УДК 004.891

Н. В. Маликов, А. П. Демиденко
E-mail: BARCAman93@gmail.com ad48@mail.ru
Кыргызско-Российский (Славянский) университет. г.Бишкек, Кыргызстан

АЛГОРИТМ ОПРЕДЕЛЕНИЯ НАИБОЛЕЕ ВЫГОДНЫХ РАЙОНОВ ДЛЯ ОТКРЫТИЯ БИЗНЕСА НА ОСНОВЕ DATA MINING КЛАСТЕРИЗАЦИИ

В работе проанализированы данные сети предприятий общественного питания города, рассмотрен алгоритм data mining обработки и визуализации данных на географической карте города, а также разработано приложение, которое позволяет решать поставленную задачу определения выгодных для открытия бизнеса районов города.

Ключевые слова: data mining, кластеризация, статистический анализ, обработка больших данных, интерактивная географическая карта.

Введение

С развитием IT-технологий интернет прочно вошел в нашу жизнь, а чем больше информации мы имеем, тем сложнее превратить эту информацию в полезную. В информационный век успех компаний и отдельных людей все чаще зависит от того, как быстро и эффективно они превращают огромные объемы данных в полезную информацию. Причем отбор информации должен происходить быстро и качественно, поскольку пользователь не желает тратить время на дополнительную фильтрацию полученных данных, а хочет увидеть именно то, что ему надо, поэтому проблема анализа и обработки данных весьма актуальна в наше время.

Постановка задачи

Имеется набор филиалов $F = \{f_1, f_2, ..., f_n\}$ сети общественного питания. Каждый из них можно представить, как вектор $[e_n, latitude_n, longitude_n]$, где e_n – выручка, $latitude_n$ – широта, $longitude_n$ – долгота. Нужно добавить еще один филиал f_{n+1}, и изменяя переменные $latitude_{n+1}$ и $longitude_{n+1}$, добиться того, чтобы его выручка e_{n+1} была максимальной

$$e_{n+1} = \Phi(latitude_{n+1}, longitude_{n+1}) -> max \quad (1)$$

Для данного предприятия существует система, которая передает информацию о результатах функционирования филиалов в центральный офис. Эта информация включает в себя сведения о доходах и расходах. Для решения поставленной задачи предлагается использовать географические координаты филиалов, а также данные, полученные за период функционирования системы.

Решение

В качестве основного подхода к решению проблемы был выбран подход Data Mining [4]): кластерный анализ и статическая обработка геоинформационных объектов.

- Задача кластеризации решает задачу группировки филиалов сети по признаку близости географического расположения.
- Статистическая обработка позволяет представить объекты внутри кластера по эффективности функционирования.
- Геоинформационный подход решает задачу визуализации рассматриваемых объектов на карте города.

Кластерный анализ – это метод классификационного анализа; его основное назначение – разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры. Это многомерный статистический метод, поэтому предполагается, что исходные данные могут быть значительного объема, т.е. существенно большим может быть как количество объектов исследования (наблюдений), так и признаков, характеризующих эти объекты. Большое достоинство кластерного анализа в том, что он дает возможность производить разбиение объектов не по одному, а по ряду признаков. Кроме того, кластерный анализ, в отличие от большинства математико-статистических методов, не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходных данных практически произвольной природы. [3]

Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятие решений. С этой целью применяется сжатие данных, если исходная выборка избыточно большая, то можно сократить её, оставив по одному, наиболее типичному представителю от каждого кластера; обнаружение новизны – выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных. Таким приемом выгодно пользоваться в бизнесе, выделяя группы филиалов и разрабатывая для каждой из них отдельную стратегию развития.

С целью выбора метода кластеризации нами было рассмотрено множество различных методов. В итоге была найдена и изучена статья “Кластеризация точек на основе регулярной сети” [5] и впоследствии метод кластеризации, описанный в данной статье, был адаптирован под имеющиеся и требуемые данные.

Филиал – исследуемый объект алгоритма, представляет собой численный показатель, характеризующий выручку. При создании кластера ему присваивается значение выручки первого филиала, входящего в него. При добавлении следующего филиала вычисляется среднее значение для кластера. Вычислять прибыль для кластера требуется для визуализации результатов на интерактивной географической карте города.

Описание алгоритма

В данной работе разбиение на кластеры происходит с учетом географических координат филиалов. Входными данными являются:

- данные о филиалах, содержащие координаты – *двумерный массив,* наименование – *строка,* выручка за выбранный период – *число* и др.
- R – радиус точки воздействия, который является пределом максимального расстояния между филиалами и указывается в пикселях относительно минимального масштаба карты 1:2500 и увеличивается с увеличением масштаба. Данный показатель может меняться пользователем и по умолчанию имеет значение 10px.

$$d = |x.latitude - x_1.latitude| + |x.longitude - x_1.longitude|, \quad (2)$$

где *latitude* – значение широты, *longitude* – значение долготы рассматриваемых филиалов, а *d* расстояние между ними.

