УДК 004.9:530.1

І.Б. Швороб[1]
[1]Національний університет «Львівська політехніка», м. Львів, Україна

ПІДХІД ДО РОБОТИ ЗІ СЛАБОСТРУКТУРОВАНИМИ МЕДИЧНИМИ ДАНИМИ НА ОСНОВІ ВИКОРИСТАННЯ ВАГ РЕБЕР ДЛЯ ДОКУМЕНТО-ОРІЄНТОВАНОЇ ГРАФОВОЇ БАЗИ ДАНИХ

В роботі запропоновано підхід до роботи з слабоструктурованими даними збереженими у вигляді документо-орієнтованої графової бази даних. Розглянуто способи зображення зваженого графа та наведено спосіб зображення документо-орієнтованого зваженого графа. Наведено алгоритм перерахунку ваг ребер результуючого графа на основі врахування оцінок якості попереднього вибору даних з результуючого графа. Проведено експериментальне дослідження роботи запропонованого алгоритму для роботи зі слабоструктурованими медичними даними.

СЛАБОСТРУКТУРОВАНІ ДАНІ, NOSQL, NEO4J, ГРАФОВІ БАЗИ ДАНИХ, ДОКУМЕНТО-ОРІЄНТ ОВАНІ ГРАФОВІ БАЗИ ДАНИХ, ЗВАЖЕНИЙ ГРАФ, ВАГА РЕБРА ГРАФА, АЛГОРИТМ РОЗРАХУНКУ ВАГИ РЕБРА ГРАФА

Вступ

Існуючі підходи до опрацювання неструктурованих та напівструктурованих даних на сьогоднішній день мають емпіричний підхід. Відсутність систематизації та теоретичного обгрунтування використовуваних методів та засобів опрацювання таких даних негативно впливає на застосування нових методик та методів. Однією з проблем в роботі зі слабоструктурованими даними є їх оптимальне зберігання. У роботі [1] запропоновано підхід до збереження слабоструктурованих медичних даних.

Істотною перевагою зберігання даних в документо-орієнтованій базі даних є зручність для подальшої обробки даних. Однак запити будуть складними і при запиті може прийти набагато більше інформації, ніж це необхідно. Це, в свою чергу це впливає на продуктивність. Дані в графовій базі даних займають великий об'єм. Час оптимізації бази даних графа є більшим. Графові бази даних значно домінують над реляційними при пошуці в реальному часі з великими обсягами даних. Таким чином, графові бази даних слід використовувати при наявності великих обсягів даних і ресурсів, за умови, що швидке виконання пошуку є дуже важливим.

У даній роботі основною метою є розроблення алгоритму роботи зі слабоструктурованими даними, збереженими за допомогою документо-орієнтованої графової бази даних запропонованої у роботі [1]. Проведено аналіз розробленого алгоритму на прикладі роботи зі слабоструктурованими медичними даними.

1. Аналіз останніх публікацій

В роботі [1] було розглянуто NoSQL бази даних та запропоновано об'єднання документо-орієнтованої та графової баз даних. Під час роботи зі слабоструктурованими даними важливо зберегти якомога більшу їх кількість в найзручнішій для використання формі. База даних на основі документо-орієнтованого графа включає складність вузла графа даних, тобто, коли вузолом є елемент з багатьма різними характеристиками. Така реалізація використовує документ для забезпечення гнучкості запитів до графових баз даних, а збереження у вигляді ключ/значення забезпечує швидкий пошук даних.

Об'єкт G такої бази даних може бути представлений наступним чином:

$$G = \left\{ \left\langle \left\{ [k, v] \right\}, \left\{ \left\langle e_1, \ldots e_n \right\rangle \right\} \right\rangle \right\}.$$

В роботі [1] розглянуто приклад збереження та обробки даних про лікарські засоби, а також побудовано документ-орієнтований граф на основі бази даних Neo4j. Створені вершини графа діляться на два типи: препарати і хвороби (показання та протипоказання). Ребра також двох типів: показання та протипоказання.

