УДК 004.89

В.В. Литвин
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

СПОСІБ ВВЕДЕННЯ МЕТРИКИ ДЛЯ ВИЗНАЧЕННЯ ВІДСТАНІ МІЖ ТЕКСТОВИМИ ДОКУМЕНТАМИ

Розглянуто метод оцінювання подібності документів на основі введення метрики між концептуальними графами, що є відповідними моделями цих документів. Для цього вводиться поняття центру ваг концептуального графа, а відстань між документами визначається як відстань між їх центрами ваг. Щоб існував шлях між центрами ваг концептуальних графів, відповідні моделі документів доповнюються контекстом з онтології.

This article considers documents similarity evaluation method, using metrics between conceptual graphs – models of these documents. The idea of "center of weight" is used for solving this task. Difference between the documents is calculated as distance between graph's centers of weight. If the way between centers doesn't exist document models are expanded with context from ontology.

Постановка проблеми у загальному вигляді

Практична побудова інтелектуальних агентів розпізнавання змісту електронних текстових документів пов'язана з формуванням та наповненням їх бази знань (БЗ). Огляд літератури свідчить про значні труднощі зі створенням баз знань, придатних для їх промислової експлуатації. Розроблено та експлуатується багато програмних засобів ручного та інтерактивного напів-автоматичного формування баз знань, проте, фактично, ручне наповнення зумовлює великі фінансові та часові затрати, переважно не сумісні з комерційним застосуванням подібних інтелектуальних систем (ІС). Підходи до моделювання інтелектуальних систем, зокрема, їх онтологій не враховують агентної природи таких систем. У зв'язку з цим не ставиться задача оцінки ваги складових елементів БЗ та нових знань, які можуть її доповнити, тим самим знижується ефективність порівняння текстових документів на предмет їх подібності за змістом до взірцевого. Тому побудова оцінки подібності електронних документів з врахуванням інформаційної ваги елементів БЗ, коли процес формування та наповнення БЗ здійснюється автоматизовано є актуальною науково-технічною задачею.

Аналіз останніх досліджень

Статистичні та семантичні методи порівняння (векторно-просторова модель, міра на основі коефіцієнта Дайса, латентно-семантичне індексування, порівняння концептуальних графів), були запропоновані свого часу П. Фолтсом, С. Думайсом, Дж. Солтоном, Е. Расмусеном, М. Монтес-Гомезом [1–4] та іншими.

Формування цілей

Розглянуто метод оцінювання подібності документів на основі введення метрики між концептуальними графами, що є відповідними моделями цих документів. Для знаходження відстані між документами пропонується моделі цих документів доповнити контекстом з онтології.

Основний матеріал.

Постановка задачі

Ефективність адаптації онтології бази знань до особливостей предметної області визначають закладені в її структуру та функції механізми оптимізації шляхом самонавчання під час експлуатації. Одним з підходів до реалізації таких механізмів є автоматичне зважування понять бази знань та семантичних зв'язків між ними під час самонавчання. Цю роль беруть на себе коефіцієнти важливості понять та зв'язків. Їх розподіл у БЗ має відповідати таким основним вимогам:

відображати семантичну вагу понять предметної області, в якій ця інтелектуальна система реально застосовуватиметься;

формуватися під час наповнення БЗ та коректуватися відповідно до визначених правил;

забезпечувати контроль цілісності БЗ;

відповідати вимогам метрики при їх використанні для порівняння семантичної близькості понять.

Стоїть задача сформулювати відповідний набір правил присвоєння вагових коефіцієнтів (інформаційної ваги) поняттям та твердженням в моделі БЗ, що забезпечить оцінку актуальної цінності її інформаційного наповнення та досліджуваних текстових документів, а також створить передумови для розв'язання задачі пошуку оптимальної кількості елементів у БЗ. На основі визначених правил розробити процедуру оцінювання подібності двох електронних текстових документів.

Правила визначення інформаційної ваги елементів онтології БЗ

Покажемо можливість вирішення сформульованої задачі шляхом введення деяких спрощень і припущень. Подамо базу знань у вигляді іменованого графу, числові семантичні характеристики вершин і ребер якого визначаються за певними правилами. Він є орієнтованим зваженим мультиграфом з такими властивостями:

1) на кожний елемент (вершину) може бути довільна кількість посилань;

2) кожний елемент може мати зв'язок з будь-якою кількістю інших елементів;

3) кожному зв'язку (ребру) у моделі відповідає певний напрям і коефіцієнт важливості зв'язку та достовірності відповідного твердження, кожному поняттю (вершині) – коефіцієнти важливості поняття.

Коефіцієнт важливості поняття (зв'язку) – це числова міра, котра характеризує значимість цього поняття (зв'язку) у конкретній предметній області і динамічно змінюється за певними правилами в процесі експлуатації системи [5].

