УДК 004.9

О.Ю. ПШЕНИЧНИЙ

ВИЯВЛЕННЯ ЗАЛЕЖНОСТЕЙ ДАНИХ

Анотація. *У статті описується розроблений алгоритм пошуку залежностей у підмножинах об'єктів досліджуваного набору даних. Даний алгоритм дозволяє ефективно виявляти асоціативні залежності за заданими критеріями якості. Застосування розробленого методу виявлення залежностей в даних можливе в багатьох предметних галузях та дозволяє виявити нові закономірності в даних, що покращує роботу фахівців та якість прийнятих ними рішень.*
Ключові слова: *залежність даних, набір даних, алгоритм, аналіз даних.*

Аннотация. *В статье описан разработанный алгоритм поиска зависимостей в подмножествах объектов исследуемого набора данных. Данный алгоритм позволяет эффективно выявлять ассоциативные зависимости в данных по заданным критериям качества. Использование разработанного метода поиска зависимостей в данных возможно в многих предметных областях и позволяет выявить новые закономерности в данных, что улучшает работу специалистов и качество принятых ими решений.*
Ключевые слова: *зависимость данных, набор данных, алгоритм, анализ данных.*

Abstract. *The worked out algorithm of data search dependencies in the object subsets of investigated dataset is described in the article. This algorithm allows effectively reveal associative dependencies based on specified quality criteria. Application of developed data dependency detection method is possible in many specializations and allows finding new data patterns which improves work and decisions quality of specialists.*
Keywords: *data dependency, dataset, algorithm, data analysis.*

1. Вступ

Аналіз результатів будь-яких досліджень чи спостережень є невід'ємною частиною процесу дослідження предметної області. Напрям аналізу інформації швидко зростає і поширюється на нові й нові галузі науки і техніки. Водночас з галузевим поширенням методи аналізу даних застосовуються до все більших наборів даних. Це потребує розробки нових підходів та алгоритмів у даній галузі. Хоча на сьогодні наявні потужні методи та засоби аналізу даних у деяких вузьких галузях, наприклад, CLASSIFI (Department of Pathology, UT Southwestern Medical Center) [1], BiNGO (Department of Plant Systems Biology, VIB/Ghent University) [2] та EASE (National Institute of Allergy and Infectious Diseases) [3], проте більшість зібраної інформації не аналізується належним чином перш за все через недостатні обчислювальні потужності для наявних алгоритмів і відсутність ефективніших методів аналізу цих даних. Внаслідок цього багато залежностей предметних областей залишаються невідомими експертам та аналітикам і це не дозволяє приймати правильні рішення у проблемних ситуаціях, що приносить великі збитки як економічні, так і екологічні, медичні та ін. Тому розробка алгоритму, застосовного до широкого спектра схем даних, є надзвичайно актуальною задачею.

Метою даного дослідження є розробка ефективного алгоритму аналізу широкого спектра даних, що дозволить фахівцям предметних областей краще зрозуміти залежності в цих галузях та приймати правильні рішення у проблемних ситуаціях.

2. Основна частина

Аналіз великих обсягів даних потребує виявлення груп атрибутів, що утворюють функціональні залежності. Проте в реальних умовах значно частіше зустрічаються набори даних, в

© Пшеничний О.Ю., 2012

ISSN 1028-9763. Математичні машини і системи, 2012, № 1

яких важливі залежності визначені тільки на деякій підмножині значень групи ключових атрибутів. Називатимемо такі залежності частковими функціональними залежностями (ЧФЗ).

Тобто, часткова функціональна залежність – це ФЗ в деякій селекції основного відношення.

$$F_p:\ K=\{a_i\}, a_i \in A, D=\{a_j\}, a_j \in A, R' \subset R: K \to D\big|R'. \quad (1)$$

Багато залежностей також мають не чітко детермінований характер. Називатимемо їх імовірнісними продукційними залежностями (ІПЗ).

