DOI 10.31891/2307-5732-2024-333-2-1
УДК 004.773.2

ШАХОВСЬКА НАТАЛІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-6875-8534
e-mail: Nataliya.b.shakhovska@lpnu.ua
ЗАГОРОДНИЙ ІВАН
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-6875-8534
e-mail: ivan.i.zahorodnyi@lpnu.ua

МЕТОД ПОШУКУ ЗАЛЕЖНОСТЕЙ У ВЕЛИКИХ НАБОРАХ ДАНИХ ІЗ ПРОПУСКАМИ НА ПРИКЛАДІ ДАНИХ ПРО ПОШИРЕННЯ COVID

В роботі наведено результати розроблення методу пошуку прихованих залежностей у великих наборах даних з пропусками на основі розробленого ансамблю кластеризації, регресійного дерева та асоціатвних правил.
Ключові слова: видобування залежностей даних, великі набори даних, асоціативне правило.

SHAKHOVSKA NATALIYA, ZAGORODNIY IVAN
Lviv Polytechnic National University

METHOD OF DEPENDENCES MINING IN BIG DATA SETS WITH OMISSIONS AS AN EXAMPLE OF DATA ON THE SPREAD OF COVID

In the article, a method of big data analysis with omissions is developed on the example of the construction of advisory rules regarding adequate state policy to reduce the spread of new Covid-19 cases. Association rules and rule generation in Big Data have a number of challenges, the main one being the presence of large numbers of vectors and multivalued datasets. That is why in the paper the system of these rules is based on a new ensemble of machine learning techniques such as associative rules, regression tree and clustering. This study used pooled data from the Government's COVID-19 Response Tracker and ECDC's Covid-19. Clustering was performed using the k-means method. Gap statistics allow finding an appropriate number of clusters, and in the case study three clusters were selected. The clusters differ in the recommendations and actions of the correspondent governments. Thus, the first countries of the cluster chose to close schools and control international travel as the main recommendations; countries in the second cluster recommended staying at home, while the main recommendations from the governments of countries in the third cluster were to stay at home and cancel public events. The same country can be attached to different clusters at different time intervals. Therefore, the clustering by countries will not be so unambiguous. That is why the time series for the detached country can be interesting and will be the subject of further study. In addition, the impact of such clustering on the spread of COVID-19, the position and duration of the peak, and the mortality rate will also be the subject of our further investigation. A regression decision-making tree was built, a set of rules was obtained from the decision tree and applied rules for generating associative dependencies. The resulting dependencies can be used for strategic planning in the healthcare system.
Keywords: extraction of data dependencies, large data sets, associative rule.

Постановка проблеми

Інформаційна технологія Велеких даних - це сукупність методів і засобів обробки різних типів структурованих і неструктурованих динамічних великих обсягів даних з метою їх аналізу та використання для підтримки прийняття рішень. Відсутні значення викликають серйозні проблеми при аналізі даних. Окрім того аналіз біл е ніж одного джерела даних вимагають попредньої інформації про їх вміст. Одним із способів попереднього аналізу даних є асоціативні правила. Але, зважаючи на відчутну обчислювальну складність різних алгоритмів вибобування асоціативних правил, доцільно використовувати ансамблювання різних моделей машинного навчання.

Аналіз останніх джерел

В [1–3] визначено, що для представлення великих даних використовуються багатовимірні та об'єктні моделі. Багатовимірне представлення даних добре використовується для задач візуалізації та аналізу даних, але через розсіювання гіперкуба обсяг даних в цьому випадку є більшим, ніж при реляційному представленні, що є неприйнятним для великих даних. Запропонований метод дозволяє зберігати об'єкти у вигляді атрибутів, їх характеристик та зв'язків між характеристиками. При певній модифікації він може бути використаний і для великих даних. Однак залишається невирішеною проблема трансформації різних типів даних (текстових, напівструктурованих) в об'єктну модель.

