УДК 004.89

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

МЕТОД ОЦІНЮВАННЯ ПОДІБНОСТІ ДОКУМЕНТІВ, ДОПОВНЕНИХ КОНТЕКСТОМ З ОНТОЛОГІЇ

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 графу, а відстань між документами визначається як відстань між їх центрами ваг. Щоб існував шлях між центрами ваг концептуальних графів, відповідні моделі документів доповнюється контекстом з онтології. Для переведення здайденої відстані в ймовірнісну оцінку використовується регресійний аналіз.

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. Transformation of distance in probabilistic observation is made using regression analysis.

Постановка проблеми у загальному вигляді

Менеджмент в наукових дослідженнях, як і будь-який інший менеджмент, передбачає облік та управління ресурсами. У науковому менеджменті доводиться керувати як традиційними ресурсами на зразок фінансів, майна та кадрів, так і ресурсами, специфічними для наукової сфери: інформацією, знаннями, авторськими правами, науковим реноме. Від ефективності адміністрування усім переліком ресурсів істотно залежить ефективність наукових досліджень. Ключем для розв’язання проблеми забезпечення ефективності наукового менеджменту є особливий вид ресурсів, здатних слугувати медіатором середовища управління – інформаційні ресурси. Від їх організації та застосування залежать прозорість, своєчасність і компетентність управлінських рішень. Для підвищення ефективності досліджень необхідно вести оперативний інформаційний пошук, облік ресурсів та забезпечити об’єктивний аналіз і оцінку результатів досліджень. Сучасний розвиток інформаційних технологій створює необхідні передумови впровадження програмно-технічних систем, здатних автоматизувати розв’язання цих задач. Зокрема, таким засобом є віртуальне автоматизоване робоче місце (ВАРМ) працівника.

Основним ресурсом, який забезпечує ефективність наукових досліджень, є актуальна науково-технічна інформація. Її опрацювання потребує відповідного інструментарію. Таким інструментарієм у складі ВАРМ працівника є його функціональне ядро – так званий інтелектуальний агент у вигляді метапошукової системи (МПС), що здатна в процесі самонавчання адаптуватися до конкретних інформаційних потреб користувача та виявляти, зберігати і використовувати релевантні до відповідних задач знання. Основою такої інтелектуальної системи є адаптивна онтологія бази знань, яка являє собою таксономію понять, пов’язаних семантичними зв’язками, здатна налаштовуватись на певну предметну область шляхом зміни своєї структури і значень параметрів (див. рис. 1).

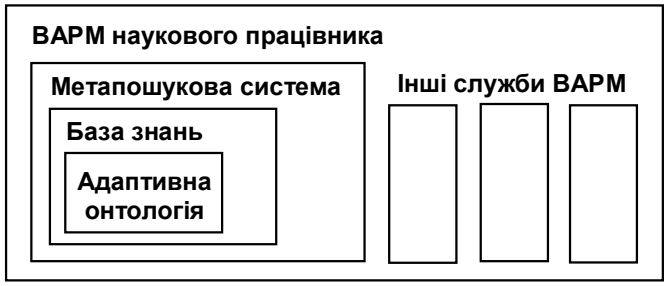

Рис. 1. Місце інтелектуальної метапошукової системи у загальній структурі АРМ працівника

Адаптована до інформаційних потреб користувачів онтологія використовується метапошуковою системою для оцінювання подібності (порівняння) текстових документів за змістом. Загальним недоліком розроблених методів є їх низька точність порівняння документів за змістом. Водночас для автоматизованих систем, в яких не передбачено інтерактивної взаємодії системи з користувачем, точність має вирішальне значення. Тому актуальною є задача розроблення методів та алгоритмів порівняння за змістом текстових документів, доповнених контекстом з адаптованої до інформаційних потреб користувача онтології.

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

Метод введення метрики для оцінювання подібності документів

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:

$$d_{ij} = \frac{Q}{L_{ij}(W_i + W_j)}, \tag{1}$$

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:

$$\overline{d}_{i^*} = \min_i \overline{d}_i. \tag{2}$$

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

$$\overline{d}_i = \frac{\sum_{j=1, j\neq i}^{n} d_{ij}^*}{n-1}, \tag{3}$$

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:

$$\overline{d}^{12} = \frac{\overline{d}^1 + \overline{d}^2}{2} \tag{4}$$

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.

$$d^{12} = \min d(C^1, C^2), \tag{5}$$

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.

