Интеллектуальный анализ больших пространственных данных в условиях неопределенности

А. Р. Гараева[1], Ф. Т. Махмутова[2], И. В. Аникин[3]
КНИТУ-КАИ имени А. Н. Туполева
[1]argaraeva@stud.kai.ru, [2]ftmakhmutova@stud.kai.ru, [3]anikinigor777@mail.ru

***Аннотация*. Пространственно-временные данные часто бывают неточными и неопределенными, особенно это характерно для потоков пространственных данных, поступающие от ГИС-служб в режиме реального времени, из-за задержек между GPS сигналами. К тому же, пространственно-временные данные относятся к категории больших данных и должны обрабатываться соответствующим образом. Однако такие инструменты обработки больших данных, как Apache Spark и Apache Hadoop, не предоставляют функции по обработке неопределенных больших пространственных данных. Цель данной работы дать обзор существующих подходов обработки пространственных данных в условиях неопределенности, а также сделать обзор существующих программных средств распределенной обработки данных, позволяющих обрабатывать большой объем пространственных информации.**

Ключевые слова: пространственный анализ в условиях неопределенности; большие неопределенные данные; пространственные запросы; нечеткая кластеризация; анализ пространственных шаблонов

I. Введение

На сегодняшний день цифровые устройства производят огромное количество пространственно-временных данных. В последние годы ученые-исследователи все больше заинтересованы в обработке таких данных и в получении полезных и потенциально важных знаний из них. Пространственный анализ относится к категории Data Mining задач, и представляет собой процесс получения скрытых знаний, пространственных отношений и интересных пространственных шаблонов. Пространственно-временные данные часто являются неточными и неопределенными, особенно потоковые данные, поступающие от ГИС-служб в режиме реального времени из-за задержек между сигналами. Поэтому, обрабатывая такие неопределенные пространственные данные, необходимо учитывать следующие особенности: пространственно-временных данные неточны и относятся к категории больших данных. Однако современные технологии анализа больших данных, такие как, Apache Hadoop и Apache Spark не поддерживают встроенные функции для обработки пространственных данных. Таким образом, основная цель данной работы – дать обзор существующих подходов обработки пространственных данных в условиях неопределенности, а также сделать обзор существующих программных средств распределенной обработки данных, позволяющих обрабатывать большой объем данных.

II. Определение неточных пространственных данных

Пространственно-временные данные часто бывают неточными и неполными. Неопределенность данных означает, что хотя бы один из параметров объекта является неопределенным (неточным). В случае пространственных данных, неопределенность подразумевает неточную геолокацию объекта. Согласно классическому определению, параметр неопределенности может быть указан с помощью двух моделей: дискретной и непрерывной. В дискретной модели геолокация неопределенного объекта определяется применением функции вероятности (англ. probability mass function), которая определяет вероятность нахождения пространственного объекта в одном из возможных местоположений. В непрерывной модели для определения геопозиции объекта, используется функция вероятности для непрерывной случайной переменной (англ. probability density function). Разница между этими моделями – это количество возможных значений параметров. Бесконечное для непрерывных моделей и конечное для дискретных.

III. Подходы интеллектуального анализа пространственных данных в условиях неопределенности

В научной литературе выделяют следующие подходы интеллектуального анализа пространственных данных в условиях неопределенности: нечеткие пространственные запросы (англ. uncertain spatial data querying) [1, 2], нечеткая пространственная кластеризация (англ. uncertain spatial clustering analysis) [3–8], анализ пространственных шаблонов в условиях неопределенности (англ. co-location pattern mining on uncertain spatial data) [9–13].

A. Нечеткие пространственные запросы

Пространственные запросы позволяют выполнять выборки пространственных объектов, с точки зрения их пространственного расположения. Первоначально неопределенные пространственные запросы были

предложены в работе [1], наиболее важные из этих запросов: запрос меры удаленности (англ. range query) и запрос ближайшего соседа (англ. nearest neighbor query).

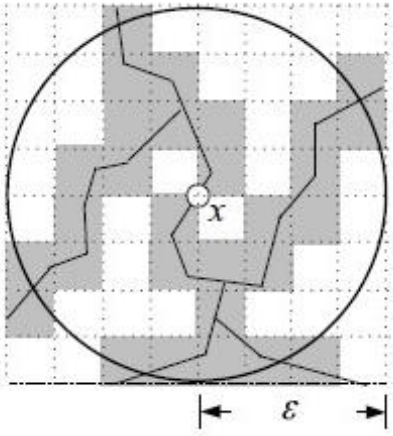

Рис. 1. Нечеткий пространственный объект

Range Query предназначен для извлечения всех пространственных объектов, которые расположены в некоторой интересующей области (англ. region of interest). Пример задачи, которая может быть решена использованием Range Query запроса: получить список ресторанов, расположенных в Нью-Йорке.