Імовірнісна продукційна залежність – це продукційне правило в селекції основного відношення, яке справджується для значущої кількості об'єктів цієї селекції. Поріг значущості повинен визначатись експертним шляхом або виходячи з розрахунків імовірності помилкового виділення цієї залежності.

$$F_I:\ K=\{a_i\}, a_i \in A, D=\{a_j\}, a_j \in A,: P\left(k \in K \to d \in D\right)=p. \quad (2)$$

Тут k та d – кортежі значень деяких груп атрибутів K та D відповідно.

Основним показником достовірності такої залежності є відношення кількості об'єктів, для яких має місце така ФЗ, до кількості об'єктів у селекції:

$$P\left(F_I\right)=\frac{\left|\sigma_{k\in K \wedge d\in D}(R)\right|}{\left|\sigma_{k\in K}(R)\right|}. \quad (3)$$

Позначатимемо ІЗ F_I з використанням цієї величини як імовірність виконання продукційного правила:

$$P\left(k \in K \to d \in D\right)=p. \quad (4)$$

При потребі неважко розширити поняття до імовірності виконання довільних продукційних правил. Але пошук складних продукційних правил у великих наборах даних є дуже складною і обчислювально важкою операцією, тому вони використовуються рідко.

Введення поняття імовірнісних залежностей, на перший погляд, надлишкове, оскільки з множини об'єктів, на яких визначена така залежність, завжди можна вибрати підмножину, на якій визначена чітка часткова функціональна залежність. Але за умови, що значення групи ключових атрибутів такої залежності мають змістовне сортування у предметній області, опис залежностей може бути здійснений значно простіше та інформативніше. Для наочності розглянемо приклад:

Приклад 1.

Таблиця 1. Залежність типу вагона, в якому їде людина, від дальності поїздки

i – номер опитаного	L – дальність поїздки, х100км	T – клас вагона
1	1	3
2	1	2
3	1	3
4	3	1
5	2	1
6	1	3
7	2	1
8	2	2
9	3	1

З табл. 1 можна зробити висновок.

Такі ЧФЗ:

$$i \to (T,L)\big|\{1..9\},$$

$$L \to T\big|\{3\},$$

$$T \to L\big|\{3\},$$

$$(L,T) \to i\big|\{(1;2),(2;2)\}.$$

Розглядаються тільки максимально розширені ЧФ, тобто, наприклад, залежність $i \to (T,L)\big|\{1,2,3\}$ не береться

до уваги, оскільки перша ЧФ з вищенаведених повністю визначає її.

Ці залежності справджуються для всіх об'єктів, які мають вказану комбінацію визначальних атрибутів залежності, але інформативність таких залежностей часто мала через надмірний детермінізм – будь-який виняток зруйнує залежність і будь-який випадковий шум даних може утворити нову залежність.

Розглянемо деякі ІПЗ з попереднього прикладу:

$$P(i=1 \to (T;L)=(1;3))=1,$$

$$P(L=1 \to T=3)=0{,}75,$$

$$P(L \in \{2,3\} \to T \in \{1,2\})=1,$$

$$P(L \in \{2,3\} \to T=1)=0{,}8,$$

$$P(T=3 \to L=1)=1,$$

$$P(T=1 \to L=3)=0{,}5.$$

Такі дані значно інформативніші і корисніші для аналізу. Вони захищені від шуму та випадкових помилок і навіть дозволяють виявити ці помилки. Тому пошук саме таких залежностей і є завданням сучасного аналізу даних. Задачу пошуку мінімального покриття часто використовують в аналізі даних [4].

Ще одним важливим параметром ІПЗ є її достовірність – імовірність відхилення величини $P(F_I)$ не більше, ніж на задане значення δ:

$$P\left(\left|P(F_I)-\delta\right| \le P^*(F_I)\right), \tag{5}$$

де $P^*(F_I)$ – справжнє значення виконання залежності F_I.