Правила асоціацій та генерація правил у Великих даних мають ряд проблем, головною з яких є наявність великої кількості векторів та багатозначних наборів даних [4]. Найпопулярнішими алгоритмами для генерації асоціативних правил є: Apriori, Apriori TID, Hybrid Apriori, FP-tree [5, 6]. Для невеликих наборів даних продуктивність Apriori вища, ніж Apriori TID, але Hybrid Apriori можна використовувати і для великих наборів даних. FP-дерево дозволяє знайти іншу залежність, ніж Apriori, оскільки створює зв'язки між рівнями в дереві. Однак ці алгоритми дуже важко використовувати для аналізу великих даних. Наприклад, часова складність Apriori становить O(d2 n), часова складність FP-дерева - O(dn), де n - кількість записів у наборі даних, a d - кількість частих наборів елементів [6]. Однопроцесорні системи зі звичайною швидкістю процесора не можуть обробляти такий обсяг даних, що робить алгоритм неефективним для використання. Останні розробки мережевих технологій і особливо хмарних платформ надали нові ідеї для генерації багатоатрибутивних правил з використанням паралельного середовища, такого як Hadoop [5].

MapReduce став популярним і більш використовуваним для обчислення великої кількості даних з тих пір, як Google запустив його на своїй платформі. Розподілена файлова система Google (GFS) та Amazon Web Services (AWS) використовують платформу Hadoop та MapReduce для надання своїх послуг [5].

Алгоритм Apriori має великі проблеми з великими обсягами великих даних, оскільки він сканує всю базу даних кілька разів [6]. Це означає, що час виконання збільшується відповідно до кількості транзакцій. У нашому дослідженні ми використовували Spark та ієрархічний метод формулювання правил для покращення алгоритму Apriori.

Метою статті є розробка алгоритму пошуку прихлваних асоціативних правил для обробки великих даних. Для зменшення обчислювальної складності пропонується двоетапний метод, де спочатку здійснюється кластеризація, а потім для створених кластерів будуються асоціативні щалежності.

Виклад основного матеріалу

Опис наборів даних

У цьому дослідженні ми використовували дані Оксфордського урядового відстежувача COVID-19, а також Європейського центру з профілактики та контролю захворювань (ECDC) Covid-19 [7]. Набір даних Oxford COVID-19 урядового відстеження (OxCGRT) надає інформацію про те, які уряди вживали заходів, які дії вони вживали та коли вони вживали. OxCGRT систематично збирає інформацію про низку спільних політичних відповідей певного уряду та визначає, наскільки уряд реалізує ці заходи. Відповідні бали об'єднані у набір показників політики. Другий набір даних містить нові публічні дані про географію випадків COVID-19 у всьому світі від Європейського центру з профілактики та контролю захворювань. Кожен рядок або запис містить кількість нових випадків на день за країною чи регіоном.

Обробка даних

Набори даних складається з атрибутів різного характеру. Також представлені дані про шум та викиди. Ось чому на першому етапі необхідні наступні кроки: вибір функцій, аналіз порожніх даних, нормалізація та масштабування даних.

Вибір ознак здійснюється на основі теорії інформації. Спільна взаємна інформація між кожною ознакою та цільовим атрибутом ConfirmedCases обчислюється як:

$$JI(f_i) = \sum_{f_j \in S}\left(I(f_i; c) + I(f_i; c|f_j)\right) \quad (1)$$

де c - цільовий клас, , *I(f,c)* - взаємна інформація, f_j - це вже обрана функція, f_i обробляється.

Вже вибрані ознаки були обрані вручну. Список уже вибраних змінних виглядає так: countryName, loaddate. В результаті було обрано 19 із 38 змінних.

Наступним кроком є аналіз порожніх даних. Набір даних складається з 66 998 рядків, 26 003 з яких мають порожні значення у вибраних атрибутах. Через невелику кількість рядків з порожніми даними ці рядки були ліквідовані.

Далі необхідно привести всі вхідні змінні до одного діапазону та нормалізувати (максимальне абсолютне значення вхідних змінних не повинно перевищувати одного). В іншому випадку помилки через змінні, що змінюються в широкому діапазоні, будуть більш впливовими, ніж помилки через змінні, що змінюються у вузькому діапазоні. Переконавшись, що кожна функція змінюється в одному і тому ж діапазоні, ми гарантуємо, що кожна з них має однаковий ефект. Тому вхідні змінні, як правило, масштабуються так, що змінні змінюються в діапазоні функції, як правило, [0,1] або [−1,1]. Використовується масштабування Softmax. Розподіл набору даних для кожної функції показано на рис.1.

