УДК 004.85

Р.Р. Даревич
Фізико-механічний інститут НАН України, м. Львів, Україна, darevych@ipm.lviv.ua

Підвищення ефективності інтелектуального аналізу тексту шляхом зважування понять у моделі онтології

Розроблено модель онтології у вигляді зваженого орієнтованого графа. Запропоновано новий метод зважування класів онтології бази знань, який полягає у рекурсивному сумуванні ваги їх підкласів та інших семантично пов'язаних об'єктів. Показано, що метод задовольняє вимогам метрики при порівнянні текстових документів, поданих у вигляді зважених концептуальних графів та доповнених контекстом. Результати чисельного моделювання процедури ранжування текстових документів, з врахуванням контексту і відповідної семантики вжитих в них термінів та словосполучень, показали, що даний підхід підвищує ефективність інтелектуального аналізу текстів.

Вступ

У даний час постійно зростає інтерес до застосування інтелектуальних систем (ІС) у різних галузях, таких як: інформаційні технології, машинобудування, медицина, біологія, екологія, географія, юриспруденція тощо [1-3]. В основі архітектури сучасних ІС використовуються бази знань (БЗ), котрі формуються відповідно до предметної області (ПО), в якій застосовується дана ІС. Основною частиною бази знань є онтологія як чітко структурована модель ПО, систематизований набір термінів, котрі пояснюють, в яких відношеннях можуть знаходитись об'єкти предметної області. Онтології є загальновизнані та широко застосовувані в таких різних галузях науки, як інженерія знань (knowledge engineering), представлення знань (knowledge representation), інформаційний пошук (information retrieval), управління знаннями (knowledge management), проектування баз даних (database design), інформаційне моделювання (information modeling) та об'єктно-орієнтований аналіз (object-oriented analysis). Зокрема, у квітні 2004 року фірма Gartner, яка займається дослідженням ринку IT-технологій, віднесла використання таксономії/онтології на третє місце в десятці передових технологій у даній галузі, передбачених на 2005 рік [4].

Постановка задачі

Складна структура взаємозв'язків між поняттями, представлених у базі знань інтелектуальної системи, а також її динамічне наповнення у процесі експлуатації вимагає застосування певних оптимізаційних процедур з метою мінімізації часу реакції на запити, місця, необхідного для розміщення бази знань, а також вирішення конфліктів між даними, внесеними з різних джерел. Для вирішення цієї задачі можна використати апарат теорії графів. Відповідно такі процедури (механізми) оптимізації (усунення конфліктів, збереження цілісності, дотримання обмежень на максимальний об'єм) бази знань полягають в ітераційній редукції її графа до заданого рівня та певної стандартної форми при максимізації суми ваг його вузлів та ребер відповідних семантичних типів. Одним з підходів до реалізації таких механізмів є

автоматичне зважування концептів (понять) та семантичних зв'язків між ними у БЗ. Вага понять має відповідати наступним основним вимогам: забезпечувати оптимізацію змісту БЗ за наперед визначеними критеріями та контроль цілісності БЗ; відповідати вимогам метрики при їх використанні для порівняння семантичної близькості понять.

Таким чином, метою роботи є розробка методу присвоєння вагових коефіцієнтів поняттям у моделі онтології, за допомогою якого можна було б реалізовувати механізми оптимізації структури бази знань. При цьому підвищиться ефективність інтелектуального аналізу тексту, його класифікації, порівняння даного тексту з іншими джерелами інформації та визначення їх подібності.

Існуючі підходи до семантичного зважування онтологій БЗ

За останні декілька років спостерігається підвищення уваги фахівців в області інформаційного пошуку та інженерії знань до об'єктної парадигми, що базується на фреймовій моделі представлення знань та моделі семантичних мереж. Таке представлення дозволяє формувати зважений, орієнтований граф [5]. Доцільність його використання забезпечується високою ефективністю процедур семантичного аналізу таких структур, а також існуванням та широким використанням відповідних стандартів представлення даних зі складною ієрархічною структурою (XML, RDF, DAML+OIL, OWL, KIF, CycL, CGs). При цьому застосування механізму оцінки семантичної ваги знань суттєво покращує результати порівняння текстових документів за їх подібністю до запиту або до деякого еталонного документа [6].