Этапы алгоритма кластеризации

- Берется филиал *x*, если он и его соседние филиалы не привязаны к кластеру, то создается новый кластер, и филиал добавляется в него.
- Берется соседний филиал x_1 и проверяется наличие его в кластерах.
- По формуле (2) находится расстояние между филиалом *x* и x_1, если расстояние <= *R*, где *R* – радиус точки воздействия, то возможны два случая:
  - если филиал x_1 пока не входит в кластер, то x_1 вносится в тот же кластер в котором находится филиал *x*;
  - если филиал x_1 уже принадлежит какому-то кластеру, отличному от текущего, то кластер с большим количеством объектов поглощает кластер с меньшим.
- Если расстояние больше *R*, то создается новый кластер, и филиал x_1 добавляется в него.

Программные средства для определения наиболее выгодных районов

Данный алгоритм использован для разработки web приложения, позволяющего найти оптимальный район расположения нового филиала на географической карте г.Бишкек в заданном масштабе. Обозначение кластеров по среднему значению эффективности функционирования отображено изменением цветовой гаммы. Интенсивность цветовой окраски связана с выручкой района: чем выше выручка, тем интенсивнее окраска.

Основные возможности программы представлены на рисунке 1.

Рис. 1. Основные возможности программы.

На рисунке 2 представлен пример визуализации в разных масштабах данных, полученных в процессе разработки:

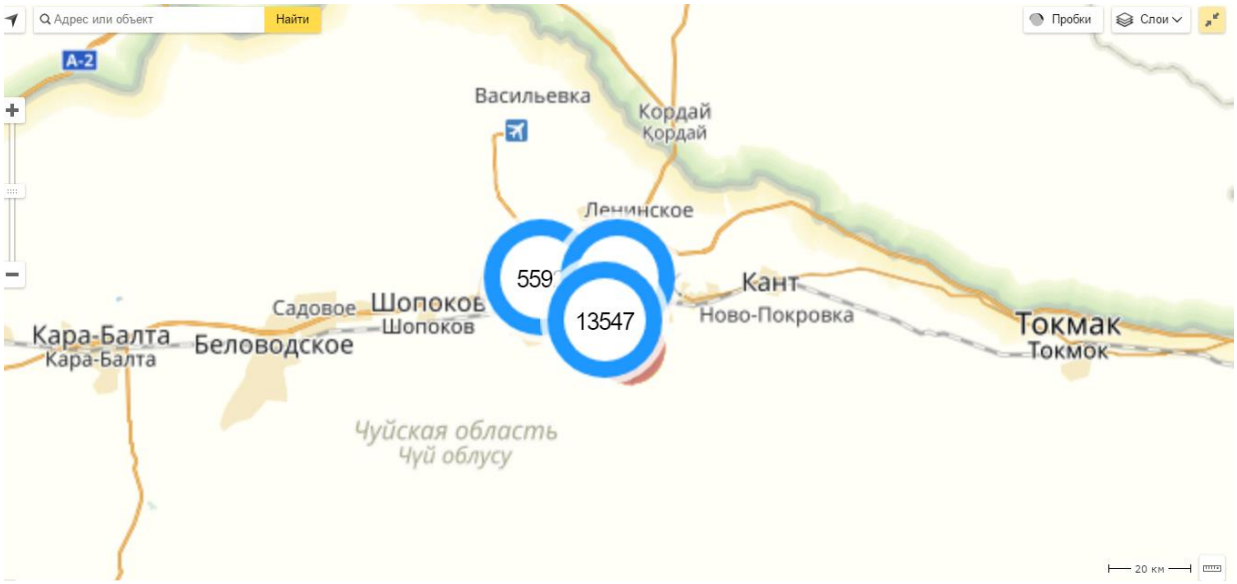

Рис. 2. Пример визуализации результата анализа данных

На рисунке 3 приведен пример изменения данных по кластерам при изменении масштаба карты.