Таким чином, база даних на основі документо-орієнтованого графа в даному випадку є дуже зручним засобом для збереження даних. Слід зазначити, що при виборі препаратів потрібно враховувати не тільки показання та протипоказання, а також інші фактори, такі як дозування, віку і ваги пацієнта та інше.

Було здійснено аналіз обробки слабоструктурованих даних для різних типів баз даних. Аналіз проводився за такими параметрами: кількість створюваних об'єктів (документів або вузлів) (N), об'єм бази даних (W), час запису в базу даних (t), час виконання запиту з декількома умовами (tc). Для аналізу було використано 100 інструкцій для медичних препаратів. За результатами аналізу для документо-орієнтованої БД було створено 100 записів, для графової БД та документо-орієнтованого графа – по 740 записів. Об'єм баз даних становить 40,2Мб, 30.9Мб та 61.1Мб відповідно, час запису в БД – 10мс, 15мс, 20.75мс, а час виконання запиту з кількома умовами становить 2с, 1.4с, 1.3с відповідно.

В залежності від вимог до проекту та для оптимізації і пришвидшення роботи із великими об'ємами даних потрібно використовувати

відповідні бази даних. Наприклад, для програмного рішення дуже важливим є швидкий пошук, тому для цього можна використати графову базу даних. В окремих випадках можна комбінувати представлення даних у вигляді документно-орієнтованих графових баз даних.

2. Постановка задачі

Зважаючи на великий обсяг збережених слабоструктурованих даних та їхню різноманітність, доволі важко здійснювати швидкий пошук, а також результати такого пошуку можуть містити надлишкові дані. Саме тому виникає потреба в розробленні методів для покращення роботи з такими даними та отримання більш якісного результату.

Для запропонованого в роботі [1] підходу до збереження слабоструктурованих даних у вигляді документо-орієнтованого графу з метою вирішення даної проблеми варто використати переваги графової структури бази даних. Саме для цього необхідно розробити алгоритим для роботи із слабоструктурованими даними, що базується на використанні ваг ребер графа, що у свою чергу дозволить скоротити об'єм результуючих даних за рахунок ігнорування даних, ваги ребер між вершинами яких будуть меншими за встановлене значення.

3. Застосування ваг ребер для документо-орієнтованої графової бази даних

Подальше узагальнення відображення зв'язків між об'єктами слабоструктурованих даних за допомогою графових баз даних складається в приписуванні ребрам та дугам деяких кількісних значень, якісних ознак чи характерних властивостей, які називають вагою.

Означення: ***Зваженим*** називають простий граф, кожному ребру e якого приписано дійсне число $w(e)$. Це число називають ***вагою*** ребра e [2].

Вагою ребра може бути: порядкова нумерація ребер та дуг, яка показує на чергу при їх розгляданні (пріоритет чи ієрархія); довжина шляху, пропускна здатність; кількість набраних очок; характер відношень між об'єктами та ін. Зважені орієнтовані графи застосовують у мережевому плануванні, у теорії ланцюгів.

Існує багато способів зображення зваженого графа. Розглянемо деякі з них.

Нехай дано граф

$$G = (V, E)$$

де $|V|=n$, $|E|=m$

Спосіб 1. Задання матриці ваг W, яка є аналогом матриці суміжності. Для такої матриці елемент w_{ij}, у випадку якщо ребро $(v_i,v_j)\subset E$, буде позначатись як

$$w_{ij} = w\left(v_i, v_j\right)$$

Якщо ж ребро $(v_i,v_j)\subset E$, то, в залежності від задачі, яку потрібно розв'язати, елемент матриці буде позначатись як

$$w_{ij} = 0 \text{ або } w_{ij} = \infty .$$

Спосіб 2. Інколи граф задають списком ребер. Для зваженого графа під кожний елемент списку E можна відвести три комірки — дві для ребра і одну для його ваги, тобто всього потрібно $3m$ комірок.