Наш підхід до представлення знань у формі зваженої семантичної мережі (концептуальних графів) полягає у тому, що будь-яке можливе узагальнення, тобто комплексне, складене поняття завжди явно артикульоване, назване і як окремий концепт фігурує в базі знань. Тому якщо деяке узагальнення має спільні властивості чи способи функціонування, вони фізично можуть бути реалізовані через властивості та обробники подій відповідного узагальнюючого концепта.

Як показано в роботі [6], модель онтології бази знань можна подати у вигляді четвірки $G(C,R,W_C,L_R)$, де C – скінченна множина вершин, яка представляє атомарні поняття ПО; $R \subseteq C \times C$ – множина дуг, семантичні зв'язки між атомарними поняттями ПО; W_C – вага вершини; L_R – вага зв'язку. Тоді інформаційну вагу елементів онтології БЗ визначають за такими правилами:

1. Перерахунок ваги вершин відбувається по вертикалі знизу догори.

2. Розрахунок вагових коефіцієнтів поняття є рекурсивною процедурою. Повна вага W_j^i класу онтології дорівнює сумі власної ваги Wo_j^i, ваги підкласів Ws_j^i та ваги суміжних класів Wn_j^i (класів, пов'язаних з цим класом не is-a зв'язком):

$$W_j^i = Wo_j^i + Ws_j^i + Wn_j^i, \qquad (1)$$

де $Ws_j^i = \sum_k Wc_k^{i+1} \cdot L_{j,k}$ − вага k підкласів j-го класу i-го рівня; $Wc_k^{i+1} = Wo_k^{i+1} + Ws_k^{i+1}$ − вага класу C_k^{i+1}; $L_{j,k}$ − вага зв'язку між класами C_j^i та C_k^{i+1}.

Схема перерахунку окремих компонент повної ваги класу показана на рис. 1.

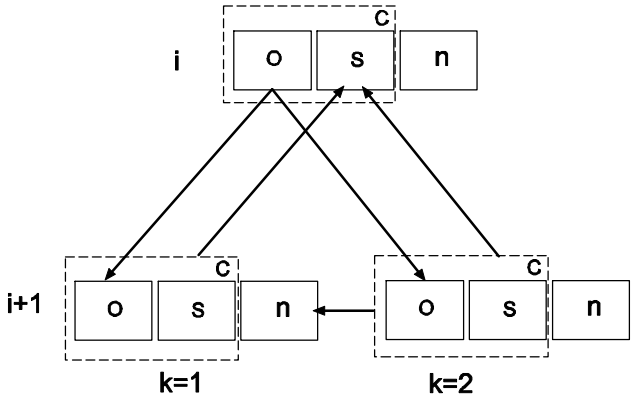

Рис.1. Схема перерахунку окремих компонент повної ваги класу

3. У момент внесення на i+1-й рівень нового підкласу йому присвоюється власна вага Wo_j^{i+1}, що дорівнює половині власної ваги класу, вищого і-го рівня.

4. Під час встановлення зв'язку між поняттями k_1 та k_2 між відповідними вершинами графа БЗ з'являється ребро, а до ваги суміжних класів Wn_1 додається вага Wc_2 і навпаки – до Wn_2 додається вага нового суміжного до нього класу Wc_1.

5. Вага екземпляра дорівнює повній вазі його класу.

Цей набір правил покладено в основу автоматичного перерахунку ваги класів онтології та екземплярів бази знань в процесі її експлуатації при реалізації методу семантичного ранжування електронних текстових документів.

Розрахунок центру ваг графу та обчислення відстані між графами

Для порівняння текстових документів запропоновано застосувати їх представлення у вигляді зважених концептуальних графів, визначити їх центр інформаційної ваги та обчислити семантичну відстань між такими центрами. Це дає змогу, по-перше, порівнювати тексти незалежно від їх розміру, по-друге, оцінювати релевантність досліджуваного тексту до заданої онтології, представленої відповідним концептуальним графом [5].

Оцінювання подібності текстів за змістом полягає у наступному:

1. Порівнювані тексти подаємо у вигляді їх концептуальних графів;

2. Графи доповнюємо відповідним контекстом та коефіцієнтами важливості з адаптивної онтології. Детально процедури адаптації онтології описано в роботі [6];

3. Відстань між двома вершинами графу C_i та C_j, якщо ці вершини з'єднані дугою, визначаємо як:

$$d_{ij} = \frac{Q}{L_{ij}(W_i + W_j)}, \qquad (2)$$

де W_i та W_j – коефіцієнти важливості вершин C_i та C_j відповідно; L_{ij} – коефіцієнт важливості зв'язку між вершинами; Q – константа, яка залежить від конкретної онтології. Приймемо, що $L_{ii} = \infty$, тоді $d_{ii} = 0$.