K – Nearest Neighbor Query используется для извлечения k-ближайших соседей в географической близости от предопределенного пространственного объекта. Для выполнения этих запросов в условиях неопределенности, местоположение объекта задается некоторой неопределенной областью R (англ. uncertain region), которая охватывает фактическое расположение объекта. Формально, область неопределенности R определяется как (1). Пример неопределенной области представлен на рис. 1, где x – последнее известное местоположение объекта, а E – максимальное отклонение от текущей позиции до x.

$$\int_{x \in R} O.pdf(x)dx = 1 \qquad (1)$$

Range Query в условиях неопределенности (англ. fuzzy range query) предназначен для решения задач следующего типа, например, получить список всех такси, находящиеся в пределах 5 км от текущего положения с вероятностью не менее 60%, где месторасположение такси является неопределенным. Формально это можно определить следующим образом (1). Предположим, что S – множество неопределенных пространственных объектов, а q – точка запроса, E – положительное отклонение, fuzzy range query определяет для каждого пространственного объекта вероятность того, что расстояние до точки q меньше E (2):

$$PRQ(q,S) = \{U \in S, P(dist(q,U) < \varepsilon)\} \qquad (2)$$

Поэтому основная задача запроса fuzzy range query состоит в том, чтобы найти вероятность $P(dist(Q,U) < \varepsilon)$ для неопределенных пространственных объектов. Nearest Neighbor Query запрос в условиях неопределенности предназначен для оценки вероятности того, что неопределенные пространственные объекты A ближе к неопределенным пространственным объектам B, чем другие неопределенные объекты C. Формально предположим, что S представляет собой набор неопределенных пространственных объектов, q – заданный неопределенный пространственный объект и δ – порог вероятности. Поэтому Fuzzy Nearest Neighbor Query находит все объекты из $S(o \in S)$, которые удовлетворяют условию $P_{fnn}(o,q) \geq \delta$, где $P_{fnn}(o,q)$ – вероятность того, что o – ближайший сосед для q [2].

B. Пространственная кластеризация в условиях неопределенности

Кластеризация – это разбиение некоторого начального набора немаркированных объектов данных на разные подмножества (кластеры), где члены одного кластера обладают одинаковыми свойствами. Как правило, кластеризация предполагает, что каждый объект принадлежит не более чем одному кластеру, но неопределенность в терминах кластеризации говорит о том, что каждый объект может принадлежать нескольким кластерам одновременно, и, обычно, это делается путем вычисления вероятности принадлежности объекта к каждому кластеру (рис. 2).

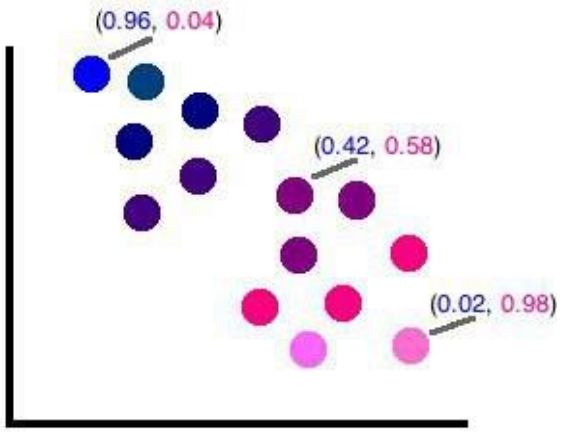

Рис. 2. Пример нечеткой кластеризации

Первый программный комплекс по нечеткой кластеризации был предложен ученым-исследователем E.R. Ruspini в 1969 году [3]. На сегодняшний день, наиболее распространенным алгоритмом нечеткой кластеризации является метод нечеткой кластеризации C-средних (FCM), который был введен в 1973 году [4], а затем усовершенствован ученым J.C. Bezdek в работе [5].

Алгоритм FCM основан на минимизации целевой функции (3):

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, 1 \leq m < \infty \qquad (3)$$

где $m:(m \geq 1)$ определяет меру нечеткости кластера, $u_{ij}:(0 \leq u_{ij} \leq 1)$ представляет меру принадлежности x_i по отношению к кластеру j, x_i – это i^{th} объект данных измерения, N – количество объектов данных, C – количество кластеров, c_j – это центр кластера j.