Спосіб 3. Граф можна подати у вигляді списку суміжностей. Для зваженого графа кожен список $Adj[u]$ містить крім вказівників на всі вершини v множини $G(u)$ ще й числа $w(u,v)$.

Розглянемо об'єкт графа, поданого в роботі [1]:

$$G = \left\{\left\langle N, E\right\rangle\right\},$$

де N – вершина графа, E – множина ребер.

Об'єкт такого графа з урахуванням ваг буде мати наступний вигляд:

$$G = \left\{\left\langle N, E, W\right\rangle\right\},$$

де W – множина ваг ребер.

Множину ваг ребер подамо у наступному вигляді:

$$W = \left\{\left\langle w_1, \ldots, w_n\right\rangle\right\}.$$

Отже, документо-орієнтований зважений граф буде подано як

$$G = \left\{\left\langle \left\{k, v\right\}, \left\{\left\langle e_1, \ldots, e_n\right\rangle\right\}, \left\{\left\langle w_1, \ldots, w_n\right\rangle\right\}\right\rangle\right\}.$$

4. Алгоритм перерахунку ваги ребер

Нехай ваги ребер документо-орієнтованого графа знаходяться в межах від 0 до 1. Початковим значенням для всіх ребер визначимо 1.

Наведемо алгоритм перерахнку ваг ребер для документо-орієнтованого графа з урахуванням початкових вхідних даних та введених оцінок якості попереднього вибору даних з результуючого графа.

1. Введення початкових даних для запиту до документо-орієнтовагої графої бази даних для отримання результуючого графа.

2. Пошук в результуючому графі вершин з найкоротшим шляхом та видалення їх з результуючого графа, якщо значення шляху менше за встановлене експертом.

Для пошуку вершин з найкоротшим шляхом використовуємо алгоритм Дейкстри [3].

3. Здійснення вибору даних з результуючого графа (вибір вершин графа, що є об'єктом бази даних).

4. Введення оцінки якості вибору даних з результуючого графу (наприклад, оцінка відповідності навиків обраного працівника до вказаних в резюме, оцінка якості лікування певних симптомів обраними медичними препаратами і т.д.). Оцінка виставляється в межах від 0 до 1 (1 – максимальна оцінка якості, 0 – мінімальна), що в свою чергу стане новою вагою ребра між вершинами вхідних даних.

5. Здійснення перерахунку ваг ребер між вхідними даними та обраними раніше вершинами результуючого графа.

5.1. Перерахунок ваг ребер між вхідними даними. Нехай в нас є множина з n вершин вхідних

даних, початкові ваги ребер w_i (де i – номер вершини (від 1 до n)) між якими становила 1. Нехай для кожного ребра між вершинами введено оцінку якості і її значення становить r_i, де i – номер вершини (від 1 до n). Оновлена вага ребра між вершинами k_i буде розраховуватись за формулою:

$$k_i = w_i - r_i .$$

Якщо значення нової ваги ребра між вершинами становить 0, то таку вершину можна видалити з результуючого графа.

5.2. Перерахунок ваги ребра між вершиною вхідних даних та вершиною обраного результату. Для цього необхідно використати всі попередньо введені оцінки якості вибору. Нехай маємо масив R попередніх оцінок якості вибору. Об'єкт такого масиву буде мати наступний вигляд:

$$R = \{\langle x, f \rangle\},$$

де x – індивідуальне значення оцінки; f – повторюваність оцінки.

Використовуючи формулу середнього арифметичного зваженого знаходимо нову вагу між вершиною вхідних даних та вибраним результатом:

$$w = \frac{\sum_{i=1}^{n} x_i f_i + r}{\sum_{i=1}^{n} f_i + 1},$$

де r – нова введена оцінка.

Далі оновлюємо масив R, додавши до нього нове значення оцінки.

