УДК 004.622

В.Б. Иванов, В.Н. Гиляров, А.А. Мусаев

ПРОСТРАНСТВЕННАЯ КЛАСТЕРИЗАЦИЯ МЕСТ ВОЗНИКНОВЕНИЯ ЧРЕЗВЫЧАЙНЫХ СИТУАЦИЙ

Иванов В.Б., Гиляров В.Н., Мусаев А.А. **Пространственная кластеризация мест возникновения чрезвычайных ситуаций.**

Аннотация. Обоснована актуальность пространственной кластеризации мест возникновения чрезвычайных ситуаций. Представлен адаптированный алгоритм STING кластеризации. На основе сравнительного анализа описан выбор вида базы данных для хранения результатов работы алгоритма. Предложен алгоритм подготовки данных для последующей визуализации.

Ключевые слова: Пространственная кластеризация, дерево квадрантов, графовая база данных.

Ivanov V.B., Gilyarov V.N., Musaev A.A. **Spatial clustering of emergency situations locations.**

Abstract. The topicality of spatial clustering of emergency situations locations is substantiated. The adapted STING clustering algorithm is presented. The work has a description of a choice of a database for storing clustering results based on comparative analysis of alternative storages. An algorithm for data preparation for following visualization is offered.

Keywords: spatial clustering, quadtree, graph database.

1. Введение. Важной составляющей комплекса программного обеспечения ситуационных центров единой дежурно-диспетчерской службы 112 является система мониторинга мест возникновения чрезвычайных ситуаций. В ее задачи входит формирование актуальной картины событий в целом регионе, в частности, отображение очагов преступности или аномальной активности при помощи геоинформационной системы. Так актуальной задачей становится разработка алгоритма кластеризации мест происшествий для их последующей визуализации.

В работе решаются следующие задачи:

1) задача пространственной кластеризации, где X – множество мест возникновения чрезвычайных ситуаций, Y – множество очагов возникновения чрезвычайных ситуаций (кластеры). Задана функция расстояния между местами возникновения чрезвычайных ситуаций. Имеется конечная обучающая выборка $X^m = \{x_1, x_2, ..., x_m\}$. Требуется разбить выборку на кластеры, так, чтобы каждый кластер состоял из мест возникновения чрезвычайных ситуаций, близких по метрике, а объекты разных кластеров существенно отличались;

2) задача динамического перестроения кластеров при поступлении информации о новых чрезвычайных ситуациях;

3) задача выборки мест происшествий и кластеров по географическому фильтру;

4) задача хранения результатов кластеризации в базе данных;

5) задача подготовки результатов кластеризации путем применения алгоритма построения выпуклой оболочки в двумерном пространстве для последующего отображения кластеров на карте.

Выполнение перечисленных выше задач обеспечивает подсистема кластерного анализа, структура которой показана на рис. 1. Цифрами на рисунке отмечены номера задач, решаемых модулями.

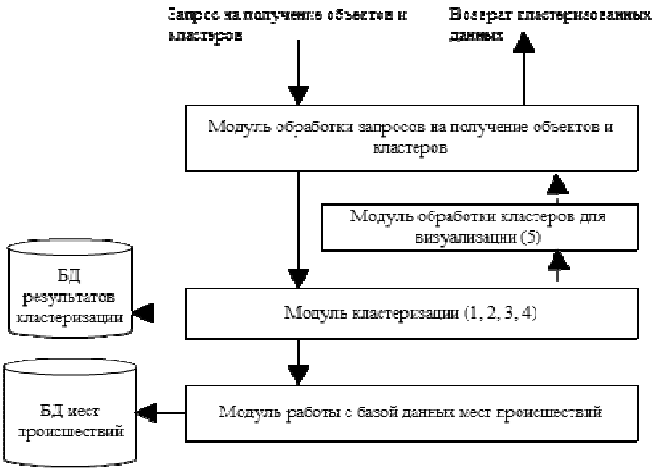

Рис. 1. Структура подсистемы кластерного анализа.