1. Знаходимо центр ваг концептуального графу. Це вершина C_{i^*}, для якої середня відстань $\overline{d}_i$ найменша:

$$\overline{d}_{i^*} = \min_i \overline{d}_i \,. \tag{3}$$

Середня відстань $\overline{d}_i$ для вершини C_i обчислюється за формулою:

$$\overline{d}_i = \frac{\sum_{j=1, j\neq i}^{n} d_{ij}^*}{n-1}, \tag{4}$$

де n – кількість вершин графу, d_{ij}^* – найкоротший шлях між вершинами C_i та C_j, який обчислюється за допомогою відомих алгоритмів, наприклад, Форда, Дейкстри, Флойда–Уоршалла [7];

2. Накладаємо порівнювані графи.

а) якщо вони мають спільні дуги, то відстань між вершинами, з'єднаними такими дугами, визначається як середня відстань двох графів:

$$\overline{d}^{12} = \frac{\overline{d}^1 + \overline{d}^2}{2}; \tag{5}$$

б) якщо дуги не є спільними, то відстань між вершинами береться із відповідного графу.

3. Обчислюємо найкоротший шлях між центрами ваг КГ, яка слугуватиме оцінкою подібності двох електронних документів.

$$d^{12} = \min d(C^1, C^2), \tag{6}$$

де C^1 – центр ваги 1-го графу, C^2 – центр ваги 2-го графу. Найкоротший шлях між вершинами обчислюємо за допомогою алгоритму Дейкстри.

За отриманою відстанню визначається подібність між двома документами, яким відповідають ці концептуальні графи.

Розглянемо приклад. Нехай граф містить чотири вершини з вагами $W_1 = W_2 = W_3 = W_4 = 1$, з'єднаних дугами, як це показано на рис. 2. $L_{12} = 5$, $L_{13} = 7$, $L_{14} = 6$, $L_{24} = 8$. Значення Q приймемо дорівнює 1.

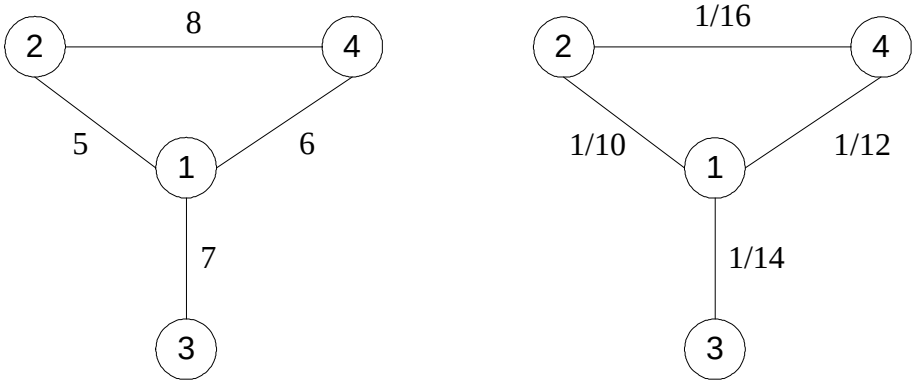

Рис. 2. Приклад зваженого графу з вагами вершин, що дорівнюють одиниці

Знайдемо відстань між вершинами та найкоротші віддалі для всіх пар вершин.

$$d_{12} = \frac{1}{5\cdot(1+1)} = \frac{1}{10},\ d_{13} = \frac{1}{7\cdot(1+1)} = \frac{1}{14},\ d_{14} = \frac{1}{6\cdot(1+1)} = \frac{1}{12},\ d_{24} = \frac{1}{8\cdot(1+1)} = \frac{1}{16}.$$

$d_{12}^* = \frac{1}{10}$ $d_{12}^* = \frac{1}{10}$ $d_{31}^* = \frac{1}{14}$ $d_{41}^* = \frac{1}{12}$

$d_{13}^* = \frac{1}{14}$ $d_{23}^* = \frac{1}{10} + \frac{1}{14} = \frac{6}{35}$ $d_{32}^* = \frac{6}{35}$ $d_{42}^* = \frac{1}{16}$

$d_{14}^* = \frac{1}{12}$ $d_{24}^* = \frac{1}{16}$ $d_{34}^* = \frac{1}{14} + \frac{1}{12} = \frac{13}{84}$ $d_{43}^* = \frac{13}{84}$