5. Приклад застосування роботи зі слабоструктурованими медичними даними на основі використання ваг ребер в документо-орієнтованій графовій базі даних

Розглянемо роботу розробленого алгоритму на прикладі роботи з даними про лікарські засоби. Для дослідження обрано 100 пацієнтів з однаковими симптомами: температурою та запаленням. Для роботи з даними обрано базу даних Neo4j [4]. Запит до такої бази даних з урахуванням вхідних даних – симптомів пацієнта, буде мати наступний вигляд:

```
MATCH(p:Disease)-[:INDICATION]->
(m:Antibiotic)
WHERE p.name = «Infection» AND p.name =
«Temperature»
RETURN m.title
```

В результаті отримаємо граф, зображений на рис. 1.

Створені вершини графа будуть трьох типів: паці-єнт, хвороби та препарати. Ребра будуть двох типів: симптоми та показання. На початку дослідження всі ваги ребер встановлюються в 1.

Встановлюємо експертне значення для порівняння з найкоротшим шляхом між вершинами рівним 0,1. Тобто всі шляхи між вершинами симптомів та вершинами препаратів, які менші за 0,1 будуть видалятись.

Враховуючи, що для даного прикладу ваги всіх ребер одинакові, пропускаємо крок з пошуком найкоротшого шляху від симптомів до препаратів.

Нехай для заданих симптомів буде обрано препарат Ампіцилін. Тоді після здійснення вибору отримуємо новий результуючий граф, зображений на рис. 2.

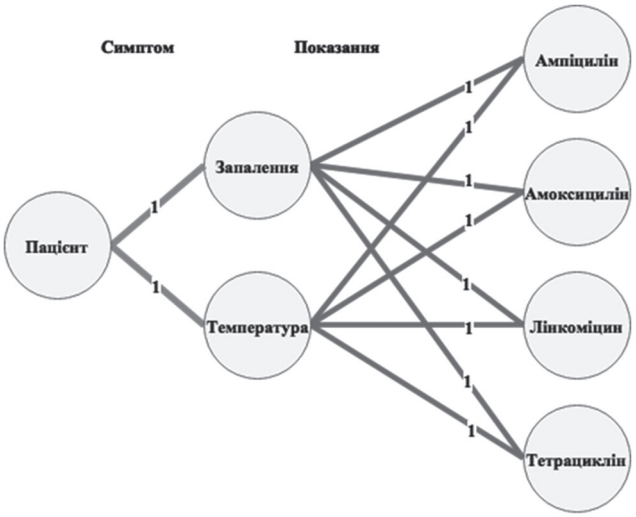

Рис. 1. Результат виконання запиту до документо-орієнтованої графової бази даних

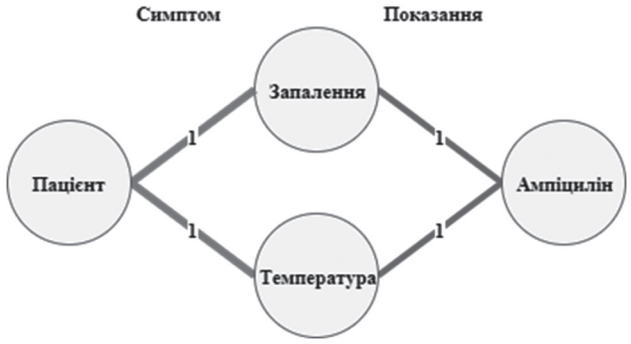

Рис. 2. Граф-відображення призначеного лікування пацієнтові

Після проведення лікування, здійснюється повторний огляд пацієнта та лікарем вводиться оцінка якості лікування даним препаратом, тобто визначається чи допоміг обраний препарат вилікувати симптоми пацієнта. Оцінка виставляється в межах від 0 до 1. Якщо симптом зник повністю, то виставляється оцінка 1, якщо зовсім не зник — то 0. Якщо симптом вилікувано частково, то оцінка виставляється на розсуд лікаря. Якщо симптом вилікувано повністю, то вага ребра між вершинами пацієнта та симптома стає рівною 0, а, отже, вершина даного симптома видаляється з результуючого графа.