Обчислення $P^*(F_I)$ напряму в більшості випадків неможливе, оскільки відсутня інформація про всі можливі комбінації значень атрибутів об'єктів відношення. Але можна оцінити достовірність залежності, виходячи з кількості об'єктів, для яких залежність виконується, комбінацій задіяних атрибутів та інших параметрів.

Легко зауважити, що ЧФ є лише частковим випадком ІПЗ при $P(F_I)=1$. Тому завдання пошуку залежностей у даних можна звести до пошуку ІПЗ.

У даній роботі пропонується метод виявлення ІПЗ в наборах даних, опираючись на задані параметри імовірності виконання та достовірності цих залежностей.

Розглянемо для початку пошук простих залежностей типу

$$P(s=s_1 \to t=t_1) \ge p_0, \tag{6}$$

де p_0 – порогове значення імовірності виконання шуканих залежностей, s, t – деякі атрибути відношення, в якому здійснюється пошук залежностей.

Для обчислення необхідних імовірностей та пошуку таких ІПЗ необхідно перебрати всі пари атрибутів $(s;t)$, $s \in A$, $t \in A$ і для кожної комбінації здійснити прохід по об'єктах відношення, заповнюючи для кожного значення атрибута s кількість його повторень в об'єктах та словник відповідних значень атрибута t із вказанням кількості повторів. Після таких операцій усі залежності виду (6) при $p_0=0$ будуть збережені в описаній структурі даних. Відбір потрібних ІПЗ з цієї структури для довільного p_0 є тривіальною задачею фільтрації. Складність описанного алгоритму – $O(m^2 \cdot n)$, де m – кількість атрибутів відношення, n – кількість об'єктів у відношенні.

Хоча складність вищеописаного алгоритму і залежить не лінійно від кількості атрибутів відношення, проте переважно навіть дуже складні дані мають не більше декількох сотень атрибутів. Окрім того, алгоритм має добрі дані розпаралелювання, тобто, його можна елементарно модифікувати для виконання на багатопроцесорних машинах, кластерах чи грідах.

Розширимо задачу (6) до пошуку залежностей виду

$$P\left(s = s_1 \to t \in T\right) \geq p_0 . \tag{7}$$

Маючи дані про залежності виду (6), можна розв'язати дану задачу, аналізуючи лише підгрупу залежностей виду (6), що має спільну частину $s = s_1$. Це легко довести, виходячи з визначення ІПЗ та формули (3): будь-який об'єкт, для якого діє залежність $P(s = s_2 \to q \in Q) \geq p$, не виконує залежність (7) і, відповідно, не змінює значення $P(s = s_1 \to t \in T)$. З іншої сторони, додавання чи вилучення об'єкта зі значенням $s = s_1$ неодмінно впливає на імовірність $P(s = s_1 \to t \in T)$, якщо це значення більше нуля. Отже, при пошуку залежностей виду (7) необхідно і достатньо розглянути всі залежності виду (6), у яких умовна частина предиката співпадає з шуканою залежністю.

Щодо імовірності виконання залежності (7) на заданому наборі даних, то з (3) випливає, що

$$P\left(s = s_1 \to t \in \{t_1, t_2\}\right) = \frac{\left|\sigma_{t \in \{t_1, t_2\}}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|} = \frac{\left|\sigma_{t=t_1}(R)\right| + \left|\sigma_{t=t_2}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|} =$$
$$= P\left(s = s_1 \to t = t_1\right) + P\left(s = s_1 \to t = t_2\right). \tag{8}$$

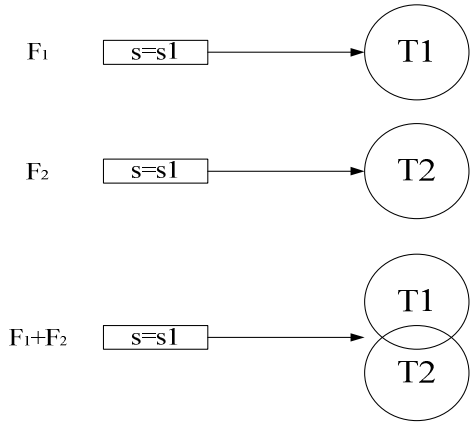

Рис. 1. Графічне представлення об'єднання ІЗ типу (7)

Тобто, об'єднання залежностей виду (6) у залежності від виду (7) збільшить імовірність виконання нової залежності. Для більш загального випадку об'єднання двох залежностей виду (7) формула обчислення імовірності виконання агрегативної залежності буде трохи складнішою, оскільки множини значень результуючої частини предиката можуть перекриватись (рис. 1).