Методи пошуку правил

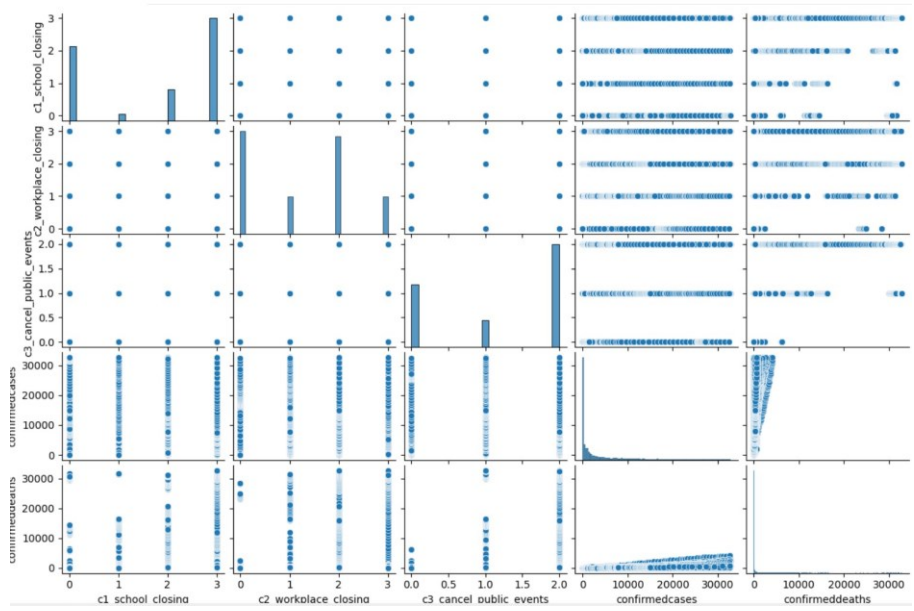
Спочатку розробляється регресійне дерево рішень. Дерева рішень ділять простір об’єктів відповідно до набору правил розділення. Ці правила є логічними твердженнями про змінну і можуть бути істинними або хибними. Тут є три обставини:

- правила дозволяють здійснювати послідовну дихотомічну сегментацію даних,
- два об'єкти вважаються подібними, якщо вони з'являються в одному сегменті розділу,
- на кожному кроці розділу кількість інформації про досліджувану змінну (відповідь) збільшується.

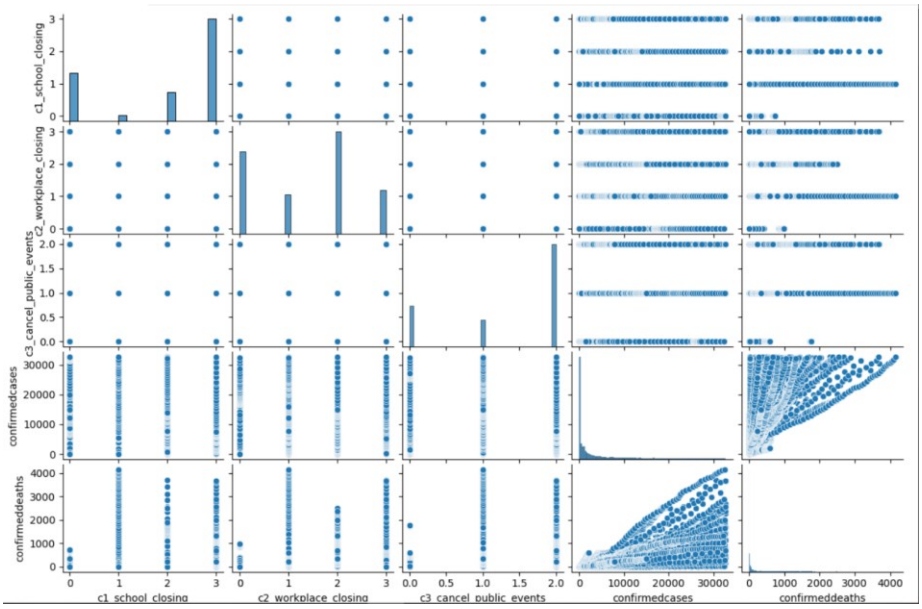
Головною особливістю запропонованого алгоритму є його k-дугова структура. Розгалуження за обраною ознакою x розділяє навчальні об'єкти на k підпроби, де k - кількість різних характеристичних значень.