В даному прикладі припустимо, що в пацієнта повністю зник симптом Температура і встановлено оцінку 1, але ознаки симптому Запалення ще проявляються і визначено оцінку якості лікування як 0,82. В зв'язку з цим, за формулою, наведеною у пункті 5.1, обраховуємо нову вагу ребра між вершинами пацієнта та симптому:

1) вага між вершиною пацієнта та вершиною симптому Температура:

$$k_1 = w_1 - r_1 = 1 - 1 = 0;$$

2) вага між вершиною пацієнта та вершиною симптому Запалення:

$$k_2 = w_2 - r_2 = 1 - 0{,}82 = 0{,}18.$$

Враховуючи, що в даному прикладі це перша експертна оцінка якості, то за формулою, наведеною у пункті 5.2. масив *R* кожного ребра буде мати 0 елементів. Обраховуємо нову вагу ребра між вершиною обраного препарату та вершиною симптому:

1) вага між вершиною симптому Температура та вершиною та вершиною препарату Ампіцилін:

$$w = \frac{0+r}{0+1} = \frac{r}{1} = \frac{1}{1} = 1.$$

В масив *R* обраного ребра додаємо новий об'єкт <1;1>.

2) вага між вершиною симптому Запалення та вершиною препарату Ампіцилін:

$$w = \frac{0+r}{0+1} = \frac{r}{1} = \frac{0{,}82}{1} = 0{,}82.$$

В масив *R* обраного ребра додаємо новий об'єкт <0,82;1>.

На рис. 3 зображено новий результуючий граф з перерахованими вагами та видаленою вершиною вилікуваного симптома.

Рис. 3. Результуючий граф з урахуванням введеної оцінки якості результату вибору

Для наступного пацієнта з такими ж симптомами, результуючий граф зображено на рис. 4.

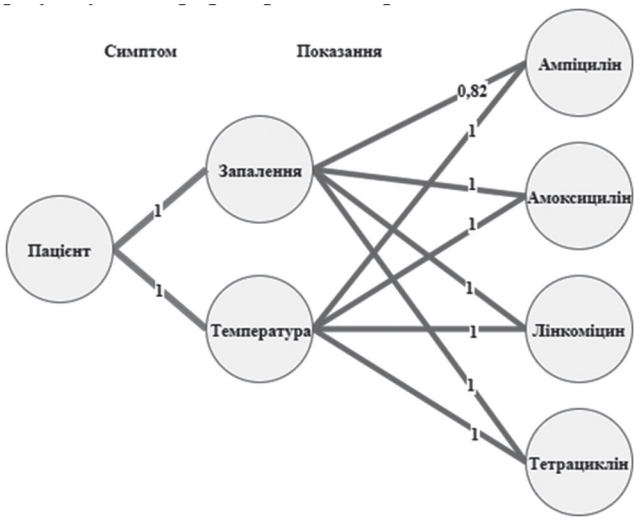

Рис. 4. Результат виконання запиту до документо-орієнтованої графової бази даних з урахуванням оцінки якості результату вибору при визначенні ваг ребер

Пошук по даному графові вершин з найкоротшим шляхом покаже, що найкоротша відстань між вершинами симптому Запалення та вершиною препарату Ампіцилін і становить 0,82. Проте, оскільки отримане значення більше за встановлене експертне значення, то вершину препарату Ампіцилін не видаляємо.

Припустимо, що лікарем для лікування нового пацієнта знову було обрано препарат Ампіцилін. Новий результуючий буде мати вигляд, наведений на рис. 5.

Нехай в даному випадку препарат повністю вилікував симптоми пацієнта, відповідно ваги між вершиною пацієнта та вершинами симптомів (пункт 5.1.) будуть становити 0, а вершини симптомів будуть видалені з результуючого графа.

За формулою, наведеною у пункті 5.2, перераховуємо ваги ребер між вершинами симптомів та вершиною препарату.

