УДК 004.89

Т.І. ЧЕРНА

ВИКОРИСТАННЯ АДАПТИВНИХ ОНТОЛОГІЙ В ЗАДАЧІ КВАЗІРЕФЕРУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ

Розглядається підхід до розроблення інтелектуальних систем з використанням онтологій у складі їх баз знань. Здійснюється класифікація таких систем з погляду їх функціонування на основі онтологій. Вводиться поняття адаптивної онтології. Модель адаптивної онтології визначається як розвиток класичної моделі додаванням ваг важливості понять та відношень, які зберігаються в онтології. Показується, як такі онтології можна використати в задачі квазіреферування текстових документів.

1. Вступ

Сучасні дослідження побудови інтелектуальних систем (ІнС) ведуться у двох напрямах: 1) ІнС класифікації (виведення за прецедентами, англ. Case-Based Reasoning); 2) ІнС планування діяльності (пошук стану мети у просторі станів) [1-4].

Вибір ІнС залежить від типу задачі. Метод виведення за прецедентами ефективний, коли основним джерелом знань про задачу є досвід, а не теорія; рішення не є унікальними для конкретної ситуації, а можуть використовуватись в інших випадках; мета розв'язування задачі - отримати не гарантований правильний розв'язок, а найкращий серед можливих. Виведення, основане на прецедентах, є методом побудови ІнС, які приймають рішення щодо проблеми або ситуації за результатами пошуку аналогій, що зберігаються в базі класів. З математичного погляду поточна ситуація S належить до класу Class_k серед множини N класів $\mathbf{Class}=\{\mathrm{Class}_1, \mathrm{Class}_2,\ldots, \mathrm{Class}_N\}$, якщо відстань від S до цього класу є найменшою, тобто

$$\mathrm{Class}_k = \arg\min_i \mathrm{d}(\mathrm{Class}_i, \mathrm{S}),\ \mathrm{i} = \overline{1, \mathrm{N}}\ . \qquad (1)$$

ІнС планування діяльності має досягти стану мети. Насамперед потрібно розробити план досягнення цього стану всіма можливими альтернативними способами. Процес планування ґрунтується на принципі декомпозиції. Задача планування ZP містить три складові: множину станів *St*, множину дій *A*, множину станів мети *Goal*, тобто

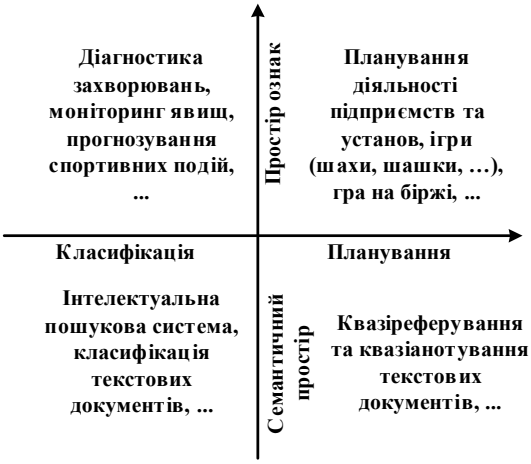
$$\mathrm{ZP} = \langle \mathrm{St}, \mathrm{A}, \mathrm{Goal} \rangle\ . \qquad (2)$$

Для ефективного планування діяльності ІнС повинна вміти оцінювати стани та дії. Як бачимо, для обох типів ІнС необхідна метрика. У першому випадку – для оцінювання близькості класу, у другому – для визначення релевантності станів та дій. Від способу побудови цієї метрики безпосередньо залежить ефективність функціонування ІнС [1].

Проаналізувавши типи задач, для яких використовують онтології, робимо висновок, що всі задачі можна поділити на два підтипи. Перший – тип задач, для яких суттєво, які значення приймають властивості понять. Сюди належать задачі діагностики захворювань, розпізнавання образів, класифікації явищ на підставі зібраних даних тощо. Такі задачі назвемо ознаковими. Для іншого типу задач не є істотним значення понять, швидше – їх семантика або частотність вживання понять у тексті і т.д. До таких задач можна зарахувати кластеризацію інформаційних ресурсів, класифікацію текстів згідно з УДК, інтелектуальні пошукові системи, квазіреферування та квазіанотування текстових документів. Задачі такого типу назвемо семантичними. У результаті отримаємо поділ ІнС за двома вимірами (напрямом розвитку та простором функціонування), як зображено на рис 1. У кожній чверті перераховано задачі, які потрапляють у відповідний тип.