Відповідно, виразимо з (3) імовірність виконання такої ІПЗ:

$$P\left(s = s_1 \to t \in T_1 \cup T_2\right) = \frac{\left|\sigma_{t \in T_1 \cup T_2}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|} =$$
$$= \frac{\left|\sigma_{t \in T_1}(R)\right| + \left|\sigma_{t \in T_2}(R)\right| - \left|\sigma_{t \in T_1 \cap T_2}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|} .$$

Використовуючи зворотне перетворення формули (3), легко помітити, що

$$\frac{\left|\sigma_{t \in T_1}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|} = P\left(s = s_1 \to t \in T_1\right),$$
$$\frac{\left|\sigma_{t \in T_2}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|} = P\left(s = s_1 \to t \in T_2\right),$$

$$\frac{\left|\sigma_{t\in T_1\cap T_2}(R)\right|}{\left|\sigma_{s=s_1}(R)\right|}=P\left(s=s_1\to t\in T_1\cap T_2\right),$$

і утворене відношення можна переписати простіше:

$$P\left(s=s_1\to t\in T_1\cup T_2\right)=P\left(s=s_1\to t\in T_1\right)+P\left(s=s_1\to t\in T_2\right)-P\left(s=s_1\to t\in T_1\cap T_2\right). \quad (9)$$

Враховуючи обчислені значення імовірностей всіх атомарних залежностей (6), кон'юнкція та диз'юнкція довільних множин значень атрибута T не потребує нових складних обчислень для знаходження імовірності виконання об'єднаної залежності.

Ще одним варіантом об'єднання ІПЗ виду (6) є залежності

$$P\left(s\in S\to t=t_1\right)\geq p_0. \quad (10)$$

Обчислення імовірності такої агрегативної залежності трохи складніше, ніж у попередньому випадку і неможливе, якщо спиратись тільки на значення $P\left(s=s_i\to t=t_1\right)$. Для цього потрібно залишати імовірності залежностей виду (6) у вигляді звичайних дробів – відношення кількості об'єктів, для яких залежність виконується до загальної кількості об'єктів, відібраних умовною частиною предиката ІПЗ, як і було запропоновано в алгоритмі пошуку ІПЗ виду (6). Тоді, виходячи з формули (3),

$$P\left(s\in\{s_i\}\to t=t_1\right)=\frac{\sum_i\left|s=s_i\wedge t=t_1\right|}{\sum_i\left|s=s_i\right|}. \quad (11)$$

І знову ж, використовуючи вищеописану структуру даних, ця формула обчислюється без додаткових проходів по таблиці.

Розглянемо тепер загальний вигляд шуканих залежностей:

$$P\left(s\in S\to t\in T\right)\geq p_0. \quad (12)$$

Обчислення імовірності виконання такої залежності ґрунтується на попередніх міркуваннях та можливості розкладу такої залежності на складові ІПЗ типу (11):

$$P\left(s\in S\to t\in T\right)=\sum_{t_i\in T}P\left(s\in S\to t=t_i\right)=\sum_{t_i\in T}\frac{\sum_j\left|s=s_j\wedge t=t_i\right|}{\sum_j\left|s=s_j\right|}. \quad (13)$$

Позначення, що використовувались до цих пір у статті, досить громіздкі, тому введемо алгебру залежностей, визначену над четвірками

$$\left(S;T;\left|s\in S\wedge t\in T\right|;\left|s\in S\right|\right)=F_I, \quad (14)$$

операціями об'єднання (+) та розщеплення (-) залежностей. Для додаткового спрощення позначень введемо операцію проекції трійки F_I на її атрибути (назвемо їх відповідно Pr (Predicate), PrS, PrT (дочірні атрибути предиката $s\in S\to t\in T$ – множини S та T), PN(Probability nominator), PD (Probability Denominator) і позначатимемо операцію квадратними дужками, наприклад, $F_I[\mathrm{Pr}]$, $F_I[\mathrm{Pr}\,S]$).

