DOI 10.37882/2223–2966.2023.03.32

# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ОБРАБОТКИ СЛОЖНЫХ ВЗАИМОСВЯЗАННЫХ СОБЫТИЙ

## INFORMATION TECHNOLOGIES FOR PROCESSING COMPLEX INTERRELATED EVENTS

***I. Rizaev***
***R. Faskhutdinov***
***E. Takhavova***
***Z. Zakharova***

*Summary.* The article considers the problem of computations when processing complex interrelated events. Such problems arise when solving the traveling salesman problem with a large number of nodes, clustering retail buyers, analysis of the consumer basket using associative rules method, etc. In some cases, approximate calculation methods help to solve the problem, but they are insufficient to get accurate results within an acceptable time in cases with a large number of parameters. Parallel computing technology based on cluster models is an approach to solve such problems.

*Keywords:* interrelated events, parallel computing, cluster systems.

***Ризаев Ильдус Султанович***
*К.т.н., доцент, Казанский научный исследовательский технический университет (КНИТУ-КАИ)*
*isr4110@mail.ru*
***Фасхутдинов Руслан Минсеферович***
*Аспирант, Казанский научный исследовательский технический университет (КНИТУ-КАИ)*
*sijeyrus@mail.ru*
***Тахавова Эльза Габдулбаровна***
*К.э.н., доцент, Казанский научный исследовательский технический университет (КНИТУ-КАИ)*
*elzzy@yandex.ru*
***Захарова Земфира Хаматовна***
*Старший преподаватель, Казанский научный исследовательский технический университет (КНИТУ-КАИ)*
*zkhzakharova@mail.ru*

*Аннотация.* В статье рассматривается проблема вычислений при обработки сложных взаимосвязанных событий. Такие проблемы возникают при решении задача коммивояжера с большим числом узлов, кластеризации покупателей розничной торговли, анализа рыночной корзины методом ассоциативных правил и др. Показано, что в ряде случаев могут быть использованы приближенные методы расчетов. Но для получения точных результатов при большом числе параметров и за приемлемое время требуются новые подходы. Для решения таких проблем предлагается использовать технологию параллельных вычислений на основе кластерных систем.

*Ключевые слова:* взаимосвязанные события, параллельные вычисления, кластерные системы.

## Введение

В настоящее время сложность технических задач возрастает постоянно, это связано с проектированием и обсчетом самых различных устройств: ракет самого различного назначения, самолетов, интеллектуальных робототехнических систем, больших нефтехимических предприятий, полеты космических объектов и т.д. Например, для расчета летательного аппарата в условиях нестационарного полета может понадобиться вычислительная система с производительностью $10^n$, где n>10.

С ростом и развитием информационно-коммуникационных технологий, повсеместной цифровизации народного хозяйства, появлением «умных» отраслей, предприятий, городов резко выросли объемы обрабатываемой информации [1,2]. Из-за постоянного роста объемов информации, базы данных стали настолько большими, что большая часть этой информации остается невостребованной человеком. Для обнаружения и извлечения новых знаний используются методы Data Mining: классификация, кластеризация, ассоциативные правила и др. В природе довольно часто встречаются взаимосвязанные события: маркетинг — анализ рыноч-

Таблица 1. Сложность вычислений

| Количество городов | Количество возможных решений |
|---|---|
| 4 | 24 |
| 5 | 120 |
| 6 | 720 |
| 7 | 5 040 |
| 8 | 40 320 |
| 9 | 362 880 |
| 10 | 3 628 800 |
| 11 | 39 916 800 |
| 12 | 479 001 600 |
| 13 | 6 227 020 800 |
| 14 | 87 178 291 200 |
| 15 | 1 307 674 368 00 |
| 16 | 20 922 789 888 000 |
| 17 | 355 687 428 096 000 |
| 18 | 6 402 373 705 728 000 |
| 19 | 121 645 100 408 832 000 |
| 20 | 2 432 902 008 176 640 000 |
| 25 | 15 511 210 043 330 985 984 000 000 |

| Метод решения задачи коммивояжера | Сложность метода |
|---|---|
| Полный перебор | $O(n!)$ |
| Метод ветвей и границ | $O(n * log_2 n)$ |
| Алгоритм ближайшего соседа | $O(n)$ |
| Алгоритм Метрополиса (имитация отжига) | $O(n^2 * log_2 n)$ |
| Генетический алгоритм | $O(t * m * n^2)$ |

Рис. 1. Методы решения задачи коммивояжера

ной корзины, медицина, страхование, военные операции и т.д. В этом случае для расчета таких систем в приемлемое время понадобятся вычислительные системы очень высокой производительности. Рост производительности вычислительных систем, конечно, всеми приветствуются, но вместе с этим, появляются новые задачи, требующие еще большей производительности и этот замкнутый процесс бесконечен. Решением таких проблем является использование суперкомпьютеров или технологии параллельных вычислений на основе кластерных систем.