Онтологія використовується для визначення подібності між атомарними та складеними поняттями, які утворюють метазнання [7]. Таксономічна структура онтології БЗ представляється зваженим орієнтованим графом, зв'язки якого мають парну структуру і кожному типу зв'язку присвоєні певні коефіцієнти подібності. Наприклад, для зв'язку типу спеціалізації («*is-a*») він дорівнює 0.9, для узагальнення («*kind-of*») – 0.4, для причинного зв'язку («caused-by») – 0.3. Цей підхід створює умови для семантичного порівняння подібності різних понять онтології шляхом знаходження добутку усіх коефіцієнтів зв'язків між ними. Він знайшов своє застосування у конструюванні запитів на основі онтології, проте одним із суттєвих його недоліків є постійність значення ваги зв'язків між поняттями і, відповідно, відсутності механізмів адаптації системи до предметної області у процесі її експлуатації.

Інший підхід передбачає зважування семантичних зв'язків і застосовується для автоматичного поділу великих онтологій на менші модулі на основі структури ієрархії класів [8]. Визначення сили залежності між поняттями грунтується на теорії соціальної мережі шляхом обчислення пропорційної сили мережі для графа. Пропорційна сила між двома вершинами описує важливість з'єднання одної вершини з рештою на основі числа наявних у вершині зв'язків. У роботі [9] за допомогою онтології автоматично створюються профілі, що дозволяє ефективніше відображати інформаційні інтереси користувача. Профіль користувача представлений зваженою ієрархією понять на основі векторів ключових слів.

Отже, існує ряд підходів для семантичного зважування зв'язків між поняттями в онтологіях, проте у них відсутні процедури зважування самих понять онтології, що є основним їх недоліком. Тому присвоєння вагових коефіцієнтів поняттям онтології,

розглядаючи їх як ресурс онтології, дозволяє визначати інформаційну цінність досліджуваних текстових документів, а також у процесі роботи та самонавчання БЗ здійснювати редукцію її надлишкових елементів.

Модель онтології бази знань

Подамо онтологію бази знань у вигляді графа, вершинам і ребрам якого присвоєно певні числові та семантичні характеристики. Він є орієнтованим зваженим мультиграфом, в якому на етапі формування бази знань допускається існування паралельних ребер, циклів, петель, дублювання вершин з аналогічними параметрами та інших особливостей.

Будуючи модель онтології, врахуємо інформаційну вагу понять та зв'язків шляхом присвоєння їм відповідних вагових коефіцієнтів. Таким чином, модель онтології подамо як зважений орієнтований граф у вигляді четвірки:

$$G(C,R,W_C,W_R),$$

де C – скінченна множина вершин, представляє зважені атомарні поняття;
$R \subseteq C \times C$ – множина дуг, семантичні зв'язки, зважені відповідно до сили зв'язку;
W_C – вага вершини;
W_R – вага зв'язку.

Визначення вагового коефіцієнта поняття в моделі онтології являє собою рекурсивну процедуру. Загальна вага W_j^i класу онтології дорівнює сумі ваги його підкласів W_k^{i-1}, помножених на величину зв'язку $L_{k,j}$ між цим класом та підкласами:

$$W_j^i = \sum_k W_k^{i-1} \cdot L_{k,j},$$

де W_k^{i-1} – вага підкласу елемента онтології;
$L_{k,j}$ – величина зв'язку між класом та його підкласом;
i – рівень; j – елемент i-го рівня; k – кількість елементів на i-му рівні.

При встановленні зв'язку між поняттями його величина визначається співвідношенням ваги понять, які пов'язуються, а перерахунок ваги по ієрархії структури відбувається у напрямку знизу вгору.

На основі розробленої моделі та процедури зважування пропонується метод семантичного порівняння текстових документів за їх подібністю до взірцевого документа, що передбачає формування інформаційного портрета досліджуваного тексту в термінах базової онтології. Пропонується доповнювати концептуальний граф тексту семантичними зв'язками з даної онтології, завдяки чому при порівнянні текстових документів також ураховується їх зміст [6]. Розпізнавання на першому етапі полягає у «впізнаванні» понять та тверджень цього документа у базі знань ІС. Множина впізнаних понять доповнюється з онтології БЗ усіма поняттями, пов'язаними з даними узагальнюючим зв'язком «is-a» аж до кореневого/спільного для всіх. Таке доповнення забезпечує розпізнаний текст понятійним контекстом.