Для типів ІПЗ, описаних формулами (6), (7), (10), (12), введемо позначення відповідно F^A, F^B, F^C, F^D. Тоді операції об'єднання та розщеплення над залежностями цих типів будуть визначені таким чином:

$$F_1^A + F_2^A = \begin{cases} F_1^A, \quad якщо\, F_1^A[\Pr S] = F_2^A[\Pr S] \wedge F_1^A[\Pr T] = F_2^A[\Pr T], \\ F^B = \left(F_1^A[\Pr S]; F_1^A[\Pr T] \cup F_2^A[\Pr T]; F_1^A[PN] + F_2^A[PN]; F_1^A[PT]\right), \\ \quad якщо\, F_1^A[\Pr S] = F_2^A[\Pr S] \wedge F_1^A[\Pr T] \neq F_2^A[\Pr T], \\ F^C = \left(F_1^A[\Pr S] \cup F_2^A[\Pr S]; F_1^A[\Pr T]; F_1^A[PN] + F_2^A[PN]; F_1^A[PT] + F_2^A[PT]\right), \\ \quad якщо\, F_1^A[\Pr S] \neq F_2^A[\Pr S] \wedge F_1^A[\Pr T] = F_2^A[\Pr T], \\ F^D = \left(F_1^A[\Pr S] \cup F_2^A[\Pr S]; F_1^A[\Pr T] \cup F_2^A[\Pr T]; F_1^A[PN] + F_2^A[PN]; F_1^A[PT] + F_2^A[PT]\right), \\ \quad якщо \left(F_1^A[\Pr S] \neq F_2^A[\Pr S]\right) \wedge \left(F_1^A[\Pr T] \neq F_2^A[\Pr T]\right), \end{cases} \tag{15}$$

$$F^B + F^A = \begin{cases} F^B, \quad якщо\, F^B[\Pr S] = F^A[\Pr S] \wedge F^A[\Pr T] \in F^B[\Pr T], \\ F_1^B = \left(F^B[\Pr S]; F^B[\Pr T] \cup F^A[\Pr T]; F^B[PN] + F^A[PN]; F^B[PT]\right), \\ \quad якщо\, F^B[\Pr S] = F^A[\Pr S] \wedge F^A[\Pr T] \notin F^B[\Pr T], \\ F^D = \left(F^B[\Pr S] \cup F^A[\Pr S]; F^B[\Pr T] \cup F^A[\Pr T]; F^B[PN] + F^A[PN]; F^B[PT] + F^A[PT]\right), \\ \quad якщо\, F^B[\Pr S] \neq F^A[\Pr S], \end{cases} \tag{16}$$

$$F^C + F^A = \begin{cases} F^C, \quad якщо\, F^A[\Pr S] \in F^C[\Pr S] \wedge F^C[\Pr T] = F^A[\Pr T], \\ F_1^C = \left(F^C[\Pr S] \cup F^A[\Pr S]; F^C[\Pr T]; F^C[PN] + F^A[PN]; F^C[PT] + F^A[PT]\right), \\ \quad якщо\, F^A[\Pr S] \notin F^C[\Pr S] \wedge F^C[\Pr T] = F^A[\Pr T], \\ F_1^D = \left(F^C[\Pr S]; F^C[\Pr T] \cup F^A[\Pr T]; F^C[PN] + F^A[PN]; F^C[PT]\right), \\ \quad якщо \left(F^A[\Pr S] \in F^C[\Pr S]\right) \wedge \left(F^C[\Pr T] \neq F^A[\Pr T]\right), \\ F_2^D = \left(F^C[\Pr S] \cup F^A[\Pr S]; F^C[\Pr T] \cup F^A[\Pr T]; F^C[PN] + F^A[PN]; F^C[PT] + F^A[PT]\right), \\ \quad якщо \left(F^A[\Pr S] \notin F^C[\Pr S]\right) \wedge \left(F^C[\Pr T] \neq F^A[\Pr T]\right), \end{cases} \tag{17}$$