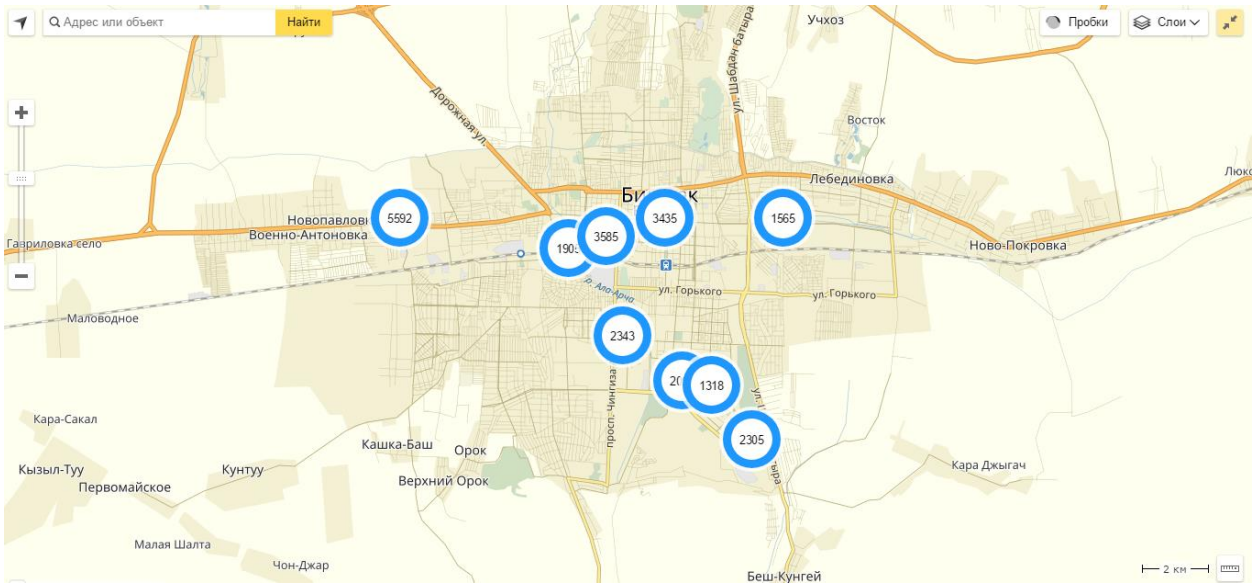

Рис. 3. Пример визуализации результата анализа данных в меньшем масштабе.

Визуализация полученных данных с помощью тепловой карты, изображенная на рисунке 4, позволяет по насыщенности цвета полагать о наиболее прибыльных филиалах и районах, но не показывает численные значения.

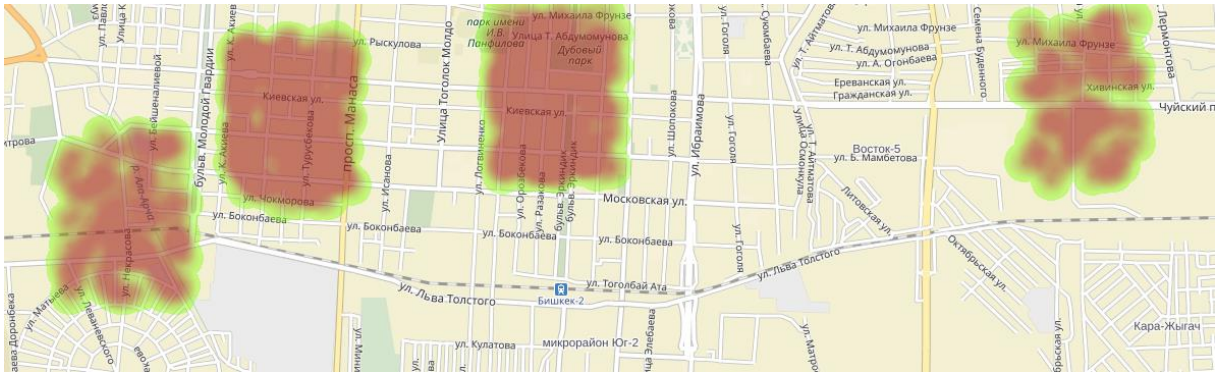

Рис. 4. Пример визуализации результата с помощью тепловой карты.

Заключение

Разработанное приложение имеет доступ к данным по всем филиалам. Обработка данных для быстрого получения результата происходит на выделенном сервере. Данные по выручке филиалов берутся за определенный период. При изменении периода карта автоматически обновляется, и кластерный анализ выполняется на основе обновленных данных. Результаты анализа данных можно получить в любое время за любой выбранный период функционирования предприятия.

Литература

1. Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004.
2. Olson C.F. “Parallel Algorithms for Hierarchical Clustering” Parallel Computing, 1995, Vol. V21, P. 1313-1325.
3. Кластерный анализ. Url: https://ru.wikipedia.org/wiki/Кластерный_анализ (дата обращения 3.02.2017).
4. Data mining. Url: https://ru.wikipedia.org/wiki/Data_mining (дата обращения 20.01.2017).
5. Кластеризация точек на основе регулярной сети. Url: https://habrahabr.ru/post/138185/ (дата обращения 10.01.2017).