Визначення відстані між двома графами та перевірка на відповідність вимогам метрики

Для порівняння двох графів G_1 та G_2, що представляють відповідні документи, визначимо відстань між ними [10]:

$$d(G_1,G_2)=W(G_1 \Delta G_2),$$

де $d(G_1,G_2)$ – відстань між графами G_1 та G_2 відповідно;
$W(G) = \sum_S \sum_j W_{S,j}$, $G_1 \Delta G_2$ – симетрична різниця двох графів,

$$G_1 \Delta G_2 = (G_1 \cup G_2)/(G_1 \cap G_2).$$

Визначена таким чином відстань задовольняє трьом властивостям метрики. Дійсно, якщо $G_1 = G_2$, то $G_1 \Delta G_2 = \emptyset$, $W(G_1 \Delta G_2) = 0$, то $d(G_1,G_2) = 0$.
Оскільки $G_1 \Delta G_2 = G_2 \Delta G_1$, то $d(G_1,G_2) = d(G_2,G_1)$.

Покажемо, що
$d(G_1,G_2) + d(G_2,G_3) \geq d(G_1,G_3)$, тобто $d(G_1,G_2) + d(G_2,G_3) - d(G_1,G_3) \geq 0$.

Тоді, на основі використання карти Карно перетину трьох графів (рис.1), запишемо:

$$W(G_1 \Delta G_2) + W(G_2 \Delta G_3) - W(G_1 \Delta G_3) = = x+y+v+w+y+z+u+v-x-z-u-w = 2y+2v \geq 0$$

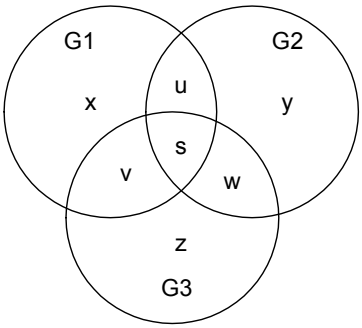


Рисунок 1 – Карта Карно перетину трьох графів

Якщо $\check{G}$ – граф, що відображає онтологію деякої ПО, то будь-яка відстань між документами даної ПО знаходиться в межах $0 \leq d \leq W(\check{G})$. Очевидно, що чим більше документи не схожі, тим більша відстань між ними. Відстань між будь-яким документом, який задається разом зі своїм доповненням графом $\hat{G}$ та онтологією, теж можна обчислювати за формулою: $d(\hat{G},\check{G}) = W(\hat{G} \backslash \check{G})$, оскільки $\check{G} \cup \hat{G} = \check{G}$, $\check{G} \cap \hat{G} = \hat{G}$.

Визначену таким чином метрику можна використовувати для ранжування текстових документів тощо.

Чисельне моделювання методу семантичного ранжування текстових документів

Подібність досліджуваних документів, представлених у вигляді концептуальних графів, до взірцевого промоделюємо на наступному прикладі (табл. 1). Нехай P – речення-прототип, задане користувачем. A, B – речення з документів, знайдених пошуковою системою Google. Порівняємо результати моделювання подібності речень на основі аналізу їх графів.

Семантичну структуру речень визначимо за допомогою аналізатора LinkParser. За його допомогою будується діаграма зв'язків речення (рис. 2). Таке представлення дозволяє встановити всі семантичні зв'язки між словами. Наприклад, речення «*The noble metal gold exists in pure kind and is more expensive than stainless steel.*» після опрацювання аналізатором матиме наступний вигляд:

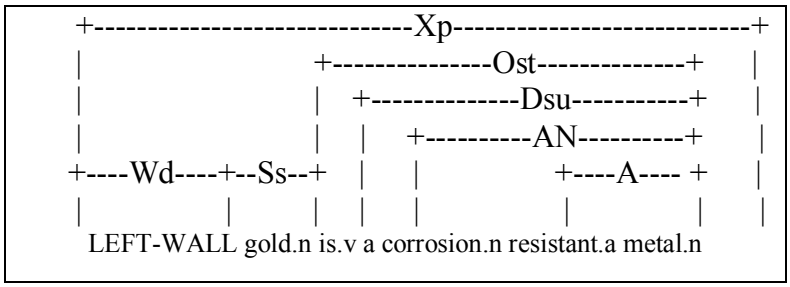
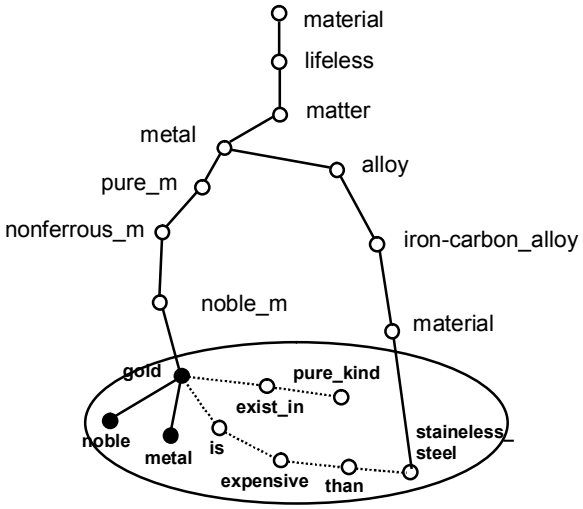
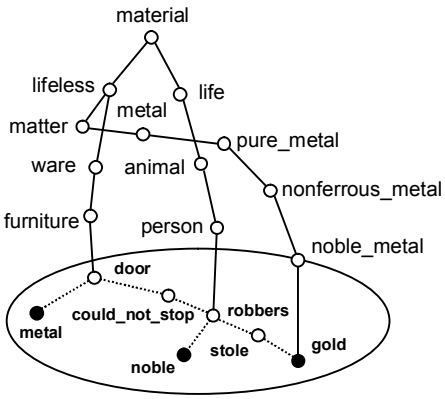
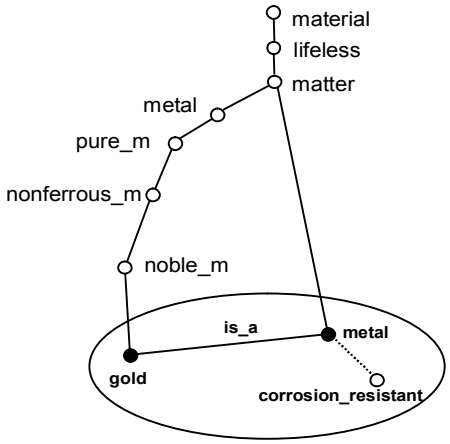


Рисунок 2 – Діаграма зв'язків речення «*The noble metal gold exists in pure kind and is more expensive than stainless steel.*»

Таблиця 1 – Моделювання подібності речень на основі аналізу їх графів

<table>
<tr><td colspan="2">Граф речення прототипу та порівнювані з ним графи двох довільних речень</td></tr>
<tr><td> 
</td><td>P: The noble metal gold exists in pure kind and is more expensive than stainless steel.</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>A: The metal door could not stop a noble robbers who stole the gold.</td><td>B: Gold is a corrosion resistant metal.</td></tr>
</table>

Наведений вище підхід до визначення подібності речень із використанням аналізу семантичних зв'язків між поняттями і врахуванням ваги цих понять при аналізі графа проілюстровано на прикладі трьох речень. Як речення-прототип, з котрим порівнюються інші, було використано речення «*P*» (табл. 1). Для нього було побудовано концептуальний граф і визначено ваги понять у ньому згідно з онтологією даної області знань. Також було побудовано відповідні графи для двох інших речень із наступним розпізнаванням понять, котрі співпали з реченням-прототипом. Для цих понять було встановлено семантичні зв'язки та їх ваги згідно з онтологією. Отримані концептуальні графи були доповнені вертикальними семантичними зв'язками з даної онтології, унаслідок чого при їх аналізі враховується зміст тексту. При аналізі отриманих графів, використовуючи вищенаведений підхід, було встановлено, що речення «*A*» є найменш подібним до прототипу, оскільки в ньому відсутні прямі зв'язки (суцільні лінії) між співпалими поняттями, і його подібність визначається тільки на основі частоти зустрічання цих понять. Тому коефіцієнт подібності цього речення до прототипу становить 15 %. На відміну від речення «*A*» у реченні «*B*» співпалих понять є менше, проте існує прямий зв'язок між ними (забрати «is» з графа), і тому воно є більш подібним до речення-прототипу. При числовому аналізі цього графа встановлено коефіцієнт його подібності 78 %. Отже, використання даного підходу дозволяє аналізувати і порівнювати текстові документи, враховуючи не тільки частоту зустрічання понять в них, а і їх зміст.