## Моделирование решения сложных задач

**Задача коммивояжера**

Задача коммивояжера (Traveling salesman problem — задача бродячего торговца) заключается в том, что при условии, что торговец должен обойти ряд городов, побывав в каждом городе один раз и вернуться в исходный город, необходимо найти путь минимальной длины [3,4]. Коммивояжер, побывав в городе X с вероятностью p (X) далее выбирает город Y с вероятностью p (Y). Математическая модель задачи имеет следующий вид:

$$F(x) = \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} X_{ij} \rightarrow min \quad (1)$$

$$\sum_{i=1}^{N} X_{ij} = 1 \text{ i} = 1..\text{N} \quad (2)$$

$$\sum_{j=1}^{N} X_{ij} = 1 \text{ J} = 1..\text{N} \quad (3)$$

$$X_{ij} \geq 0$$

Существуют различные методы решения задачи: метод перебора, ветвей и границ, отжига, генетический, муравьиный алгоритм. Метод перебора дает оптимальный результат, но сложность его составляет 0 (n!). В таблице 1 показано, как быстро растет сложность решений от количества узлов.

В реальной ситуации при движении коммивояжера по крупному агломерату с населением в несколько миллионов, где нужно учесть сотни узлов, задача становится очень сложной. Для решения можно использовать приближенные методы, при этом сложность решения различна.

На рис.1.приведен перечень методов решения задач коммивояжера и показана сложность их решения

Видно, что даже при небольшом числе узлов точные методы требуют значительных временных затрат. Так для расчета поиска расстояния методом перебора время поиска будет в 60тыс раз больше, чем поиск «жадным» алгоритмом. Если понадобиться расчет маршрута в городской среде с населением в несколько миллионов с числом дорожных узлов в 100 и более значений, то потребуются значительные временные затраты и соответствующие вычислительные ресурсы. Эффективным решением данной проблемы может быть использование технологии параллельных вычислений.

**Кластерный анализ**

Методы кластерного анализа применимы в задачах, когда необходимо провести анализ больших объемов информации в различных областях: анализ предпочтений клиентов в сети сотовой связи, интернете; банкинг-анализ профиля клиентов, подающих заявки на потребительские кредиты; выявление симптомов, препаратов и методов лечения при анализе заболеваний в медицине; выявление предпочтений клиентов и их разбивка на сегменты в розничной торговле [4,5] .

Кластеризация заключается в разбиении множества объектов на однородные группы, имеющих общие свойства. Объекты объединяются в кластеры с учетом расстояния, при этом процесс объединения является итерационным. Трудность процессов кластеризации заключается в том, что все множество объектов может быть разбито на совершенно различные кластеры. Это связано с тем, что данные, описывающие свойства объектов могут иметь самые различные типы, обычно представляемые как числовые, так и категориальные [6] .

При выполнение кластерного анализа могут быть использованы методы как иерархические, так и неиерархические. К последней группе можно отнести:

- Метод k - средних;
- Метод G - средних;
- Сети Кохонена;
- Адаптивные методы.

Большой популярностью при проведении кластерного анализа является использование метода k — средних. При этом методе интуитивно устанавливается число возможных кластеров и соответственно произвольно назначаются объекты как центры кластеров. Затем с помощью Евклидово расстояния определяются расстояния от указанного центра до каждого объекта. Определяется круг объектов, находящихся ближе всего к заданному центру. Для найденного кластера определяется новый центр. Далее пересчитываются снова расстояния до центов кластеров и так повторяется, пока не окажется, что центы больше не смещаются. Ниже приведена формула Евклидова расстояния:

$$\rho(X,Y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (4)$$

Для остановки поиска кластеров используется мера сходимости центров тяжести по формулу (5):

$$E = \sum_{i=1}^{k}\sum_{p\in S_i}(p - m_i)^2 \quad (5),$$

где $P$ — произвольный объект,
принадлежащая $i$-му кластеру;
$m_i$ — центр тяжести данного кластера.