Не втрачаючи загальності, будемо вважати, що ознака x має значення від {0, 1,…, k-1}, $k \geq 2$. У цьому випадку при побудові дерева рішень з вершини are є k дуги позначені цифрами з {0, 1,…, k-1}. Нехай σ - мітка однієї з дуг, що залишають вершину x, $\sigma \in$ {0, 1,…, k-1}. Щоб сформувати нову поточну підмножину об’єктів та новий поточний набір ознак, ті об’єкти з T видаляються, для яких значення ознаки x не дорівнює σ, а також сама функція x видаляється з набору ознак.

Нехай v - звисаюча вершина, породжена гілкою дерева з внутрішніми вершинами x_{j_1},…, x_{j_r} і нехай дуга, що виходить з вершини x_{j_i}, $i \in \{1,\ldots, r\}$, позначити як σi. Далі нехай $T(v)$ - поточний набір об’єктів, які потрапили у вершину v. Вершина v асоційована з парою $(B, w(v))$, де $w(v)$ дорівнює середньому значенню цільової змінної для всіх об’єктів з $T(v)$, а B - елементарний зв’язок форми $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$. Якщо вершина v не є підвісною, тоді ми призначаємо їй сполучнику $B = xx_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ ... Інтервал істини елементарного сполучення B позначаємо через NB. Нехай S - впізнаваний об’єкт. Для кожної висячої вершини $(B, w(v))$ виконується перевірка належності опису об’єкта тесту до інтервалу істини NB.



(*a*)



(*b*)

Рис. 1. Розподіл даних до (а) та після (б) попередньої обробки

Якщо опис S належить NB, то об'єкт S пов'язаний зі значенням цільової змінної w (v). Об'єкту S присвоюється значення цільової змінної.

$$W = \frac{\sum_{i=1}^{l} w(v_i) * I_{B_i}}{\sum_{i=1}^{l} I_{B_i}}, \quad \text{де} \quad I_{B_i} = \begin{cases} 1, if\ S \in N_{B_i}, \\ 0, \text{інакше}. \end{cases}$$

Далі на основі регресійних дерекв випадкового лісу побудовано асоціативні правила.

У випадку обчислення рівня підтримки для асоціативних правил умовний та результуючий предикат залежності об'єднуються знаком кон'юнкції:

$$Supp(S \to T) = Supp(S \wedge T) = \frac{|\sigma_{S \wedge T}(r)|}{|r|}. \tag{2}$$

З використанням цього поняття рівень довіри можна обчислити, як

$$Conf(S \to T) = \frac{Supp(S \to T)}{Supp(S)}. \tag{3}$$

Результати

Наступним кроком є кластеризація. Використовується метод k-середніх значень. Статистика ліктя дозволяє знайти відповідну кількість кластерів.

Кластерні центроїди дозволяють знаходити "середній" об'єкт у кожній групі та створювати правила регуляризації. Кластер №3 показує країни, на які поширюється обмеження, кластер №1 складається з країн без майже будь-яких обмежень (табл. 1).

Таблиця 1

Зважені характеристики для класів

Cluster	School closing	Workplace closing	Cancel public events	Close public transport
1	0.04368254	0.02120635	0.07822222	0.005587302
2	2.89865433	2.26230025	1.94790791	1.109703532
3	1.88113534	1.02194531	1.30058844	0.256905504

Cluster	Stay at home req.	Restrictions on internal movement	International travel controls
1	0.0351746	0.01853968	1.137778
2	1.7232443	1.68355761	3.438709
3	0.5112496	0.50931118	2.850190

Cluster	Income support	Public information campaigns	Testing policy	Contact tracing
1	0.02933333	0.7869206	0.6651429	0.6673016
2	1.00667578	1.9615749	1.5702271	1.3380467
3	1.17223953	1.9540325	1.8985808	1.6339218

Система рекомендацій на основі розподілу об'єктів у кожному кластері зображена на рис. 2.