Блок-схему алгоритму показано на рис. 2:

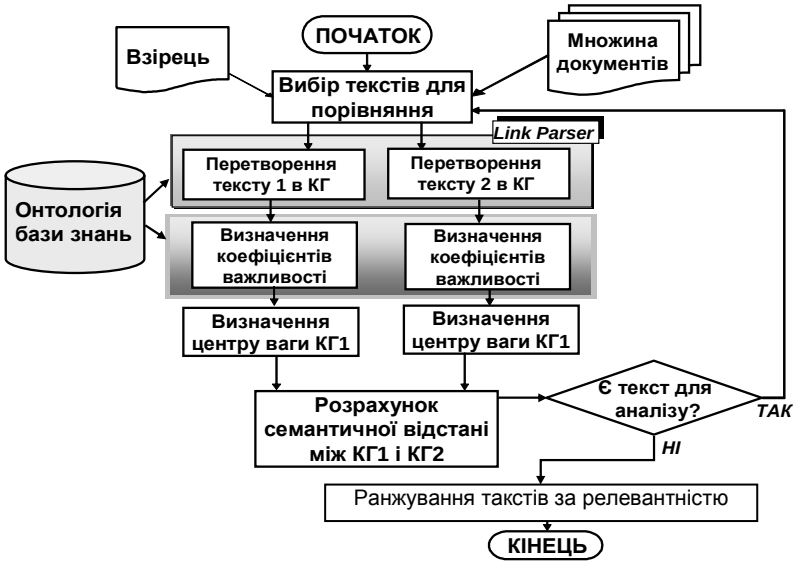

Рис. 2. Блок-схема алгоритму порівняння текстових документів з використанням онтології

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:

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

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.

Визначення ймовірності подібності двох документів методом лінійної регресії

Для розв'язання задачі знаходження ймовірності релевантності знайденого документа залежно від отриманої відстані використано регресійний аналіз. Нехай до того в процесі своєї діяльності деякий користувач отримав такі результати про релевантність документа залежно від відстані (d_i,p_i), i=1,…,n.

Для значення p_i можна використовувати двозначну логіку (0 – ні, 1 – так) або k-значну логіку, наприклад (0 – ні, 0,25 – скоріше нерелевантне, ніж релевантне, 0,5 – більш-менш, 0,75 – скоріше релевантне, ніж нерелевантне, 1 – так). Тоді, побудувавши лінію регресії для наперед заданих значень, ми зможемо в майбутньому визначати ймовірність релевантності документа на основі відстані.

Лінія регресії має вигляд: $p = a_o + a_1(d - \overline{d})$.

Коефіцієнти прямої знаходимо за формулою:

$$a_0 = \overline{p},\ a_1 = \frac{\sum_{j=1}^{n}(p_j - p)(d_j - d)}{\sum_{j=1}^{n}(d_j - \overline{d})^2}.$$

Приклад

Нехай в процесі роботи користувач отримав результати, подані у табл. 1.