$$F^D + F^A = \begin{cases} F^D, \quad якщо \left(F^A[\Pr S] \in F^D[\Pr S]\right) \wedge \left(F^A[\Pr T] \in F^D[\Pr T]\right), \\ F_1^D = \left(F^D[\Pr S]; F^D[\Pr T] \cup F^A[\Pr T]; F^D[PN] + F^A[PN]; F^D[PT]\right), \\ \quad якщо \left(F^A[\Pr S] \in F^D[\Pr S]\right) \wedge \left(F^A[\Pr T] \notin F^D[\Pr T]\right), \\ F_2^D = \left(F^D[\Pr S] \cup F^A[\Pr S]; F^D[\Pr T] \cup F^A[\Pr T]; F^D[PN] + F^A[PN]; F^D[PT] + F^A[PT]\right), \\ \quad якщо\, \left(F^A[\Pr S] \notin F^D[\Pr S]\right). \end{cases} \tag{18}$$

Залежності типів F^B, F^C, F^D можна представити у вигляді суми ІПЗ типу F^A, виразивши їх параметри зворотними перетвореннями формул (15)–(18). Тому наводити формули обчислення сум $F^B + F^C$, $F^D + F^D$ та ін. тут немає сенсу.

Використовуючи (15)–(18), можна елементарно довести, що операція об'єднання залежностей має властивості комутативності та асоціативності на множині ІПЗ.

Використовуючи (3) та (15)–(18), визначимо операцію розщеплення ІПЗ як обернену до операції об'єднання:

$$F_1 + F_2 = F_3 \Rightarrow F_3 - F_2 = F_1. \tag{19}$$

Таким чином, операції розширення та звуження множин значень умови та результату продукційного правила ІПЗ дозволяють гнучко модифікувати ці залежності без потреби проведення додаткових складних обчислень. Необхідно тільки провести попередній аналіз простих залежностей виду (6) вищеописаним алгоритмом.

Наведені операції та алгоритм дозволяють виявити область визначення ІПЗ та знайти залежності за заданою імовірністю виконання. Недоліком залишається знаходження залежностей тільки виду (12) (вони включають (6), (7) та (10)), тобто ІПЗ, що мають лише один атрибут в умовній частині предиката та один – у результуючій. Тобто, неможливий пошук залежностей

$$P\left(s_1 \in S_1 \wedge s_2 \in S_2 \wedge \ldots \wedge s_l \in S_l \rightarrow t_1 \in T_1 \wedge t_2 \in T_2 \wedge \ldots \wedge t_k \in T_k\right) \geq p_0 . \qquad (20)$$

Для забезпечення пошуку таких ІПЗ введемо дві нові операції: введення нового атрибуту в умовну частину предиката залежності та введення нового атрибуту в результуючу частину предиката. Позначимо їх $\otimes$ та $\oplus$ відповідно.