Для ефективного функціонування ІнС необхідно побудувати метрику, на основі якої можна визначати релевантність станів чи класів. Побудова такої метрики прямо залежить від типу задач: семантичні вони чи ознакові. Отже, загалом виділяють чотири різні типи задач, які розв'язують за допомогою ІнС. Зріз за напрямом досліджень потребує двох різних функціональних моделей (пошук класу та планування діяльності), зріз за типом задачі

– використання різного роду метрик для їх розв'язування та оцінювання якості отриманих розв'язків. Розглянемо всі ці типи задач, насамперед ввівши поняття адаптивної онтології (АО) (рисунок).



Типи задач, для розв'язування яких використовують ІСППР

Ефективність адаптації онтології БЗ до особливостей ПО визначають елементи її структури та механізми її адаптації через самонавчання під час експлуатації. Одним з підходів до реалізації таких механізмів є автоматизоване зважування понять БЗ та семантичних зв'язків між ними під час самонавчання. Цю роль виконують ваги важливості понять та зв'язків. Вага важливості поняття (зв'язку) – це числова міра, котра характеризує значущість певного поняття (зв'язку) у конкретній ПО і динамічно змінюється за певними правилами під час експлуатації системи. Запропоновано розширити модель онтології (1), ввівши в її формальний опис ваги важливості понять та відношень [1, 5-7]. Таку онтологію визначено як:

$$\hat{O} = \left\langle \hat{C}, \hat{R}, F \right\rangle, \tag{3}$$

де $\hat{C} = \left\langle C, W \right\rangle$, $\hat{R} = \left\langle R, L \right\rangle$; W – вага важливості понять C; L – вага важливості відношень R.

Визначену у такий спосіб онтологію названо адаптивною, тобто такою, що адаптується до ПО за допомогою задання ваг важливості понять та зв'язків між ними. Така онтологія однозначно подається у вигляді зваженого концептуального графа. Тому метрику побудовано на таких графах.

Переваги моделі (3) полягають у можливості: 1) будувати метрики на основі онтології; 2) адаптувати базу знань ІнС до потреб користувача; 3) задавати важливість знань з точки зору експерта ПО. АО на відміну від звичайної онтології відображає не лише експліцитні (явні) знання, а й імпліцитні (неявні, приховані). Методи інтелектуального аналізу даних (дерева рішень, байєсівські мережі, *k*-найближчих сусідів) є окремим випадком АО залежно від правил задання ваг важливості понять та відношень. З точки зору побудови БЗ ІнС отримуємо такий підхід – експерту або користувачу системи надається готова БЗ, ядром якої є онтологія, а їх задача зводиться лише у налаштуванні цієї БЗ під себе шляхом задання ваг важливості її елементів.

Детальніше розглянемо використання адаптивних онтологій для задач, які знаходяться у 4-й чверті (планування-семантичний простір).

Отже, метою даного дослідження є розроблення ефективного методу автоматизованого квазіреферування природомовного документа. Для досягнення мети пропонується використати адаптивну онтологію предметної області, до якої належить текст, що квазіреферується.

2. Зважування міри TF-IDF

Для задач планування у семантичному просторі про стан мети Goal наперед щось важко сказати. Наприклад, для задачі реферування текстових документів станом мети є кінцевий реферат, однак ми лише можемо собі уявляти, як він приблизно має виглядати. Оцінювання

стану в такій задачі збігається з оцінюванням важливості концепту (слово, лексема, речення), залежно від задачі [8].

Отже, для оцінки станів необхідно використати інший метод. Запропоновано такий метод – зважування міри TF-IDF адаптивною онтологією ПО.