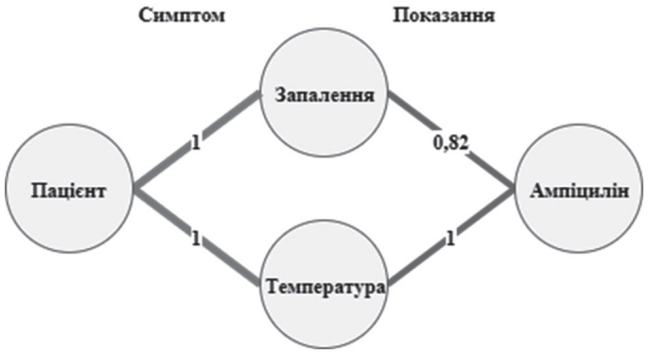

Рис. 5. Граф-відображення призначеного лікування новому пацієнтові

1) вага між вершиною симптому Температура та вершиною та вершиною препарату Ампіцилін:

$$w = \frac{\sum_{i=1}^{1} x_i f_i + r}{\sum_{i=1}^{1} f_i + 1} = \frac{1*1+1}{1+1} = \frac{2}{2} = 1.$$

В масиві *R* обраного ребра оновлюємо значення кількості оцінок в існуючому об'єкті <1;2>.

2) вага між вершиною симптому Запалення та вершиною препарату Ампіцилін:

$$w = \frac{\sum_{i=1}^{1} x_i f_i + r}{\sum_{i=1}^{1} f_i + 1} = \frac{0{,}82*1+1}{1+1} = \frac{1{,}82}{2} = 0{,}91.$$

В масив *R* обраного ребра додаємо новий об'єкт <1;1>.

В результаті, для наступного пацієнта з такими ж симптомами, результуючий граф буде мати вигляд як на рис. 6.

Після опрацювання даних 50-ти пацієнтів отримаємо граф, зображений на рис.7.

Знаходимо найкоротший шлях між вершинами. За алгоритмом Дейкстри це буде значення ваги графа між вершиною симптома Температурв та Вершиною препарату Лінкоміцин і становить 0,2, проте це значення більше за встановлене експертне значення, тому вершину препарату не видаляємо.

Для наступного пацієнта було обрано препарат Амоксіцилін. Результуючий граф зображено на рис. 8.

Припустимо, що в результаті лікування обраний препарат частково вилікував симптоми пацієнта. Здійснюємо перерахунок ваг ребер-симптомів:

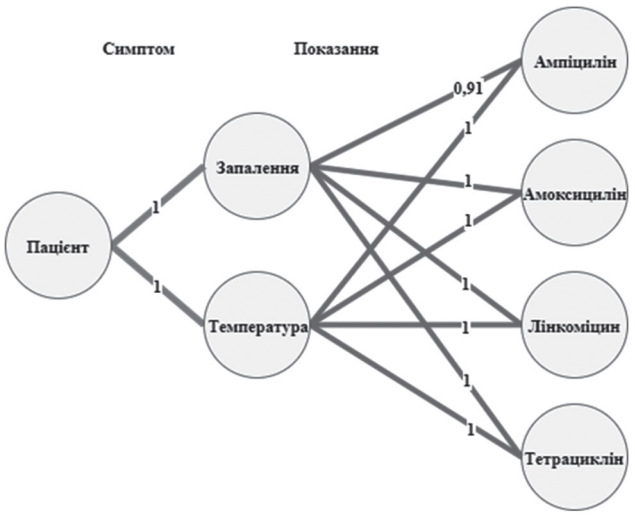

Рис. 6. Результат виконання запиту до документо-орієнтованої графової бази даних з урахуванням оцінки якості результату вибору при визначенні ваг ребер