Знайдемо середні відстані для кожної із вершин:

$$\bar{d}_1 = \frac{\frac{1}{10} + \frac{1}{14} + \frac{1}{12}}{3} = \frac{107}{1260} \approx 0,085,\ \bar{d}_2 = \frac{\frac{1}{10} + \frac{6}{35} + \frac{1}{16}}{3} = \frac{187}{1680} \approx 0,111$$

$$\bar{d}_3 = \frac{\frac{1}{14} + \frac{6}{35} + \frac{13}{84}}{3} = \frac{167}{1260} \approx 0,13,\ \bar{d}_4 = \frac{\frac{1}{12} + \frac{1}{16} + \frac{13}{84}}{3} = \frac{91}{1008} \approx 0,09.$$

Отже, центром ваг цього графу буде вершина C_1. Дійсно, такий результат був передбачуваний, оскільки ця вершина з'єднана з усіма іншими вершинами і сила зв'язків є досить значною, незважаючи на те, що сила зв'язку між 2-ю та 4-ю вершиною є найбільшою.

Подивимось, як зміниться центр ваг, якщо вершини графу змінять свою вагу. Для цього розглянемо такий приклад: візьмемо структуру того самого графу, тільки змінимо ваги вершин: $W_1 = 0,2$; $W_2 = 0,3$; $W_3 = 0,8$; $W_4 = 0,5$ (див. рис. 3).

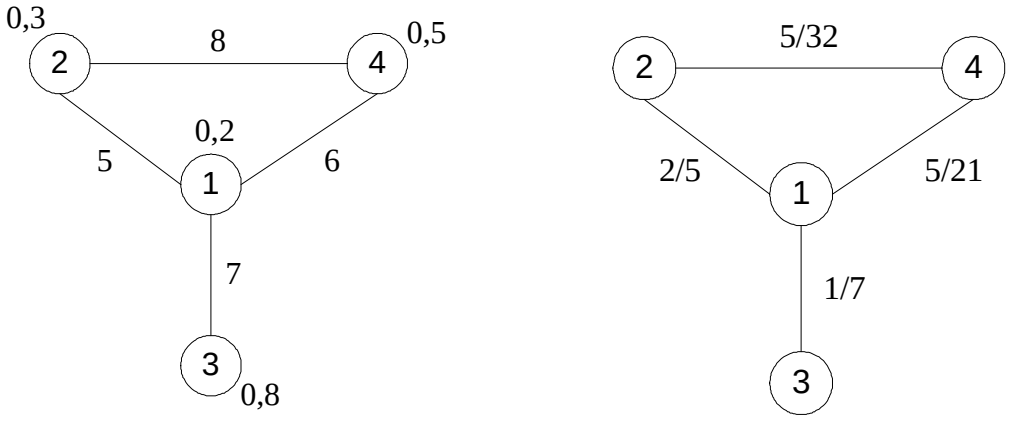

Рис. 3. Приклад зваженого графу з різними вагами вершин

Тоді $d_{12} = \frac{1}{5 \cdot (0,3+0,2)} = 0,4$, $d_{13} = \frac{1}{7 \cdot (0,2+0,8)} = 0,143$, $d_{14} = \frac{1}{6 \cdot (0,2+0,5)} = 0,238$,

$d_{24} = \frac{1}{8 \cdot (0,3+0,5)} = 0,156$.

$d_{12}^* = 0,4$ $d_{12}^* = 0,4$ $d_{31}^* = 0,143$ $d_{41}^* = 0,238$

$d_{13}^* = 0,143$ $d_{23}^* = 0,543$ $d_{32}^* = 0,543$ $d_{42}^* = 0,156$

$d_{14}^* = 0,238$ $d_{24}^* = 0,156$ $d_{34}^* = 0,38$ $d_{43}^* = 0,38$

$\bar{d}_1 = 0,26$; $\bar{d}_2 = 0,366$; $\bar{d}_3 = 0,356$; $\bar{d}_4 = 0,258$.

Отже, для таких значень ваг вершин та дуг графу центр ваг буде у вершині C_4. Наслідком такої зміни центру ваг (з вершини C_1 у вершину C_4) стала різниця ваг між цими вершинами.

Основною вимогою до запропонованого методу оцінювання подібності (семантичного порівняння чи ранжування) електронних документів є його відповідність аксіомам метрики.

Дійсно, згідно з визначенням відстані, автоматично виконуються дві перші аксіоми:

$$d(C_i, C_i) = 0;$$
$$d(C_i, C_j) = d(C_j, C_i).$$

Нехай R_{ij}^* – шлях між вершинами C_i та C_j, який відповідає відстані між ними. Тоді $d_{ij} = d_{ik} + d_{kj}$, якщо вершина C_k лежить на шляху R_{ij}^* і $d_{ij} < d_{ik} + d_{kj}$, якщо вершина C_k не лежить на шляху R_{ij}^*. А це означає, що виконується третя аксіома метрики.

