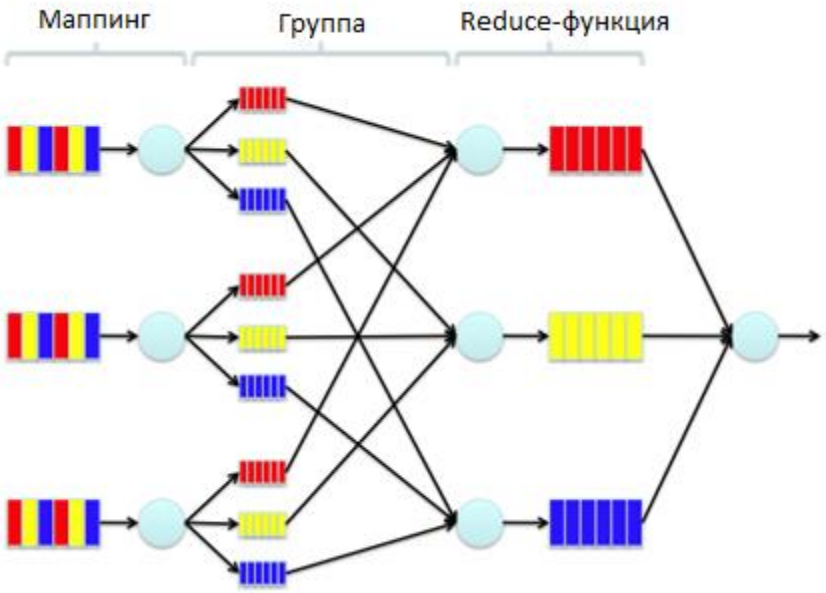

Рис. 1. Модель MapReduce

Полученные данные были кластеризированы и классифицированы по заранее заданным классам. В данной работе в качестве метода кластеризации был рассмотрен наиболее популярный на сегодняшний день метод k-means.

K-means применим к объектам в d-мерном векторном пространстве, представленных как набор D = {xi | i=1, …, N}, где xi ∈ Rd – i-й объект. Суть алгоритма состоит в объединении D так, чтобы каждая xi попала только в один k раздел. В результате образуется кластерный составной вектор m длиною N, где mi – номер кластера xi. Параметр k – входное значение для работы алгоритма и является конечным числом кластеров [4].

Методы этого вида находят широкое применение из-за простоты реализации на алгоритмических языках и большого быстродействия. Это

итеративный алгоритм, который делит данное множество объектов на k кластеров, элементы, которых являются максимально приближенными к их центрам, а сама кластеризация происходит за счет смещения этих же центров.

Процесс простейшей кластеризации методом k-средних состоит из следующих шагов:

1. Выбрать в пространстве объектов точки, которые будут центроидами (точками в центре кластера) соответствующих k кластеров. Выборка начальных центроидов может быть, как случайной, так и по определенному алгоритму.
2. В цикле, который продолжается до тех пор, пока центроиды кластеров не перестанут изменять свое положение, оцениваем каждый объект и смотрим, к какому центроиду какого кластера он является близлежащим.
3. Если найден близлежащий центроид, привязываем объект к кластеру этого центроида.
4. Когда все объекты перебраны, высчитываем новые координаты центроидов k кластеров.
5. Проверяем координаты новых центроидов. Если они соответственно равны предыдущим центроидам — выходим из цикла, кластеризация завершена, если нет, возвращаемся к пункту 2.

Результаты, полученные в данной работе путем сбора данных и их кластеризации, будут основой при дальнейшем хранении, извлечении, манипулировании, анализе и выводе данных с географической привязкой в режиме реального времени. С помощью этих данных можно узнать местоположение человека в конкретное время, либо увидеть проблемы загруженности дорог.

Литература

1. International Conference on Big Data for Official Statistics Organized by UNSD and NBS China 28-30 October 2014, Beijing China Concept Note (as of 13 August 2014)
2. Майер-Шенбергер Виктор Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим.
3. Крылов В.В., Крылов С.В. Большие данные и их приложения в электроэнергетике от бизнес-аналитики до виртуальных электростанций 2013. – 137 с.
4. Чубукова И.А. Data Mining - Интернет-университет информационных технологий, Бином. Лаборатория знаний ISBN 978-5-94774-819-2; 2008 г.