2. Выбор алгоритма кластеризации. К алгоритмам и структурам данных для хранения результатов пространственной кластеризации предъявляются дополнительные требования:

1) число кластеров не должно являться входным параметром алгоритма;

2) результаты работы алгоритма должны храниться в иерархических структурах данных для быстрого получения кластеров при изменении уровня приближения к карте и изменении участка отображения карты;

3) должна быть предусмотрена возможность перестроения кластеров при поступлении новых данных без необходимости перезапуска алгоритма на всем объеме данных.

В таблице представлена пространственная и временная сложность алгоритмов, пригодных для решения задачи пространственной кластеризации, где k означает число кластеров, t - число итераций, N - число объектов для кластеризации, L - число уровней в алгоритме STING [1].

Пространственная и временная сложность алгоритмов

Алгоритм	Пространственная сложность	Временная сложность
k-средних	$O(N^2)$	$O(tkN)$
DBSCAN	$O(N^2)$	$O(N^2)$
Алгоритм одной связи	$O(N^2)$	$O(kN^2)$
STING	$O(L)$	$O(NL)$

Для решения поставленных задач был выбран алгоритм STING, так как он обладает наибольшим быстродействием из всех рассмотренных алгоритмов, а результаты его работы легко представимы в иерархических структурах данных.

3. Использование дерева квадрантов. Алгоритм STING предполагает предварительное разбиение объектов кластеризации при помощи структуры данных «дерево квадрантов» [2]. Такое разбиение для рассматриваемой предметной области состоит из следующих шагов:

1) Создать набор уровней, соответствующий числу уровней приближения к карте. Первый уровень состоит из одного квадранта. На каждом нижележащем уровне квадрант разбивается на 4 части.

2) Распределить места возникновения происшествий по квадрантам вплоть до нижнего уровня.

Сложность такого разбиения является линейной и равна $O(NL)$.

Фрагмент дерева квадрантов представлен на рис. 2. Число внутри квадранта означает количество мест происшествий, координаты которых находятся внутри квадранта.

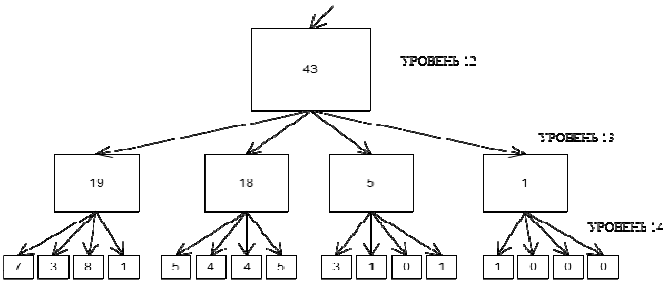

Рис. 2. Фрагмент дерева квадрантов после распределения географических объектов.

Каждый уровень дерева соответствует уровню приближения к карте. Современные картографические сервисы, такие как Google Maps и OpenStreetMap предполагают 18 уровней приближения. Поиск кластеров на определенном уровне сводится к поиску смежных квадрантов, число объектов внутри которых выше некоторого установленного значения. Соединение квадрантов в кластеры осуществляется алгоритмом взвешенного быстрого объединения [3].

При помощи такой структуры данных можно быстро осуществлять получение объектов и кластеров для определенного участка карты по заданному географическому фильтру. Для этого осуществляется проверка квадрантов на предмет пересечения с заданным участком карты на определенном уровне приближения. Если пересечение найдено, то объекты квадранта и кластеры, включающие в себя текущий квадрант, должны быть возвращены в качестве результата фильтрации.

Дерево квадрантов позволяет производить динамическое перестроение кластеров без перезапуска алгоритма на всем множестве объектов. Добавление или удаление объекта изменяет число объектов в квадранте, после чего алгоритм быстрого взвешенного объединения запускается только для текущего квадранта.

4. Хранение результатов кластеризации. На сегодняшний день существует множество типов хранилищ данных для пространственной информации. Для хранения результатов кластеризации, представленных в виде дерева квадрантов, наилучшим образом подходят графовые базы данных, преимущество которых заключается в хранении прямых ссылок связанных объектов друг на друга. Так, в

графовой базе данных поиск дочернего квадранта осуществляется за время $O(1)$. В случае же реляционной базы данных поиск по вторичному ключу может занимать время от $O(\log N)$ (при индексации ключа) до $O(N)$ (без индексации).