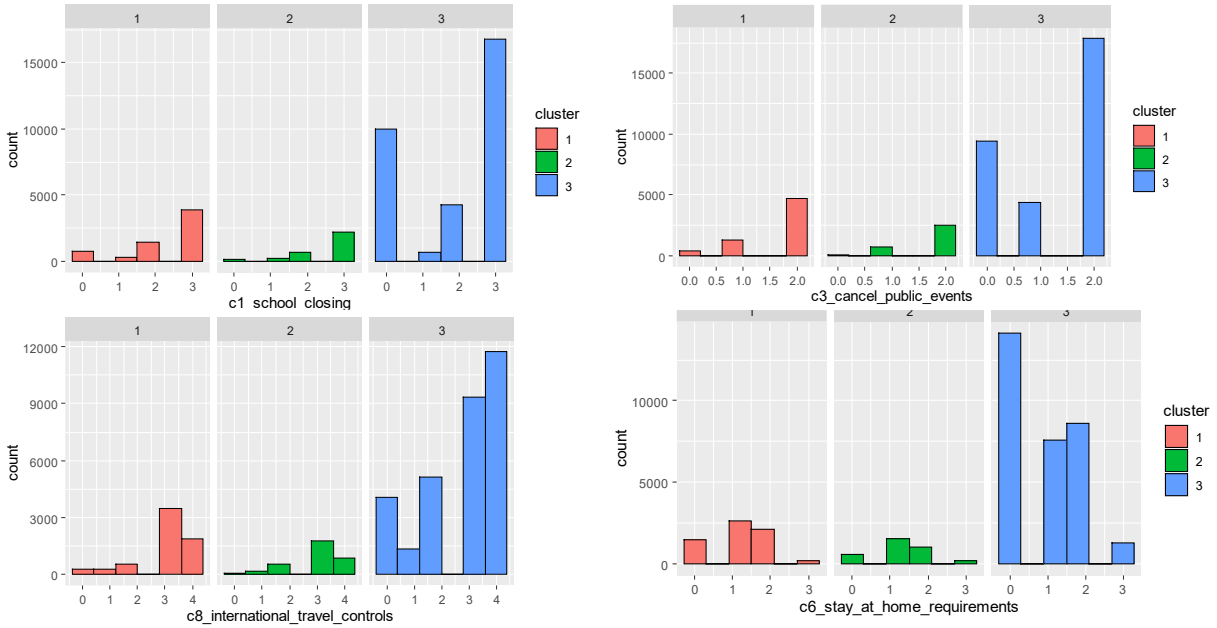


Рис. 2. Розподіл даних між кластерами за рекомендацією на основі розподілу об'єктів

Як видно з рис. 2, кластери відрізняються найбільш чіткими рекомендаціями, які можна узагальнити наступним чином:

• Кластер №1: рекомендується закривати школи та контролювати міжнародні поїздки,

• Кластер №2: рекомендується залишатися вдома,

• Кластер №3: рекомендується залишатися вдома та скасовувати публічні заходи.

Вплив такого скупчення на розповсюдження COVID-19, положення та тривалість піку, а також рівень смертності буде предметом нашого подальшого дослідження.

За результатами аналізу визначено, що одна і та ж країна може бути приєднана до різних кластерів у різні часові інтервали. Отже, кластеризація за країнами не буде настільки однозначною. Ось чому часові ряди для відокремленої країни можуть бути цікавими і будуть предметом подальшого вивчення. На цьому етапі ми використовували процедуру класифікації набору даних.

Проаналізовано три класифікатори: Random Forest (500 дерев), логістична регресія та XGBOOST (алгоритм навчання дерева). Бали моделей наведені в таблиці 2.