Визначену так відстань можна використовувати для ранжування текстових документів, знаходження їх подібності до взірцевого документа тощо.

Приклад використання методу

Як приклад розглянемо три анотації статей із журналів “Штучний інтелект”.

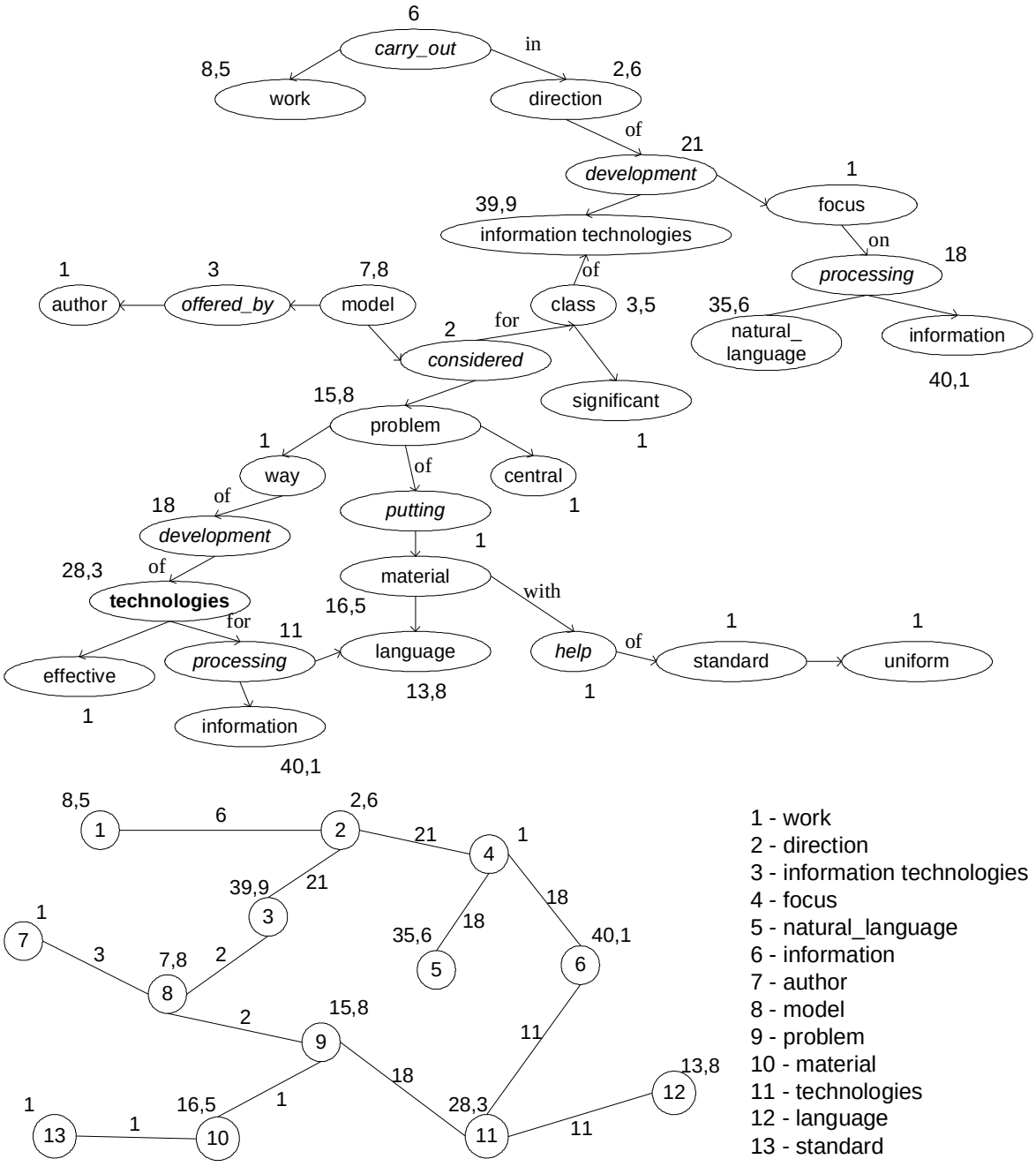

Рис. 4. Концептуальний граф 1-ї анотації

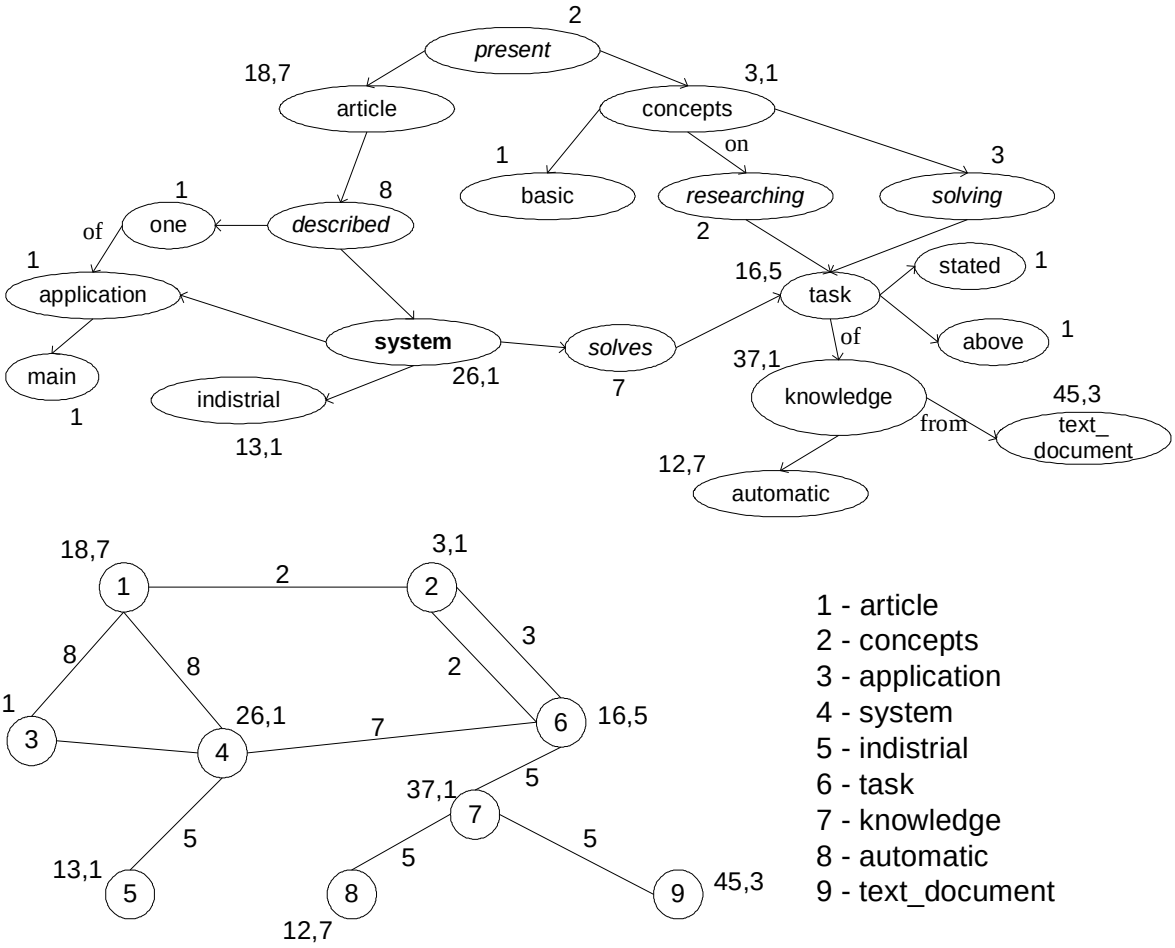

Рис. 5. Концептуальний граф 2-ї анотації

1. *The work is carried out in a direction of information technologies development focused on the natural language information processing. On position of author offered model the problem of putting a language material in order with the help of the uniform standard, it is considered rather significant for the given class of technologies. It is one of the central problem on way of development of the effective technologies for language information processing.* (ШІ, 2004, № 4, с. 613).

2. *The article presents the basic concepts on researching and solving the task of automatic knowledge retrieval from text documents. Industrial system that solves the stated above task is described as well as one of its main application.* (ШІ, 2004. – № 3. – С. 668).

3. *The paper is dedicated to the problem of automated text consistency analysis. It is proposed to implement text consistency via text logic analysis with the attraction of the knowledge of application domain, natural language and normative base.* (ШІ, 2004, № 3, с. 660).

Відповідні концептуальні графи подані на рис. 4–6, ваги зв’язків взяті із наперед побудованої онтології.

Якщо у концептуальному графі немає зв’язку, то вага такого зв’язку дорівнювала 5, що є середнім значенням ваг зв’язків.

Користуючись формулою (2), в якій Q дорівнювало 100, отримаємо такі графи (див. рис. 7).

Використовуючи алгоритм Дейкстри та роблячи відповідні обчислення за формулами (3), (4), отримаємо, що

$\overline{d}^1 = \{11\} = \{\text{'technologies'}\}$, $\overline{d}^2 = \{4\} = \{\text{'system'}\}$, $\overline{d}^3 = \{5\} = \{\text{'knowledge'}\}$,

де верхній індекс вказує номер тексту.