Таблиця 1

Результати роботи користувача

d	*p*	$d-\overline{d}$	$p-\overline{p}$	$(d-\overline{d})^2$	$(p-\overline{p})(d-\overline{d})$	*p з прямої регресії*
1	*2*	*3*	*4*	*5*	*6*	*7*
0,1	1	-0,7	0,525862	0,49	-0,3681034	0,910345
0,15	0,75	-0,65	0,275862	0,4225	-0,1793103	0,879187
0,2	1	-0,6	0,525862	0,36	-0,3155172	0,84803
0,25	1	-0,55	0,525862	0,3025	-0,2892241	0,816872
0,3	0,75	-0,5	0,275862	0,25	-0,137931	0,785714
0,35	0,5	-0,45	0,025862	0,2025	-0,0116379	0,754557
0,4	0,75	-0,4	0,275862	0,16	-0,1103448	0,723399
0,45	1	-0,35	0,525862	0,1225	-0,1840517	0,692241
0,5	0,5	-0,3	0,025862	0,09	-0,0077586	0,661084
0,55	0,25	-0,25	-0,22414	0,0625	0,0560345	0,629926
0,6	0,5	-0,2	0,025862	0,04	-0,0051724	0,598768
0,65	0,75	-0,15	0,275862	0,0225	-0,0413793	0,567611
0,7	1	-0,1	0,525862	0,01	-0,0525862	0,536453
0,75	0,25	-0,05	-0,22414	0,0025	0,0112069	0,505296
0,8	0,5	0	0,025862	0	0	0,474138
0,85	0,5	0,05	0,025862	0,0025	0,0012931	0,44298
0,9	0,5	0,1	0,025862	0,01	0,0025862	0,411823
0,95	0,25	0,15	-0,22414	0,0225	-0,0336207	0,380665
1	0,25	0,2	-0,22414	0,04	-0,0448276	0,349507
1,05	0,25	0,25	-0,22414	0,0625	-0,0560345	0,31835
1,1	0,25	0,3	-0,22414	0,09	-0,0672414	0,287192
1,15	0,25	0,35	-0,22414	0,1225	-0,0784483	0,256034
1,2	0,25	0,4	-0,22414	0,16	-0,0896552	0,224877
1,25	0	0,45	-0,47414	0,2025	-0,2133621	0,193719
1,3	0	0,5	-0,47414	0,25	-0,237069	0,162562
1,35	0	0,55	-0,47414	0,3025	-0,2607759	0,131404
1,4	0,25	0,6	-0,22414	0,36	-0,1344828	0,100246
1,45	0,5	0,65	0,025862	0,4225	0,0168103	0,069089
1,5	0	0,7	-0,47414	0,49	-0,3318966	0,037931

$$\overline{d} = 0,8;\ \overline{p} = 0,47;\ \sum_{j=1}^{n}(p_j - \overline{p})(d_j - \overline{d}) = -3,16;\ \sum_{j=1}^{n}(d_j - \overline{d})^2 = 5,075.$$

Тоді

$$a_0 = 0,47\,;\, a_1 = -0,62\,.$$

Рівняння прямої регресії матиме вигляд:

$$p = 0,47 - 0,62(d - 0,8)$$

В останньому стовпчику табл. 1 відображено значення ймовірності, отримане з прямої регресії за відповідними значеннями *d*. Поточково дано пряму, наведену на рис. 3, також на цьому рисунку бачимо початкові задані точки.

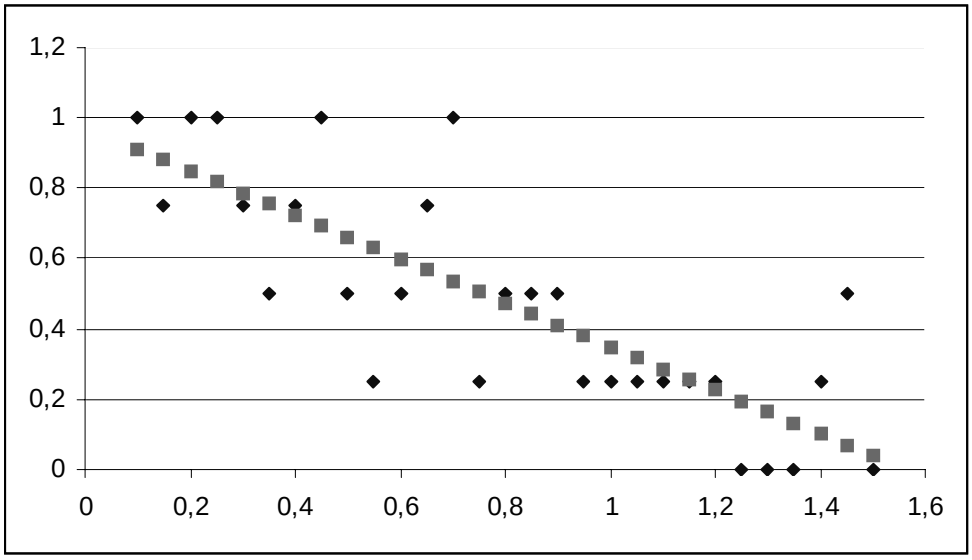

Рис. 3. Пряма регресії (поточково) та розкидані точки в координатній площині "відстань–ймовірність"

Тепер можем знайти значення *p* відносно *d*. Так, якщо *d*=0,77, то *p*=0,49.