Для залежностей (20) розширимо поняття операцій $F_I[\Pr]$, $F_I[\Pr S]$. Вони повертають множину значень кортежу атрибутів умови та результуючої частини предиката відповідно. Ця множина значень є декартовим добутком множин допустимих значень окремих атрибутів кортежу:

$$\begin{aligned} F_I[\Pr S] &= S_1 \times S_2 \times \ldots \times S_l, \\ F_I[\Pr T] &= T_1 \times T_2 \times \ldots \times T_k. \end{aligned} \qquad (21)$$

Операції з ІПЗ тепер визначимо як

$$F_I \oplus (x \in X) = F_I^{'} = \left(F_I[\Pr S]; F_I[\Pr T] \times X; \left|\sigma_{x \in X}\left(\sigma_{k \in F_I[\Pr S]}(R)\right)\right|; \left|\sigma_{k \in F_I[\Pr S] \wedge d \in F_I[\Pr T]}(R)\right|\right), \quad (22)$$

$$F_I \otimes (x \in X) = F_I^{'} = \left(F_I[\Pr S] \times X; F_I[\Pr T]; \left|\sigma_{x \in X}\left(\sigma_{k \in F_I[\Pr S]}(R)\right)\right|; \left|\sigma_{x \in X}\left(\sigma_{k \in F_I[\Pr S] \wedge d \in F_I[\Pr T]}(R)\right)\right|\right). \quad (23)$$

Тут $\sigma_{\Pr}(R)$ – операція селекції по предикату Pr над відношенням R.

Поки не вдалось розробити ефективний метод обчислення множин $\sigma_{x \in X}\left(\sigma_{k \in F_I[\Pr S]}(R)\right)$ та $\sigma_{x \in X}\left(\sigma_{k \in F_I[\Pr S] \wedge d \in F_I[\Pr T]}(R)\right)$ по наявних $\sigma_{k \in F_I[\Pr S]}(R)$ та $\sigma_{k \in F_I[\Pr S] \wedge d \in F_I[\Pr T]}(R)$ (їх можна зберігати для кожної аналізованої залежності в ході обчислень для спрощення побудови мультиатрибутних залежностей). Єдиним виходом залишається пряме виконання вказаних операцій селекції.

Отже, алгоритм пошуку ІПЗ розширюється додатковим кроком – пошуком мультиатрибутних залежностей. Для цього залежності з кроку 1 алгоритму аналізуються на наявність кореляцій з іншими атрибутами відношення. Повний перебір можливих розширень ІПЗ до мультиатрибутних буде мати складність $O\left(z^2 \cdot m \cdot sz\right)$, де z – кількість знайдених ІПЗ, m – кількість атрибутів відношення, sz – середній розмір множини кортежів відношення R, на яких визначена ІПЗ.

Для оптимізації цього процесу можна перебирати пари залежностей виду (6) – $\left(F_1^A; F_2^A\right)$, які мають значне перекриття множин значень результуючої частини предиката:

$$\frac{F_1^A[\Pr T] \cap F_2^A[\Pr T]}{F_1^A[\Pr T] \cup F_2^A[\Pr T]} \geq c . \qquad (24)$$

Значення c визначається експертним шляхом, залежно від структури даних та типів залежностей, які слід шукати.

Пари залежностей, що задовольняють умову (24), комбінуються операціями об'єднання (якщо умовні частини предиката мають однакові атрибути) чи додавання атрибуту в умову предиката (якщо умовні частини предиката мають різні атрибути), а утворені залежності додаються у множину ІПЗ для подальшого агрегування.

Використання умови (24) дозволяє значно скоротити перебір можливих мультиатрибутних залежностей і забезпечити складність виконання другого кроку алгоритму $O\left(z^2 \cdot \log(sz) + z^* \cdot sz\right)$, де z^* – кількість пар ІПЗ, що задовольняють умову (24). Обчислення формули (24) може бути виконано з використанням структури даних binomial heap або Fibonaccі heap зі складністю $O\left(\log(sz)\right)$. Таким чином, якщо кількість початкових залежностей не надто велика (порядку декількох тисяч), досягається прийнятний час аналізу мультиатрибутних залежностей на сучасних обчислювальних системах. Контролювати кількість одноатрибутних залежностей можна вибором порогового значення імовірності p_0, а перебір мультиатрибутних залежностей – зміною константи c з умови (24).