1-й і 3-й тексти легко пов'язуються за допомогою вершини **natural_language** (5-та в 1-му тексті і 7-ма в 3-му тексті). Тоді $d^{1*}_{5,11} = 0,42$, $d^{3*}_{5,7} = 0,23$. А відстань між 1-м і 3-м текстом дорівнює 0,65. Тоді $\overline{d}^{13} = 0,42 + 0,23 = 0,65$.

2-й і 3-й тексти легко пов'язуються за допомогою вершини **knowledge** (7-ма в 2-му тексті і 5-та в 3-му тексті). Тоді $d^{2*}_{4,7} = 0,71$, що є відстанню між 2-м та 3-м текстом, оскільки **knowledge** – центр ваг 3-го графу Тобто $\overline{d}^{23} = 0,71$.

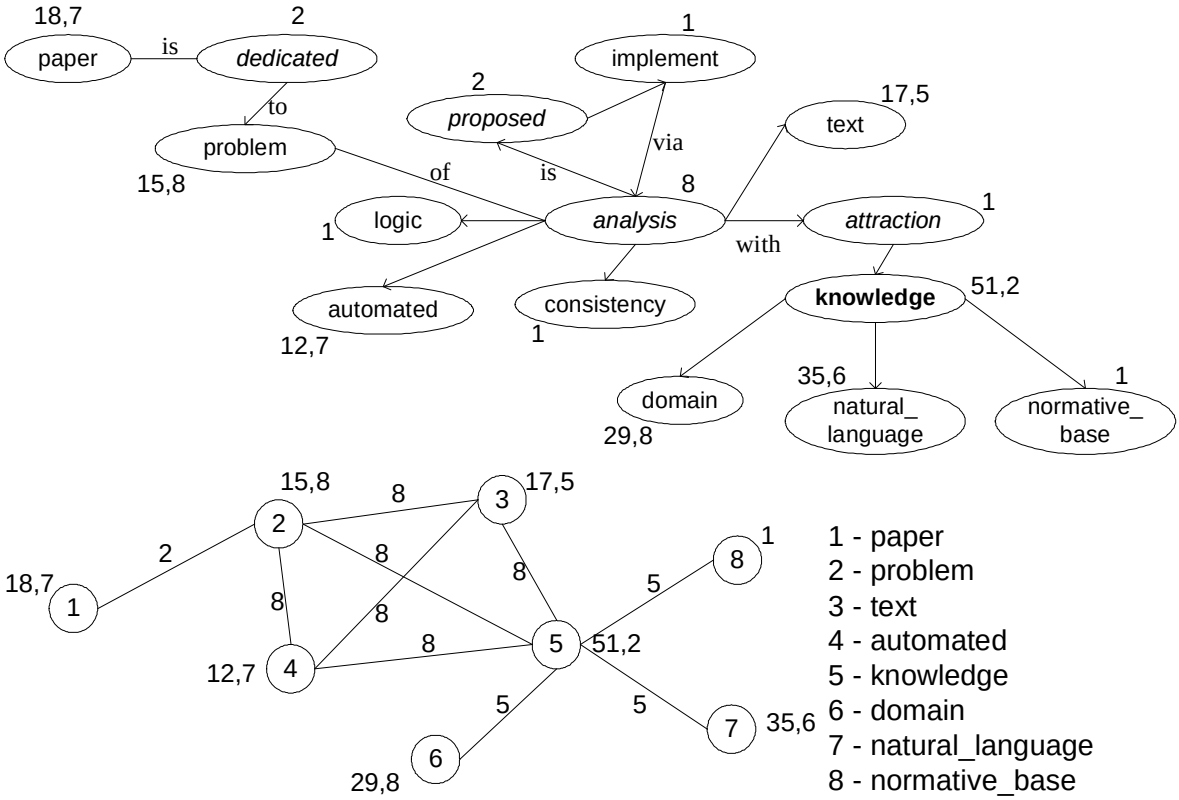

Рис. 6. Концептуальний граф 3-ї анотації

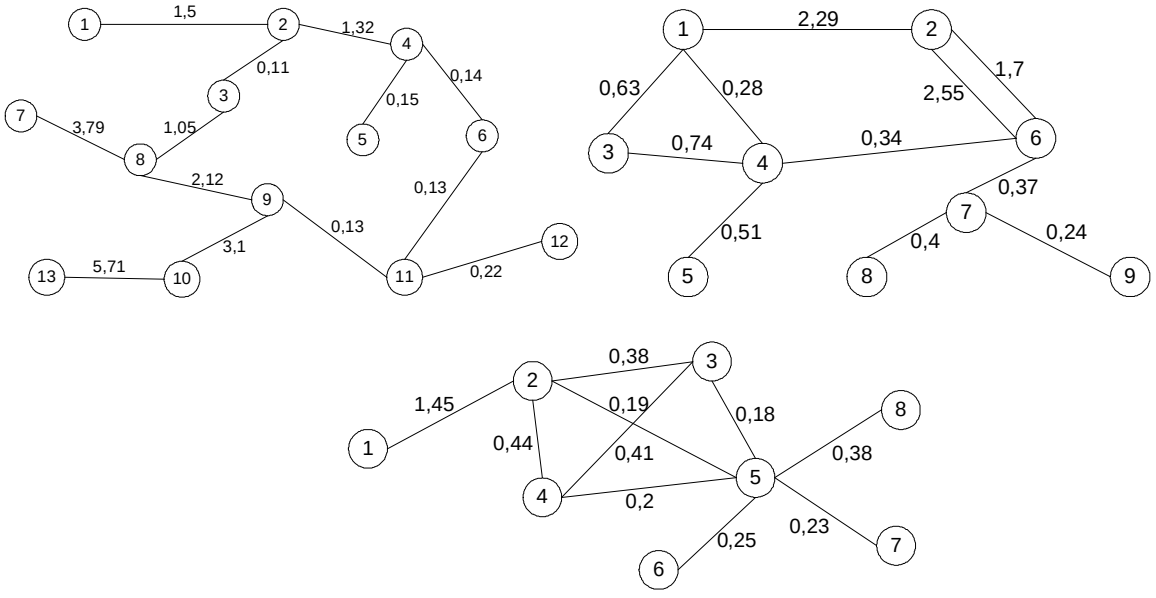

Рис. 7. Зважені графи для трьох анотацій

Для обчислення відстані між 1-м і 2-м текстом скористаємось 3-м текстом, оскільки в перших двох текстах немає спільних вершин. Тоді $\bar{d}^{12} = \bar{d}^{13} + \bar{d}^{23} = 0,65 + 0,71 = 1,36$.

Отже, найближчими за змістом є 1-й і 3-й текст, найвіддаленішими є 1-й і 2-й текст.

Треба зауважити, що вагу вершин та ребер відповідних концептуальних графів для обчислення відстані було взято з тестової модельної онтології, що не гарантує надійності отриманих результатів.

Висновки

Запропонований спосіб введення метрики дає змогу оцінити семантичну близькість двох текстових документів. Під час обчислення відстані між тестовими документами враховується контекст документів і відповідна до контексту семантика вжитих у них термінів та словосполучень. А також цей спосіб ґрунтується на правилах визначення інформаційної ваги елементів бази знань. Це дає можливість здійснювати автоматичний пошук документів, котрі найбільше відповідають запиту-прототипу, і відкидати такі, що мають малу вагу і не відповідають предметній області.