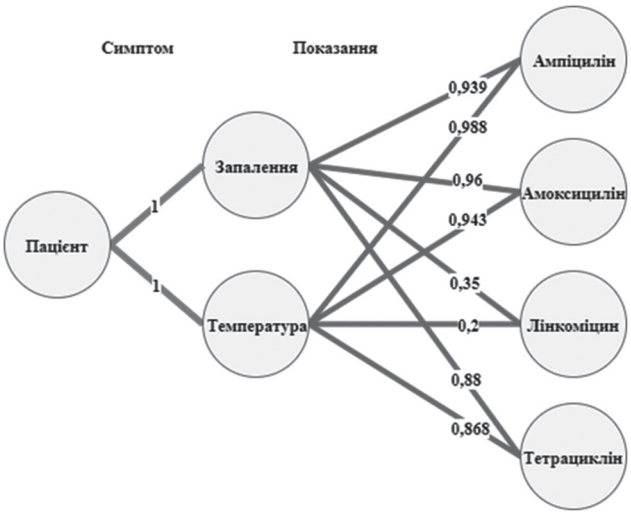

Рис. 7. Документо-орієнтований граф після опрацювання вибірки з 50 пацієнтів

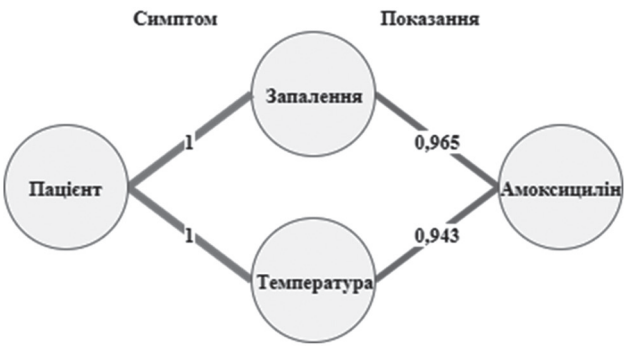

Рис. 8. Граф-відображення призначення лікування 51 пацієнтові

1) вага між вершиною пацієнта та вершиною симптому Температура зі вказаною оцінкою якості $r_1 = 0,91$:

$$k_1 = w_1 - r_1 = 1 - 0,91 = 0,09;$$

2) вага між вершиною пацієнта та вершиною симптому Запалення зі вказаною оцінкою якості r_2=0,95:

$$k_2 = w_2 - r_2 = 1 - 0,95 = 0,05.$$

Обраховуємо нову вагу ребра між вершиною обраного препарату та вершиною симптому:

1) вага між вершиною симптому Температура та вершиною та вершиною препарату Амоксіцилін:

$$w = \frac{1*5 + 0,95*5 + 0,91*3 + 0,73*1 + 0,91}{5+5+3+1+1} = 0,941.$$

В масиві R обраного ребра оновлюємо значення існуючого об'єкта <0,91;4>.

2) вага між вершиною симптому Запалення та вершиною препарату Амоксіцилін:

$$w = \frac{1*7 + 0,95*5 + 0,88*2 + 0,95}{7+5+2+1} = 0,964.$$

В масиві R обраного ребра оновлюємо значення існуючого об'єкта <0,95;6>.

Результуючий граф для даного пацієнта після введення оцінок якості результату вибору зображено на рис. 9.

Рис. 9. Граф-відображення обраного лікування для пацієнта після введення оцінок якості результатів вибору лікування

Після опрацювання вибірки зі 100 пацієнтів з однаковими симптомами, отримуємо граф, зображений на рис. 10.