Например, в больших супермаркетах при розничной торговле проводят ежедневно множество клиентов, совершающих самые различные покупки товаров. Наверное, достаточно важной задачей является сбор сведений о клиентах, совершающих покупки и их предпочтениях. Например, достаточно важной информацией являются сведения о возрасте клиентов, времени посещения супермаркета, размерах покупок, видов покупаемых товаров. Виды и размеры покупок, совершаемых по половы и возрастным признакам.

Для решения задачи анализа предпочтений покупателей кластеризации могут быть использованы различные статистические методы. Удобным средством для проведения анализа является интегрированная среда Deductor Studio. Данная аналитическая платформа имеет средства для проведения кластерного анализа и визуализации результатов [11–13] .

Например, в таблице 2 представлен кластерный анализ по цене приобретения товаров. Ценовая политика товаров, является важным показателем эффективности торговли. Можно «задрать» цены и потерять клиентов,

Таблица 2. Результат анализа

| Цена товара | Номер кластера | Расстояние до центра кластера |
|---|---|---|
| 1500 | 1 | 511,53846153846 |
| 450 | 1 | 1561,53846153846 |
| 1000 | 1 | 1011,53846153846 |
| 6500 | 2 | 1044,44444444444 |
| 15000 | 0 | 3062,5 |
| 10000 | 0 | 1937,5 |
| 100 | 1 | 1911,53846153846 |
| 12000 | 0 | 62,5 |
| 1200 | 1 | 811,53846153846 |
| 13000 | 0 | 1062,5 |
| 6900 | 2 | 1444,44444444444 |
| 4500 | 2 | 955,55555555556 |
| 3500 | 1 | 1488,46153846154 |
| 12000 | 0 | 62,5 |
| 1400 | 1 | 611,53846153846 |
| 4400 | 2 | 1055,55555555556 |
| 3200 | 1 | 1188,46153846154 |
| 4500 | 2 | 955,55555555556 |
| 3200 | 1 | 1188,46153846154 |
| 12000 | 0 | 62,5 |
| 6000 | 2 | 544,44444444444 |
| 3500 | 1 | 1488,46153846154 |
| 2700 | 1 | 688,46153846154 |
| 7800 | 2 | 2344,44444444444 |
| 2900 | 1 | 888,46153846154 |
| 11500 | 0 | 437,5 |
| 4500 | 2 | 955,55555555556 |
| 10000 | 0 | 1937,5 |
| 4000 | 2 | 1455,55555555556 |
| 1500 | 1 | 511,53846153846 |

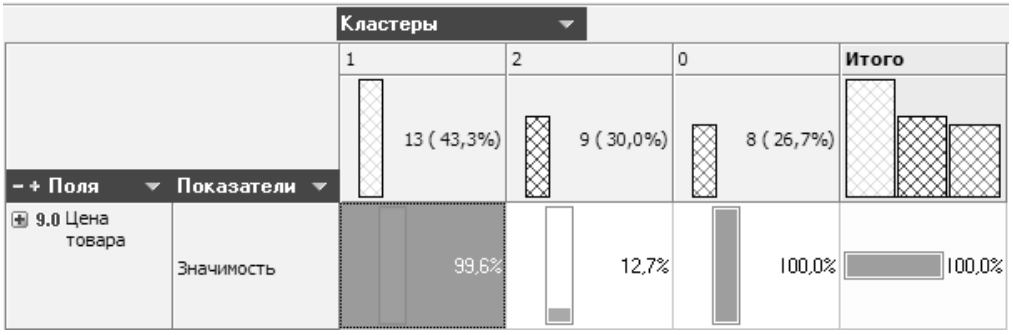


Рис. 2. Профили кластеров

низкие цены дадут малую прибыль и проигрыш конкурентам. При проведении кластеризации были указаны три кластера в соответствие с ценами: низкая, средняя и высокая цена.

Результаты анализа (таблица 2) представлены тремя кластерами. В нулевой 0й кластер вошли клиенты с высокой ценовой направленностью, в первый 1й — низкой категории, а 2й — средней категории.