Перед отображением кластеров на карте необходимо вычислить координаты области, которую необходимо визуализировать. Для решения этой задачи был выбран алгоритм построения выпуклой оболочки Грэхема, входящий в набор средств географического анализа графовой базы данных Neo4j.

Рис. 3. Результат работы кластеризации при уровне приближения 13.

5. Заключение. Разработанные алгоритмы пространственной кластеризации и иерархические структуры данных внедрены в подсистеме мониторинга мест возникновения чрезвычайных ситуаций в Курской и Свердловской области.

Литература

1. *Марманис Х., Бабенко Д.* Алгоритмы интеллектуального Интернета. Передовые методики сбора, анализа и обработки данных. СПб.: Символ-Плюс, 2011. 480 с.
2. *Yannis Manolopoulos, Alexandros Nanopoulos* R-Trees: Theory and Applications. Springer, 2006.
3. *Роберт Седжвик, Кевин Уэйн* Алгоритмы на Java. М. : ООО "И.Д. Вильямс", 2012. 848 с.

Иванов Виталий Борисович — аспирант кафедры систем автоматизированного проектирования и управления, Санкт-Петербургский государственный технологический институт (технический университет), инженер-программист компании НТЦ Протей. Область научных интересов: анализ данных, программирование, высокопроизводительные вычисления. Число научных публикаций — 7. vitalyivanov88@gmail.com, http://protei.ru; НТЦ Протей, Большой Сампсониевский

проспект, д.60А, г. Санкт-Петербург, 194044, РФ; р.т. +7 (812) 449-47-27 доб. 5057. Научный руководитель — В.Н. Гиляров.

Ivanov Vitaly Borisovich — postgraduate student of subdepartment of systems for computer-aided design and control of Saint-Petersburg state institute of technology (technical university), software engineer at company SRC Protei. Research interests: data analysis, programming, high-performance computing. The number of publications — 7. vitalyivanov88@gmail.com, http://protei.ru; SRC Protei, Bolshoy Sampsonievskiy prospect, 60A, Saint-Petersburg, 194044, Russia; office phone +7 (812) 449-47-27 add. 5057. Tutor – V.N. Gilyarov.

Гиляров Владимир Николаевич — к.т.н., доцент кафедры систем автоматизированного проектирования и управления, Санкт-Петербургский государственный технологический институт (технический университет). Область научных интересов: применение методов "мягких вычислений" (нечеткая логика, искусственные нейронные сети, генетические алгоритмы)для классификации, диагностики и прогнозирования в технических системах. Число научных публикаций — 114. giljarow@mail.ru; Санкт-Петербургский государственный технологический институт (технический университет), Московский пр., д.26, г.Санкт-Петербург, 190013, РФ; р.т. +7 (812) 4949370

Gilyarov Vladimir Nickolaevich — candidate of technic science, associate professor of subdepartment of systems for computer-aided design and control of Saint-Petersburg state institute of technology (technical university). Research interests: soft computing (fuzzy logic, artificial neural networks, genetic algorithms) for classifying, diagnostics, and prediction at technical systems. The number of publications — 114. giljarow@mail.ru; Saint-Petersburg state institute of technology (technical university), Moskovskiy prospect, 26, Saint-Petersburg, 190013, Russia; office phone +7 (812) 4949370

Мусаев Александр Азерович — д.т.н., профессор; ведущий научный сотрудник научно-исследовательской группы информационных технологий в образовании СПИИРАН, научный консультант ОАО Специализированная инжиниринговая компания «Севзапмонтажавтоматика». Область научных интересов: анализ данных, управление и прогнозирование в сложных динамических системах, стохастические хаотические системы. Число научных публикаций — 188. amusaev@szma.com, www.szma.com; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)350-5885, факс +7 (812)350-1113.

Musaev Alexander Azerovich — Dr. in Appl. Math., professor; leading researcher, Education Information Technology Group, SPIIRAS, expert, public corporation Specialized engineering company "Sevzapmontageautomatica". Research interests: data analysis, complicated dynamic systems prognosis and control, stochastic chaos systems. The number of publications — 188. amusaev@szma.com, www.szma.com; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)350-5885, fax +7(812)350-1113.