Зауваження. Для прикладу вибрано випадок, коли значення d змінюється з кроком 0,05. Очевидно, що на практиці таких випадків не буває, однак це не зменшує загальності наведеної тут теорії.

Результати порівняння анотацій наукових статей

Вихідними даними для системи є файли в електронному вигляді із англомовними анотаціями наукових публікацій, отримані з локальної метрежі (архіви наукових журналів інституту) та з серверів on-line журналів у мережі Інтернет. Розроблювана система орієнтована на аналіз англійської мови, яка сьогодні є найпоширенішою в Інтернеті та часто застосовується науковцями усього світу для розуміння та простоти як друга мова при написанні анотацій статтей. Використання анотацій як запиту-прототипу для пошуку інформації викликано такими причинами:

- анотація містить найінформативніші терміни, які достатньо повно характеризують проблематику статті;
- відсутні формули, символи, рисунки, які важко аналізувати;
- безкоштовно доступні, на відміну від повнотекстових статтей;
- як правило, анотації статтей перекладені англійською мовою.

Проведено такий експеримент з анотаціями наукових публікацій в області штучного інтелекту, який засвідчив, що запропонований у цій роботі підхід на основі адаптивної онтології підвищує точність пошуку документів у середньому на 20%.

Для цього з ключових слів анонотації-взірця сформовано запит в мережу Інтернет. У результаті отримано 25 анотацій з сайтів наукових публікацій. За трьома методами: методом концептуальних графів (Монтеза–Гомеса), коефіцієнтом Дайса (варіант векторно-просторової моделі) та методом,

розробленим у цій роботі, проводився порівняльний аналіз отриманих анотацій із взірцевою. Ефективність цих методів для інформаційного пошуку оцінено за параметром “точність пошуку”.

$$точність = \frac{кількість\_знайдених\_релевалентних(експерт)}{кількість\_усіх\_знайдених(програма)}$$

Отримані результати наведено в табл. 2.

Таблиця 2

Результати порівняння методів

Методи	Точність
Метод за коефіцієнтом Дайса	10/15=0,66, (66%)
Метод Монтеса–Гомеса	9/12=0,75, (75%)
Метод зважених концептуальних графів (розроблений)	11/12=0,916, (92%)

Метод порівняння за Дайсом у 40% випадків визначав найбільш подібними до взірця ті анотації, що мали найбільшу кількість спільних слів, при цьому найменше відповідали прототипу за змістом. Метод концептуальних графів, враховуючи лише кількість спільних зв’язків, також не дав задовільного результату. Водночас врахування апріорної інформації про предметну область через зважування вершин та зв’язків концептуальних графів еталонної та досліджуваної анотації дало змогу виділити найвідповідніші до взірця анотації.

Цей експеримент ілюструє ефективність застосування розробленого в роботі підходу для автоматизації пошуку документів, котрі найбільше відповідають запиту-прототипу і може бути застосований для побудови інтелектуальних метапошукових систем.

Висновки

Отже, розглянуто метод оцінювання подібності документів на основі знаходження відстані між концептуальними графами, що є відповідними моделями цих документів. Для цього введено поняття центру ваг концептуального графу, 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. Для переведення здайденої відстані в ймовірнісну оцінку використано регресійний аналіз. Порівняння розробленого методу з іншими (метод за коефіцієнтом Дайса, метод Монтеса–Гомеса) показує, що результати, отримані запропонованим методом, є на порядок кращими.

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. Sowa J.F. “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. – 2007. Вип. 1, Т. 6. 7. Седжвик Р. Фундаментальные алгоритмы на C++. Алгоритмы на графах: Пер. с англ./Роберт Седжвик. – СПб: ООО "ДиаСофтЮП", 2002. – 496 с.