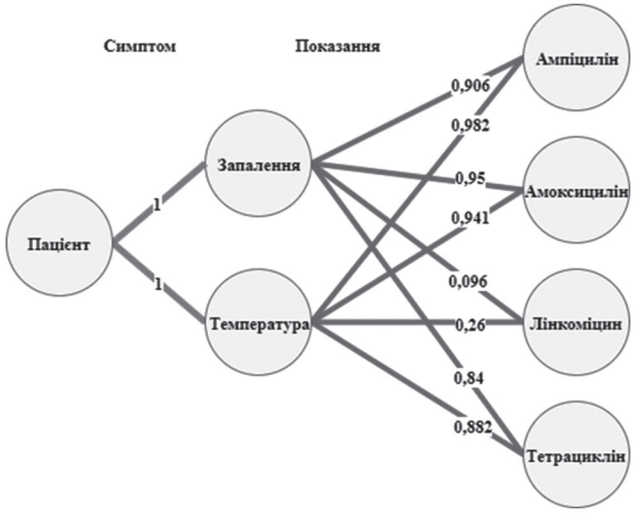

Рис. 10. Результат запиту до документо-орієнтованої графової бази даних

Після пошуку найкоротшого шляху між вершинами бачимо, що найкоротшим шляхом є відстань між вершинами симптому Запалення та вершиною препарату Лінкоміцин і становить 0,096, що менше за експертне значення. Отже, вибрану вершину необхідно видалити з графу і при наступному пошуку дана вершина препарату відображатись не буде. Кінцевий граф після проходження запропонованого алгоритму зображено на рис. 11.

Отже, запропонований алгоритм є оптимальним рішенням для роботи з великими об'ємами слабострктурованих даних, оскільки дозволяє відкидати дані з низькою ефективністю при певному вхідному наборі параметрів.

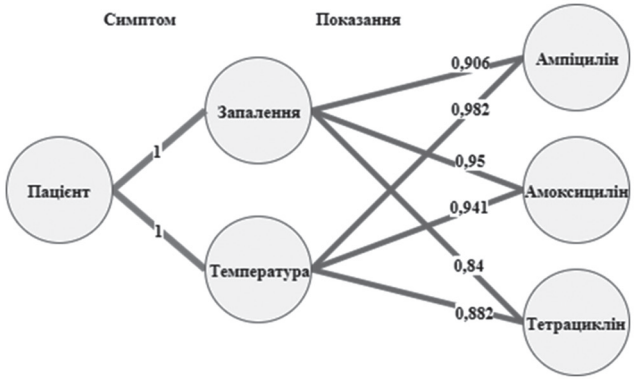

Рис. 11. Граф-відображення проведеного дослідження запропонованого алгоритму

Висновки

В роботі запропоновано алгоритм роботи зі слабоструктурованими даними, збереженими за допомогою документо-орієнтованої графової бази даних [1], на основі застосування ваг ребер графа.

Використання ваг ребер графа, розрахованих на основі оцінок якості результатів вибору даних, дозволяє визначати оптимальні для подальшої роботи дані та відкидати неефективні при певному вхідному наборі параметрів дані.

Було проведено дослідження на основі слабоструктурованих медичних даних, де за запропонованим алгоритмом здійснювався вибір лікарем препаратів для лікування симптомів пацієнта та, на основі введення оцінок якості лікування розраховувались ваги ребер результуючого графа та визначались найбільш ефективні препарати, а найменш ефективні не брались до уваги. Слід зазначити, що при виборі препаратів потрібно враховувати не тільки показання та протипоказання, а також інші фактори, такі як дозування, вік і вага пацієнта та інше. Подальші дослідження будуть спрямовані на розширення та оптимізацію базу даних для врахування цих факторів, що дозволить точніше вибирати препарат, а також прогнозувати більш ефективний препарат при заданих умовах.

Запропонований підхід можна використати як засіб для допомоги прийняття рішень лікарю або в іншій сфері пов'язаній із роботою зі слабоструктурованими даними.

Список літератури:

1. *Шворoб, І.Б.* Новий підхід до збереження слабоструктурованих медичних даних / І.Б. Шворoб. – Науковий вісник НЛТУ України – 2016 р. – Вип. 26.4 – 382-390 с. **2.***Нікольський Ю.В., Пасічник В.В., Щербина Ю.М.* Дискретна математика: Підручник. – Львів: «Магнолія2006», 2008. – 608 с. **3.** *Dijkstra E. W.* A note on two problems in connexion with graphs // Numer. Math — Springer Science+Business Media, 1959. — Vol. 1, Iss. 1. — P. 269–271. **4.** *Robinson I.* Graph Databases/ Robinson I., Webber J., Eifrem E. – O'Reilly Media, Inc., 2015. – pp.25-53.

Надійшла до редколегії 30.05.2017.