Рекомендовано ИГИТО СПИИРАН, рук. ктн, доц. А.В. Тишков.
Статья поступила в редакцию 20.12.2012.

РЕФЕРАТ

Иванов В.Б., Гиляров В.Н., Мусаев А.А. **Пространственная кластеризация мест возникновения чрезвычайных ситуаций.**

Современные системы мониторинга географических объектов сталкиваются с проблемой отображения десятков тысяч сущностей на карте. Примером такой системы является система мониторинга мест возникновения чрезвычайных ситуаций службы 112. В ее задачи входит отображение мест возникновения чрезвычайных ситуаций в целом субъекте Российской Федерации. Перерисовка десятков тысяч объектов при навигации в пользовательском интерфейсе системе занимает длительное время. Распространенным решением проблемы отображения географических объектов является их предварительная кластеризация с целью отрисовки мест скопления объектов, а не отдельных объектов. Однако, не существует единственного алгоритма кластеризации, подходящего для решения этой задачи. Его выбор зависит от предъявляемых к системе требований.

К алгоритму кластеризации в составе системы мониторинга службы 112 предъявляются следующие требования: алгоритм должен уметь динамически перестраивать кластеры при поступлении информации о новых чрезвычайных ситуациях. Результаты кластеризации должны храниться таким образом, что на них можно было бы наложить географический фильтр. Результаты кластеризации должны храниться на энергонезависимом носителе для исключения затрат на перестроение кластеров при перезапуске системы. При этом результаты кластеризации должны выгружаться в оперативную память для наискорейшего доступа к ним.

Для удовлетворения этих требований нужно разработать алгоритм кластеризации с дальнейшим сохранением результатов в структуры данных пригодные для географической фильтрации. Также требуется выбрать хранилище данных для наискорейшего доступа к хранимым объектам.

В связи с этим предлагается использовать алгоритм кластеризации на основе решеток с хранением результатов его работы при помощи структуры данных «дерево квадрантов». В качестве хранилища данных рассматривается использование графовой базы данных neo4j, как наиболее подходящей для хранения древообразных структур данных. Также neo4j позволит использовать надстройку для работы с пространственными данными, включающую алгоритм построения выпуклой оболочки в двумерном пространстве для последующего отображения кластеров на карте, и встроенные механизмы выгрузки данных в оперативную память для наискорейшего доступа к ним.

SUMMARY

Ivanov V.B., Gilyarov V.N., Musaev A.A. **Spatial clustering of emergency situations locations.**

Modern systems of geographical objects monitoring are faced with problem of displaying of tens of thousands of entities on the map. An example of such a system is a system for monitoring emergency locations of service 112. Its objective is to display locations of emergency situations in a whole region of the Russian Federation. Redrawing of thousands of objects in the user interface during a navigation takes a long time. A common solution of this problem is to cluster locations in order to render the concentrations of objects rather than individual objects. However, no single clustering algorithm is suitable for this task. Choice of algorithm depends on system requirements.

Requirements to the clustering algorithm in a monitoring system of a service 112 are the following: the algorithm should be able to dynamically rearrange the clusters when the information about the new emergency situation comes in. Clustering results should be stored in such a way that they can be filtered by geographical attachment. Clustering results should be stored in non-volatile storage to avoid the cost of rebuilding the clusters during a system restart. During the work the results should be stored in the random access memory for the fastest access.

To meet these requirements it is necessary to develop a clustering algorithm to further saving the results in a data structures suitable for the geographic filtering. Also it is needful to choose data storage for the quickest access to stored objects.

Therefore, it is proposed to use a grid-based clustering algorithm that saves the results of its work with data structures called “tree quadrants”. As a data storage it is proposed to use a graph database neo4j, as the most suitable for storing tree based structures. Neo4j also allows to use an add-in for working with spatial data, including a convex hull algorithm to prepare clusters for displaying on the map, and built-in mechanisms for uploading data into RAM for fastest access.