Алгоритм FCM требует обновления центров кластеров (4) и обновления мер принадлежности (5):

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (4)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (5)$$

В настоящее время алгоритм нечеткой кластеризации С-средних используется в различных областях интеллектуального анализа данных [6], таких как сегментация изображений, распознавание изображений, анализ ДНК и многих других. Однако в основном кластеризация используется для обработки пространственных данных. Например, для дистанционного зондирования, для грамотного ведения сельского хозяйства [7] для оценки социальной уязвимости на уровне поселков [8] и т. д.

Взятие в рассмотрение того, что обрабатываемые данные могу являться нечеткими, позволяет нам выбирать соответствующие алгоритмы нечеткой кластеризации для их анализа, тем самым предоставляя возможность извлечения более значимой информации и принятия более точных решений.

Интересно, что применение техник нечеткой кластеризации позволяет нам частично решить задачу выявления шаблонов в условиях пространственной кластеризации, которая будет рассмотрена в следующем разделе.

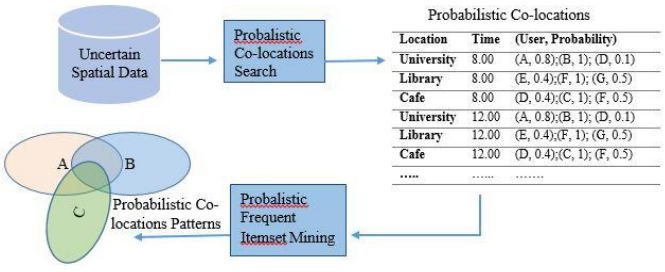

Рис. 3. Анализ вероятностных пространственных шаблонов

C. *Анализ пространственных шаблонов в условиях неопределённости*

Анализ пространственных шаблонов сосредоточен на поиске интересных отношений между пространственными объектами, на обнаружении тенденций некоторых пространственные объектов находиться в непосредственной географической близости [9]. В нашей предыдущей работе [10] по обнаружению шаблонов совместного размещения данных мы не рассматривали свойство нечеткости входных данных. В работе [11] была предложена вероятностная модель анализа пространственных шаблонов, где неопределенность каждого пространственного объекта характеризуется вероятностью находиться в некоторой точке. Примером такой задачи может быть следующее [2]: нахождение группы людей, которые имеют тенденцию находиться в одних и тех же местах, где геопозиция каждого человека представляет собой вероятность нахождения данного человека в предопределенном месте и времени. Общий процесс анализа таких шаблонов представлен на Рис. 3. На первом этапе необходимо найти все пространственные объекты, которые имеют наибольшую вероятность нахождения в пространственной близости друг к другу. Этот этап решается применением fuzzy range query запроса. Второй шаг – найти вероятностные пространственные шаблоны путем применения алгоритмов машинного обучения [12]. Подход анализа частых шаблонов (англ. frequent patterns) в условиях неопределенности был представлен в работе [13].

IV. Обзор программных средств для обработки больших пространственных данных

Неопределенные пространственные данные относятся к категории больших данных, что требует применения технологий распределённой обработки. Тем не менее, стек технологий для обработки больших данных не поддерживает встроенных функций для анализа пространственных данных. Поэтому были разработаны библиотеки поверх платформы Apache Spark, предоставляющие API для обработки пространственных данных: SpatialSpark [14], GeoSpark [15], Magellan [16], LocationSpark [17]. Однако перечисленные библиотеки не предназначены для обработки неопределенных данных.

В работах [18] и [19] были представлены библиотеки Elite и TrajSpark, соответственно, как средства распределенной обработки пространственно-временных траекторий. Сравнительный анализ программных средств (таблица) показал, что на данный момент отсутствуют средства распределенной обработки неопределенных пространственных данных.

ТАБЛИЦА I СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРОГРАММНЫХ СРЕДСТВ РАСПРЕДЕЛЕННОЙ ОБРАБОТКИ ДАННЫХ

Distributed Processing Tool	Type	Certain Spatial Data	Uncertain Spatial Data	Comments
Spatial Spark	RAM	Indexing, Range Queries		
GeoSpark	RAM	Indexing, Range Queries, Co-Location Pattern Mining		
Magellan	RAM	Indexing, Range Queries		
LocationSpark	RAM	Indexing, Range Queries		Provides Spatial Bloom Filter
Elite	Disk		Range Query, Storage	Uncertain trajectories mining
TrajSpark	RAM	Indexing, Range Queries over trajectory data		

V. ЗАКЛЮЧЕНИЕ

В этой работе мы провели обзор существующих подходов анализа больших пространственных данных в условиях неопределенности, а именно, рассмотрели нечеткие пространственные запросы, нечеткие методы пространственной кластеризации, методы анализа пространственных шаблонов в условиях неопределенности. Интересно, что в научных трудах, описывающих вышеупомянутые подходы, не затрагивается тема объема данных анализируемых этими подходами, что говорит о неизученности свойства масштабируемости этих подходов, которое временами является решающим, поскольку обрабатываемый набор данных может быть очень большим.