Статистические данные отображены в виде профиля кластеров (рис.2), где наглядно представлено, что большее число клиентов входят в низкую ценовую категорию (43,3%). Средняя категория составила 30%, категория с высокой ценовой политикой составила 26,7%.

Для наглядности и упрощения анализа система Deductor позволяет профили кластеров представить в графическом виде. Так на рис.3 результаты анализа

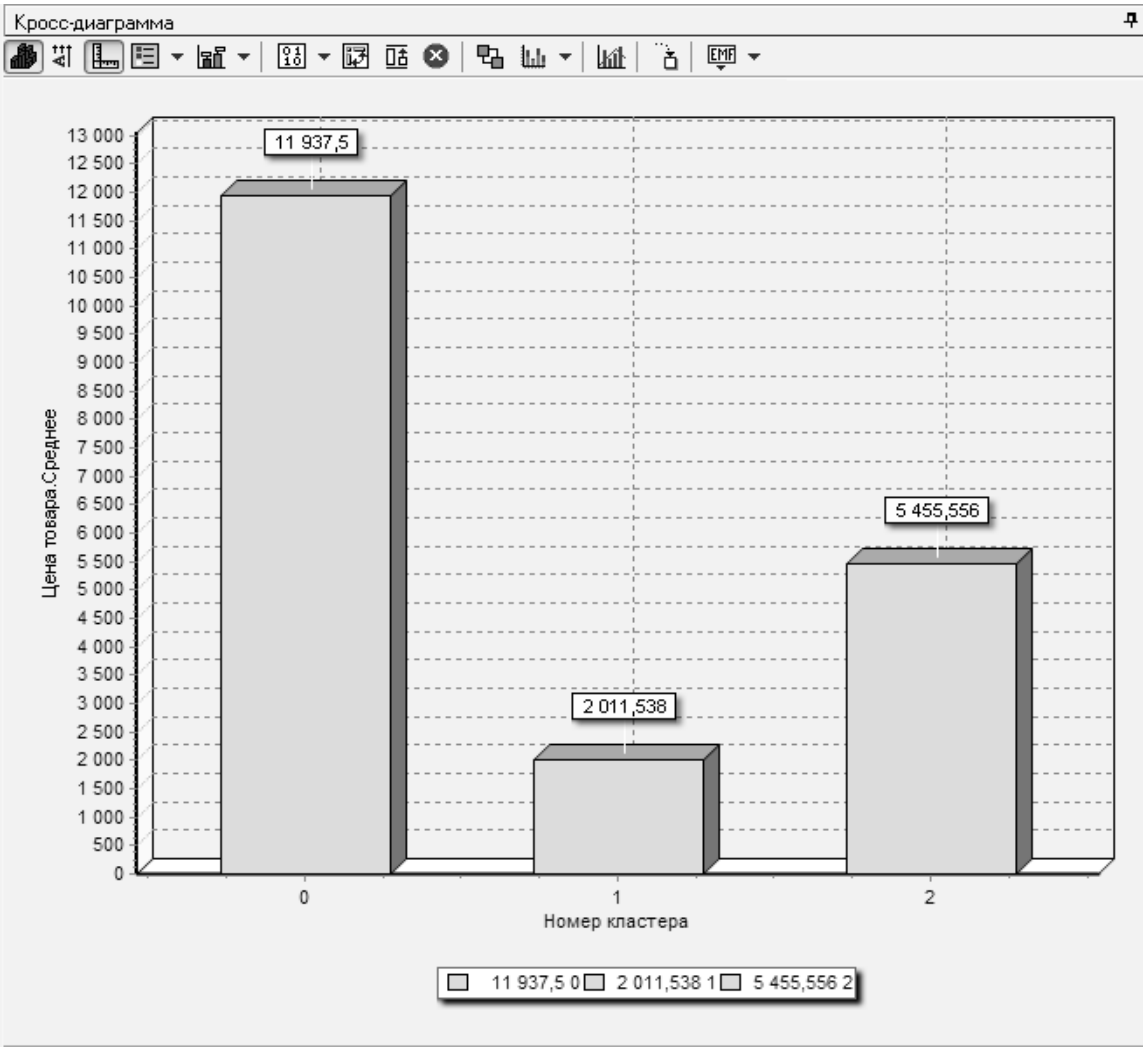


Рис. 3. Кросс-диаграмма

представлены в виде кросс-диаграммы, где по высоте диаграмм можно наглядно увидеть средние значения цен для каждого кластера.