Таким чином, описаний у даній статті алгоритм дозволяє проводити ефективний аналіз однотипних даних на наявність мультиатрибутних ІПЗ та ЧФ, сумарна складність якого $O\left(m^2 \cdot n + z^2 \cdot \log(sz) + z^* \cdot sz\right)$. При розумному виборі значень p_0 та c на сучасних обчислювальних центрах він дає змогу аналізувати набори даних з тисячами атрибутів та мільйонами кортежів відношення.

Розглянемо продовження прикладу 1 для демонстрації пошуку мультиатрибутних залежностей. Виберемо значення $c = 0,7$. Розглянемо одноатрибутні ІЗ $P\left(L = 1 \to T \in \{2,3\}\right) = 1$ та $P\left(i \in \{1,2,3\} \to T = 3\right) = \frac{2}{3}$; за формулою (24) обчислюємо їх перекриття:

$$\frac{F_1^A[\Pr T] \cap F_2^A[\Pr T]}{F_1^A[\Pr T] \cup F_2^A[\Pr T]} = \frac{3}{4} \geq c.$$

Отже, обчислюємо мультиатрибутну залежність, утворену додаванням атрибуту i на множині значень $\{1,2,3\}$ в залежність $P\left(L = 1 \to T \in \{2,3\}\right) = 1$:

$$P\left(L = 1 \wedge i \in \{1,2,3\} \to T \in \{2,3\}\right) = \frac{\left|\sigma_{i \in \{1,2,3\}}\left(\sigma_{L=1}(R)\right)\right|}{\left|\sigma_{i \in \{1,2,3\}}\left(\sigma_{L=1 \wedge T \in \{2,3\}}(R)\right)\right|} = \frac{3}{3} = 1.$$

Таким чином, утворилась нова мультиатрибутна ІПЗ.

3. Висновки

Описаний у даній статті алгоритм дозволяє проводити аналіз великих обсягів даних, які можуть бути представлені у вигляді відношення (на сучасних обчислювальних центрах за кілька годин можна провести аналіз відношення з тисячами атрибутів та мільйонами записів). Алгоритм достатньо легко можна модифікувати для паралельного виконання на декількох обчислювальних машинах, що при потребі дозволить зменшити час аналізу до хвилин.

Система аналізу даних, яка використовуватиме вищеописаний алгоритм, дозволить значно розширити межі застосування методів аналізу даних у напрямку розміру аналізованих відношень, що, у свою чергу, дозволить застосовувати їх до предметних галузей, в яких накопичуються надзвичайно великі архівні бази даних, та виявити нові важливі залежності в цих галузях.

Основним недоліком розробленого алгоритму є те, що він не використовує вже наявних знань про специфіку предметної галузі і дозволяє виявити далеко не всі типи можливих залежностей (його завданням є пошук імовірнісних продукційних залежностей). Ще одним недоліком є недостатня ефективність пошуку мультиатрибутних залежностей і для

оптимізації цього процесу доводиться вводити додатковий коефіцієнт, на якому базується евристичний алгоритм пошуку таких залежностей.

І все ж, не зважаючи на описані недоліки, алгоритм дозволяє виявляти дуже широкий клас залежностей у даних і може бути застосований у найрізноманітніших галузях науки і техніки.

СПИСОК ЛІТЕРАТУРИ

1. Головний сайт департаменту патології, UT Southwestern Medical Center [Електронний ресурс]. – Режим доступу: http://pathcuric1.swmed.edu/pathdb/classifi.html.
2. Опис утиліти BiNGO, сайт університету Гент [Електронний ресурс]. – Режим доступу http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html.
3. Офіційний сайт National Institute of Allergy and Infectious Diseases (NIAID), NIH [Електронний ресурс]. – Режим доступу http://david.abcc.ncifcrf.gov/content.jsp?file=/ease/ease1.htm&type=1.
4. Мейер Д. Теория реляционных баз данных / Мейер Д.; пер. с англ. – М.: Мир, 1987. – 610 с.

Стаття надійшла до редакції 14.07.2011