Метод перевірено на відповідність трьом вимогам метрики. Введено метрику подібності двох документів, на основі якої робиться висновок про релевантність цього документа. За допомогою розробленого методу можна здійснювати автоматичний пошук документів, котрі найбільше відповідають запиту-прототипу в мережі Інтернет, а також виконувати інтелектуальний аналіз тексту, його класифікацію та ранжування за релевантністю до заданої ПО.

1. Foltz P., Dumais S. Personalised Information Delivery: Analysis of Information Filtering Methods. Communications of the ACM 35(12), 1992. 2. Rasmussen E. Clustering Algorithms. Information Retrieval: Data Structures & Algorithms. William B. Frakes and Ricardo Baeza-Yates (Eds.), Prentice Hall, 1992. 3. Montes-y-Gómez M., Gelbukh A., López-López A. Comparison of Conceptual Graphs. Mexican International Conference on Artificial Intelligence MICAI 2000, Acapulco, Mexico, April 2000. Lecture Notes in Artificial Intelligence N 1793, Springer-Verlag, 2000. 4. Montes-y-Gómez M., Gelbukh A., López-López A., Baeza-Yates. R. Flexible Comparison of Conceptual Graphs. 12th International Conference on Database and Expert Systems Applications DEXA 2001, Munich, Germany, September 2001. Lecture Notes in Computer Science, vol. 2113, Springer-Verlag, 2001. 5. John F Sowa. "Knowledge Representation: Logical, Philosophical and Computational Foundations". 1-st edition, Thomson Learning, 1999. 6. Даревич Р.Р. Підвищення точності пошуку текстових документів на основі адаптивної онтології // Компютінг. – Т. 6. 2007. Вип. 1. 7. Седжвик Р. Фундаментальные алгоритмы на С++. Алгоритмы на графах: Пер. с англ. / Р. Седжвик. – СПб: ООО "ДиаСофтЮП", 2002. – 496 с.