Такой подход можно использовать для решения самых различных задач, возникающих в деятельности супермаркетов. Это и решение задач товарооборота, и посещения клиентов и их предпочтения. Кластерный анализ может позволить вести деятельность товарооборота более эффективным образом.

Но проблема кластеризации заключается в том, что в случае больших массивов информации, имеющих миллионы строк, требуется значительные вычислительные затраты.

Решением этих проблем может быть использование суперкомпьютеров или применения технологии параллельных вычислений.

**Поиск ассоциативных правил**

Задачей ассоциативных правил, является поиск типичных шаблонов, например, покупка совместных товаров, что нашло определение как анализ рыночной корзины. То есть, если покупатель купил молоко, то он с вероятностью 75% купит сметану и творог. Задачей ассоциативных правил является поиск закономерностей между связанными событиями [7,10]. Поиск связанных событий может быть использован при решении достаточно многих задач: в медицине — применение лекарственных средств, в геологии — поиск сопутствующих минералов, полиции — писк преступников по определенным категориям и т.д. Но наибольшее применение находит в розничной торговле для учета предпочтений покупателей. Если учесть, что количество разнообразных товаров в супермаркетах составляет десятки тысяч наименований, то весьма важным является иметь оценки предпочтений клиентов и наборы совместных

покупок. Такие наборы совместных покупок называют транзакциями. Например, транзакциями являются наборы Т 1 (хлеб, сыр, молоко, масло), Т 2 (помидоры, огурцы), Т 3 (яблоки, апельсины, виноград) и т.д. Наборы транзакций составляют базу данных: БД (Т 1, Т 2, …, $Т_к$).Аналитики должны учитывать ассоциативные связи при выполнении покупок. Например, если покупатель приобрел чай, то скорее всего он корзину дополнит конфетами и печеньем. Здесь возникает ассоциативное правило: «если условие, то следствие» из А следует В. То есть, если клиент взял огурцы, то скорее всего возьмет и помидоры. Если взял бутылку пива, то возьмет скорее всего и чипсы. Такую последовательность покупок могут учитывать ассоциативные правила. Таких правил может быть великое множество. Например, если имеем k =5 — товаров и хотим рассмотреть сочетания только двух предметных (бинарных) наборов, то получим 80 сочетаний

$$k \cdot 2^{k-1}$$

При k=10 N=5120.

Видим, что число ассоциаций растет экспоненциально.

В общем случае, при N элементов товаров число транзакций составит:

$$R = \sum_{d=1}^{N} \binom{N}{d} * \sum_{j=1}^{N-d} \binom{N-d}{j}$$

При рассмотрении множества транзакций, надо иметь ввиду, что одни транзакции могут быть частыми, а другие практически пустыми (не во требуемыми). Поэтому важным является учесть полезность ассоциативных правил. Для оценки полезности правил используются понятия: полезности (support) и достоверности (confidens) [10]:

$$Supp(P) = \frac{\left|D_j\right|}{\left|D\right|}$$

$$Conf(A \rightarrow B) = \frac{\left|D_{P(A \cap B)}\right|}{\left|D_A\right|}$$

Задав определенные порог на значения поддержки и достоверности можно исключить какие-то транзакции, не являющиеся правилами. Не смотря на введенные ограничения все равно приходится рассматривать десятки и сотни всевозможных ассоциаций, что является довольно затруднительным. Для уменьшения числа ассоциаций можно использовать алгоритм Apriori, суть которого состоит в рассмотрении частого набора, это наборы, со значением превышающим заданный порог [10]. Алгоритм Apriori использует принцип антимонотонности, суть которого заключается в том, что если предметный набор не является частым, то добавление нового предмета не делает этот набор частым. Таким образом можно уменьшить пространство поиска. Но несмотря на это число возможных транзакций может быть очень большим, что не позволит провести исследование вручную.

Здесь также эффективным решением данной проблемы может быть использование суперкомпьютеров или применение технологии параллельных вычислений.