Обзор программных средств по интеллектуальному анализу пространственных данных показал, что на данный момент не существует подходящих инструментов для обработки больших данных. Что порождает спрос на разработку масштабируемых подходов для интеллектуального анализа нечетких пространственных данных. Поэтому, в будущем, мы планируем разработать библиотеку для платформы Apache Spark, которая будет иметь возможность обрабатывать большие пространственные данные.

СПИСОК ЛИТЕРАТУРЫ

[1] Tao Y., Range and Nearest Neighbor Queries on Uncertain Spatiotemporal Data, *Manaaging and mining uncertain data*, Kluwer Academic Publishers, Boston/Dordrecht/London, 2009, pp. 327-351.

[2] Zufle A. *Similarity Search and Mining in Uncertain Spatial and Spatio-Temporal Databases*, Diss., Munich, 2013. 397p.

[3] Huang Y., Shekhar S., Xiong H. Discovering Co-location Patterns from Spatial Datasets: A General Approach, IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 16, i. 12, pp. 1472-1485. DOI: 10.1109/TKDE.2004.90.

[4] Garaeva A., Makhmutova F., Anikin I., Sattler K.-U. A Framework for Co-location Patterns Mining in Big Spatial Data, *Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements*, Saint Petersburg, DOI: 10.1109/SCM.2017.7970622. pp. 477-480.

[5] Wang L., Wu P., Chen H. Finding probabilistic prevalent colocations in spatially uncertain data sets. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(4). pp. 790–804.

[6] Bernecker T., Kriegel H.-P., Renz M., Verhein F., Zufle A. Probabilistic frequent itemset mining in uncertain databases, *Proc. 15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, 2009. 9p.

[7] Tong Y., Chen L., Cheng Y., Yu P. S. Mining frequent itemsets over uncertain databases. *Proceedings of the 38th International Conference on Very Large Data Bases (VLDB)*, 2012, 12p.

[8] Young M. The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989. 232pp.

[9] Dunn J.C. A fuzzy relative of the ISODATA process and itsuse in detecting compact well-separated clusters, *J.Cybernetics*, 3(3), 1973. pp. 32-57.

[10] Bezdek J.C. *Pattern recognition with fuzzy objective function algorithms*, Plenum New York, 1981. 272 p.

[11] Benmouiza Kh., Cheknane A. Density-based spatial clustering of application with noise algorithm for the classification of solar radiation time series., IEEE, Algeria, 2016.

[12] Russ G., Kruse R. Machine Learning Methods for Spatial Clustering on Precision Agriculture Data, *Eleventh Scandinavian Conference on Artificial Intelligence*, IOS Press, 2011. pp. 40-49.

[13] Lin W.-Y., Hung C.-T. Applying spatial clustering analysis to a township-level social vulnerability assessment in Taiwan, *Geomatics, Natural Hazards and Risk*, v. 7, i. 5, 2015. pp. 1659-1676.

[14] S. You, J. Zhang, and L. Gruenwald, , "Large-Scale Spatial Join Query Processing in Cloud"

[15] J. Yu, J. Wu, and M. Sarwat. 5, "Geospark: A cluster computing framework for processing large-scale spatial data.," In SIGSPATIAL GIS, 2015.

[16] "Magellan," [Online]. Available: https://github.com/harsha2010/magellan.

[17] Tang, M., Yu, Y., Malluhi, Q.M., Ouzzani, M., Aref, W.G. LocationSpark: a distributed in-memory data management system for big spatial data. PVLDB 9(13), 1565–1568 (2016)

[18] Xie, X., Mei, B., Chen, J., Du, X., Jensen, C.S.: Elite: an elastic infrastructure for big spatiotemporal trajectories. VLDB J. 25(4), 473–493 (2016).

[19] Zhigang Zhang, Cheqing Jin, Jiali Mao, Xiaolin Yang, Aoying Zhou, TrajSpark: A Scalable and Efficient in-memory Management System for Big Trajectory Data, *APWeb-WAIM Joint Conference on Web and Big Data*, 2017.