Висновки

Запропонований метод зважування класів онтології бази знань дозволяє підвищити ефективність інтелектуального аналізу текстів. При порівнянні текстових документів враховується не тільки частота зустрічання понять в них, а і семантичні зв'язки між ними. Використовуючи даний підхід, враховується контекст документів і відповідна до контексту семантика вжитих у них термінів та словосполучень. Це дає можливість здійснювати автоматичний пошук документів, котрі найбільше відповідають запиту-прототипу, і відкидати такі, що мають малу вагу і не відповідають предметній області.

Перспективи подальшої роботи полягають у можливості здійснення оптимізації змісту та структури бази знань завдяки відповідно розробленій моделі онтології БЗ та методу обчислення інформаційної ваги понять у ній. Наступним кроком досліджень у цьому напрямку буде створення базових алгоритмів та процедур редукції бази знань за наперед заданими критеріями.

Література

1. Dalkilic M., Costello J. BioKnOT: Biological knowledge through ontologies andTFIDF. In Proceedings, Workshop on Search and Discovery in Bioinformatics, SIGIR-Bio, 2004.
2. Léger A., Presidís A., Kervella Ph. OntoWeb: Ontology-based information exchange knowledge management and electronic commerce. IST-2000-29243 Date: 28Th February 2003.
3. De-Carvalho M., Domingue J., Pertusson H. «Alice: An ontology based architecture for supporting online shopping», K-CAP 2001 - Workshop Knowledge in E-Business. – Canada: Victoria, 2001 // http://kmi.open.ac.uk/projects/alice.
4. Michael Denny. Ontology Tools Survey, Revisited, 2004. XML.com // http://www.xml.com/pub/a/2004/07/14/onto.html.

5. Sengupta A., Dalkilic M., Costello J. «Semantic Thumbnails a Novel Method for Summarizing Document Collections» // Proc. 22nd ACM Annual International Conf. on Design of Communication (SIGDOC 2004). – Memphis (TN. USA). – 2004.
6. Досин Д.Г., Даревич Р.Р. Метод визначення подібності текстів, представлених у вигляді зважених концептуальних графів // Відбір та обробка інформації. – 2004. – Вип. 21 (97).
7. Henrik Bulskov, Rasmus Knappe, Troels Andreasen. On Querying Ontologies and Databases. FQAS 2004.
8. Stuckenschmidt H., Klein M. Towards automatic partitioning of class hierarchies // Proc. of the 1st International Conf. on Knowledge Management and Decision Support (ICKMDS'04). – Porto (Portugal). – 2004.
9. Trajkova J., Gauch S. «Improving Ontology-Based User Profiles» // Proc. of RIAO 2004, University of Avignon (Vaucluse). – France. – 2004.
10. Даревич Р.Р., Досин Д.Г., Литвин В.В. Метод автоматичного визначення інформаційної ваги понять в онтології бази знань // Відбір та обробка інформації. – 2005. – Вип. 22 (98).

Р.Р. Даревыч

Повышение эффективности интеллектуального анализа текста путем взвешивания понятий в модели онтологии

Разработана модель онтологии в виде взвешенного ориентированного графа. Предложен новый метод взвешивания классов онтологии базы знаний, который заключается в рекурсивном суммировании веса их подклассов и других семантически связанных объектов. Показано, что метод удовлетворяет требованиям метрики при сравнении текстовых документов, представленных в виде взвешенных концептуальных графов, дополненных контекстом. Результаты численного моделирования процедуры ранжирования текстовых документов, с учетом контекста и соответствующей семантики употребленных в них понятий и словосочетаний, показали, что данный подход повышает эффективность интеллектуального анализа текстов.

R.R. Darevych

Raising the Efficiency of Intellectual Text Analysis by Weighing of Concepts in the Ontology Model

The ontology model as weighed oriented graph is developed. The new method of the weighing of knowledge base ontology classes consists in recursive summing of the weight of their subclasses and other semantic-linked objects is offered. It is shown that method fulfills the properties of a metrics at the text documents comparison represented as weighted conceptual graphs and complemented by context. The results of the numeral modeling of the procedure of text documents ranking, taking into account a context and proper semantics of the terms and predicates showed that the given approach raise efficiency of intellectual texts analysis.

Статья поступила в редакцию 12.07.2005.