**Технология параллельных вычислений**

Для решения достаточно сложных вычислительных задач, требующих значительных временных затрат, могут быть использованы суперкомпьютеры, являющиеся дорогими многопроцессорными системами. Альтернативой суперкомпьютерам явились более дешевые кластерные системы, позволяющие выполнять как простые, так и сверхсложные задачи за счет параллельной обработки данных. Суть параллельной обработки заключается в том, что если 1000 операций, выполняемых одним устройством за 1000 единиц времени, распараллелить на пять устройств, то общее время обработки составит 1000/5=200 единиц.

Для решения таких задач выступают процессоры, объединенных коммуникационной сетью, которые и обеспечивают обмен данными. При этом процессоры работают независимо друг от друга.

Такая схема с процессорами, объединенных с помощью коммуникационной сети и представляет современны кластер. Для организации обмена данными между процессорами и организации соответствующей нагрузки используется интерфейс MPI (Message Passing Interface). На рисунке 4 представлена модель вычислений кластером на основе MPI программы [8,9] .

Например, при поиске оптимального маршрута в задаче коммивояжера, центральный процессор организует перебор вариантов и рассылает их на различные узлы для расчетов. На рисунке 5 представлена схема маршрутов, число вариантов (n-1)!. В данном случае (4–1)!=6

Имеем следующие варианты:

ABCDA 5+8+6+7=26

ABDCA 5+4+6+3=18

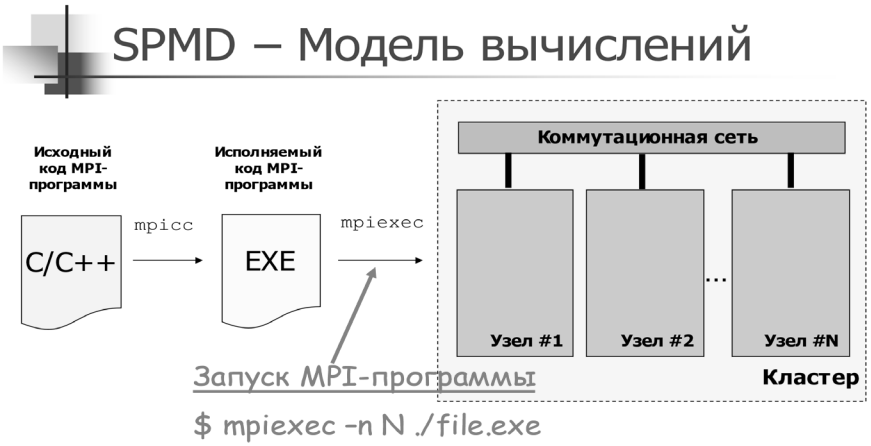


Рис. 4. Модель вычислительного кластера

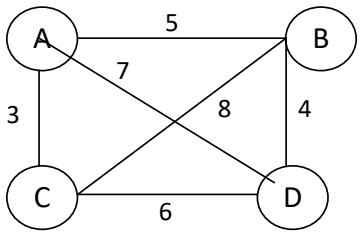


Рис. 5. Схема маршрутов

ACBDA 3+8+4+7=22

ACDBA 3+6+4+5=18

ADCBA 7+6+8+5=26

ADBCA 7+4+8+3=22

При большом числе вариантов необходимо использовать конвейерный метод распараллеливания.

Для решения задач ассоциативного поиска можно использовать аналогичный подход. Для уменьшения числа переборов проверять транзакции на поддержку и достоверность.

При кластерном анализе, получив модули кластеров, можно вести параллельную обработку модулей на различных узлах вычислительной системы.

## Заключение

Повсеместное использование вычислительной техники, информационно-коммуникационных систем во всех сферах народного хозяйства привело к взрывному росту объемов информации. Современный 21 век характеризуется технологическим ростом наукоемких предприятий, требующих для расчетов значительных вычислительных ресурсов высокой производительности. Такие требования возникают из-за повсеместной цифровизации промышленности, созданием умных городов, появлением интеллектуальных робототехнических систем. сложных высокотехнологичных изделий: в медицине, в исследовании космоса, в военной сфере и многих других. Во многих ситуациях возникает необходимость для обработки информации мощные производительные вычислительные системы. Для решения задач, требующих длительных затратных ресурсов, могут быть использованы суперкомпьютеры, или распределенные кластерные системы с использованием интерфейса передачи данных MPI.

В работе показано, что такие задачи как расчет длительности оптимального маршрута с большим числом узлов в задачи коммивояжера может быть использована технология параллельных вычислений.