Таблиця 2

Оцінка різних класифікаторів

Random forest	Logistic regression	XGBOOST
0.96	0.91	0.89

Отримаі асоціативні правила, наведені нижче, представляють стратегію, засновану на дереві рішень та асоціативних правилах, і можуть бути використані для стратегічного планування в системі охорони здоров'я, щоб уникнути смертей та тяжких наслідків епідемії.

```
if ( confirmedcases <= 11577.5 ) {
   if ( confirmedcases <= 3806.5 ) {      [[confirmeddeaths=20.21931171]]    }
     else {     [[ confirmeddeaths =206.84760845]]   }
 } else {
   if ( confirmedcases <= 17153.5 ) {
     if ( e1_income_support <= 1.5 ) {      [[confirmeddeaths=370.77655172]] }
      else {
       if ( c8_international_travel_controls <= 2.5 ) {
         if ( c6_stay_at_home_requirements <= 0.5 ) {   [[confirmeddeaths=452.20754717]] }
          else {    [[confirmeddeaths=1410.8028169]]      }
       } else {    [[confirmeddeaths=501.40616622]]     }
     }
   } else {
     if ( e1_income_support <= 1.5 ) {
       if ( confirmedcases <= 24529.0 ) {  [confirmeddeaths= [503.03383459]] }
        else {         [[confirmeddeaths=780.60655738]] }
     } else {
       if ( c8_international_travel_controls <= 2.5 ) {
         if ( c5_close_public_transport <= 0.5 ) {       [[confirmeddeaths=898.76146789]] }
          else {     [[confirmeddeaths=1899.2962963]]          }
        else {
         if ( c6_stay_at_home_requirements <= 0.5 ) {
           if ( c7_flag <= 0.5 ) {
             if ( c7_restrictions_on_internal_movement <= 0.5 ) {
[[confirmeddeaths=1346.375]]  }
              else {             [[confirmeddeaths=413.04]]  }
           } else {              [[confirmeddeaths=3256.]] }
         } else {
           if ( confirmedcases <= 29128.0 ) {
             if ( c2_workplace_closing <= 2.5 ) {
              [[confirmeddeaths=665.88839286]]  }
              else {
               if ( c5_flag <= 0.5 ) {
                 if ( c1_school_closing <= 2.5 ) {                [[confirmeddeaths=2504.]] }
                  else {                 [[confirmeddeaths=815.2952381]]  }
               } else {          [[confirmeddeaths=1627.93548387]] }
             }
           } else {
             if ( h2_testing_policy <= 2.5 ) {
               if ( c8_international_travel_controls <= 3.5 ) {
                [[confirmeddeaths=1457.]] }
                else {           [[confirmeddeaths=530.2173913]]  }
             } else {              [[confirmeddeaths=708.86764706]]         }
           }
         }
       }
```

Висновки

У статті розроблено метод аналізу великих даних з пропусками на прикладі побудови рекомендаційних правил щодо адекватної державної політики щодо зменшення розповсюдження нових справ Covid-19. Система цих правил базується на новому ансамблі методів машинного навчання, таких як асоціативні правила, регресійне дерево та кластеризація. У цьому дослідженні використовувались об’єднані дані з урядового відстежувача відповідей COVID-19 та ECDC Covid-19.

Кластеризацію проводили методом k-середніх. Статистика розривів дозволяє знайти відповідну кількість кластерів, і у випадку дослідження було обрано три кластери. Кластери відрізняються рекомендаціями та діями урядів-кореспондентів. Таким чином, перші країни кластеру обрали закриття шкіл та контроль міжнародних поїздок як головні рекомендації; країни другого кластеру рекомендували залишатися вдома, тоді як головні рекомендації урядів країн, що входять до третього кластеру, полягали в тому, щоб залишатися вдома та скасовувати публічні заходи.

Одна і та ж країна може бути приєднана до різних кластерів у різні часові інтервали. Отже, кластеризація за країнами не буде настільки однозначною. Ось чому часові ряди для відокремленої країни можуть бути цікавими і будуть предметом подальшого вивчення. Крім того, вплив такого скупчення на поширення COVID-19, положення та тривалість піку, а також рівень смертності також буде предметом нашого подальшого дослідження.

Було побудовано регресійне дерево прийнятя рішень, набір правил було отримано з дерева рішень та застосованих правил генерації асоціативних залежностей. Отримані залежності може бути використано для стратегічного планування в системі охорони здоров’я.

References

1. Association mining. IEEE Transactions on Knowledge and Data Engineering, vol. 12, issue 3, pp. 372-390, 2000.

2. J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation. in ACM Sigmod Record, 2000, pp. 1-12.

3. J. Woo, "Apriori-Map/Reduce algorithm," in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), 2012, pp. 1.

4. X. Y. Yang, Z. Liu, Y. Fu, "MapReduce як модель програмування для алгоритму асоціативних правил на Hadoop", в Proceedings of IEEE 3rd International Conference on Information Sciences and Interaction Sciences (ICIS'2010), 2010, pp. 99-102.

5. R. Agrawal, T. Imieliński, A. Swami, "Mining association rules between sets of items in large databases," in ACM Sigmod Record, 1993, pp. 207-216.

6. M. Delgado, M. D. Ruiz, & D. Sánchez, "New approaches for discovering exception and anomalous rules," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 19, issue 2, pp. 361-399, 2011..

7. COVID-19 Data Lake. URL: https://azure.microsoft.com/en-US/services/open-datasets/catalog/covid-19-data-lake/