Технология параллельных вычислений может быть использована при решении задач кластерного анализа при большом числе значений информации. В этом случае разбив множество исходных данных на кластеры

(модули) можно распараллелить их для решения на узлах сети.

Приведен пример поиска ассоциативных правил для установления закономерностей между связанными событиями. Задачей ассоциативных правил, является поиск типичных шаблонов, например, покупка совместных товаров, что нашло определение как анализ рыночной корзины. При большом числе товаров количество возможных правил может быть очень большим. Опять, таки, для решения таких задач можно использовать технологию параллельных вычислений.

ЛИТЕРАТУРА


1. Шалагина, Г.Э. Информационно-коммуникационные технологии как предмет социогуманитарных исследований // Г.Э. Шалагина, С.В. Шалагин — Вестник Московского государственного областного университета. Серия: Философские науки. — 2019. — № 2. — С. 154–163. — DOI 10.18384/2310–7227–2019–2–154–163.
2. Шалагин, С.В. Когнитивные проблемы проектирования на основе компьютерных моделей: технический и социо-гуманитарный аспекты // С.В. Шалагин, Г.Э. Шалагина — Онтология проектирования. — 2016. — Т. 6. — № 3 (21) . — С. 368–376. — DOI 10.18287/2223–9537–2016–6–3–368–376.
3. Ерзин А.И. Задачи маршрутизации: учеб. пособие // А.И. Ерзин, Ю.А. Кочетов. — Новосиб. гос ун-т. — Новосибирск: РИЦ НГУ, 2014. — 95 с.
4. Семенов С.С. Анализ трудоемкости различных алгоритмических подходов для решения задачи коммивояжера // С.С. Семенов, А.В. Педан, В.С. Воловиков, И.С. Климов — Системы управления, связи и безопасности. 2017. № 1. С. 116–131.
5. Ризаев И.С. Группировка объектов на основе кластерного анализа.// И.С. Ризаев, К.В. Потапова, Е.В. Январева — Инновационные технологии XXI века материалы Международной научно-прктической конференции. — Казанский государственный технический университет им. А.Н. Туполева. 2015.С.160–162.
6. Ризаев И.С. Кластеризация покупателей розничной торговли на платформе Deductor. // Ризаев И.С., Кирпичников А.П., Сафаров Н.И., Анишкина В.Н. — Вестник технологического университета. 2019. Том 22, № 7, с. 158–161
7. Ризаев И.С. Поиск закономерностей между взаимосвязанными событиями на основе ассоциативных правил. // И.С. Ризаев, Л.М. Шарнин, З.Т. Яхина — Вестник КГТУ им. А.Н. Туполева,№ 4, 2014, с.314–317
8. Параллельное программирование на основе MPI [Электронный ресурс] // ННГУ. Центр суперкомпьютерных технологий: [сайт]. [2004]. URL: http://www.hpcc.unn.ru/mskurs/RUS/DOC/ppr04.pdf (дата обращения: 04.05.2022) .
9. Райхлин В.А. Эффективность консервативных СУБД больших объемов на кластерной платформе // В.А. Райхлин, Р.Ш. Минязев, Р.К. Классен — Кибернетика и программирование, 2018, № 5, с.44–62
10. Barsegyan A.A. Data analysis technology: Data Mining, Visual Mining, Text Mining, OLAP. — St. Petersburg: BHV-Petersburg, 2008. — 384p.
11. Cherezov D.S. N.A. Tyukachev "Overview of the main methods of classification and clustering of data // D.S. Cherezov, N.A. Tyukachev Voronezh. Bulletin of Voronezh State University, Series: System Ana Liz and Information Technology, 2009,
12. Chubukova I.A. Data Mining: a tutorial // I.A. Chubukova — Moscow: Internet-University of Information Technologies: BINOM: Laboratory of Knowledge, 2008. — 382 p.
13. Eibe F. Data mining. Practical machine learning tools and techniques. F. Eibe, I. Witter. — 2005. — 525p.


© Ризаев Ильдус Султанович ( isr4110@mail.ru ), Фасхутдинов Руслан Минсеферович ( sijeyrus@mail.ru ),
Тахавова Эльза Габдулбаровна ( elzzy@yandex.ru ), Захарова Земфира Хаматовна ( zkhzakharova@mail.ru ).
Журнал «Современная наука: актуальные проблемы теории и практики»