TF-IDF (від англійського TF – term frequency, IDF – inverse document frequency) – статистична міра, що використовується для оцінювання важливості слова в контексті документа. Важливість деякого слова пропорційна кількості його вживання у документі і обернено-пропорційна частоті вживання слова у інших документах колекції. Ця міра часто використовується у задачах аналізу текстів та інформаційного пошуку, наприклад, як один з критеріїв релевантності документа пошуковому запиту, під час розрахунку міри близькості документа, під час кластеризації.

TF (term frequency – частота слова) – відношення числа входження деякого слова до загальної кількості слів документа. Отже, оцінюється важливість слова a_i в межах окремого документа: $TF = \frac{n_i}{\sum_k n_k}$, де n_i – кількість вживання слова у документі, а у знаменнику – загальна кількість слів у цьому документі. IDF (inverse document frequency – зворотна частота документа) – інверсія частоти, з якою деяке слово зустрічається у документах колекції. Врахування IDF зменшує важливість широковживаних слів: $IDF = \log \frac{|T|}{|T_j \supset a_i|}$, де $|T|$ – кількість текстових документів у колекції; $|T_j \supset a_i|$ – кількість текстових документів, в яких зустрічається слово a_i (коли $n_i \neq 0$).

Отже, міра TF-IDF є добутком двох множників: TF і IDF: $TF\text{-}IDF = TF \cdot IDF$.

Більшу важливість у TF-IDF отримають слова з високою частотою у межах конкретного документа і з низькою частотою вживання в інших документах.

Для оцінювання станів зважуватимемо міру TF-IDF важливістю понять, відображених в адаптивній онтології, тобто $v(St) = (TF\text{-}IDF) \cdot W$.

Таке оцінювання містить істотні переваги порівняно з іншими оцінюваннями, оскільки у ній одночасно враховується як частотний аналіз зустрічання термінів у тексті (TF-IDF), так і специфіка ПО, до якої належить тематика цього тексту.

3. Процес квазіреферування

Перероблення інформації, яку подано у вигляді текстів природною мовою, має багато аспектів. Сюди належать такі види інформаційних процесів, як розуміння текстів, їх переклад, стиснення семантичної інформації. Особливе значення має останній тип перероблення; сюди входять класифікація і індексування документів, їх анотування та реферування.

Завдання автоматизації реферування текстової інформації сьогодні залишається дуже актуальним, незважаючи на величезну кількість робіт, що зроблені за останні роки в цьому напрямі. Це зумовлено, насамперед, необхідністю в умовах постійного зростання інформації ознайомлювати спеціалістів та інших зацікавлених людей з необхідними їм документами, поданими стисло, але із збереженням їх змісту. Крім того, анотування й реферування є невід'ємною частиною сучасного видавничого процесу. Будь-яке видання, чи це монографія, підручник, аналітичний огляд тощо, завжди випереджуються вторинним документом (рефератом або анотацією). Реферування використовується не тільки для економії часу під час ознайомлення з великою кількістю джерел, але й з метою пришвидшення повнотекстового пошуку по множині документів, оскільки обсяг реферату у декілька разів менший, ніж обсяг вхідного документа чи їх множини.

Реферування - це одержання найважливішої інформації з одного або декількох джерел для складання їхньої скороченої версії для потреб певних користувачів або задач [9].

Реферат - це семантично адекватний виклад основного змісту первинного документа, що відрізняється ощадливим знаковим оформленням, сталістю лінгвістичних і структурних характеристик і призначений для виконання різноманітних інформаційно-комунікативних функцій у системі наукової комунікації.

Текст складається з послідовності речень $A_1, A_2, ..., A_k$ та утворює кортеж $T = (A_1, A_2, ..., A_k)$, а речення $A_i, i = \overline{1,k}$ – з послідовності слів $a_{ij}, i = \overline{1,k}, j = \overline{1,n}$, яке, своєю чергою, зображається кортежем $A_i = (a_{i1}, a_{i2}, ..., a_{ik})$. Зміст (семантику) тексту T позначимо $S(T)$.

Реферат (summary) текстового документа T позначимо $\hat{T}$ і визначимо як текст, який містить кортеж $\hat{T} = (\hat{A}_1, \hat{A}_2, ..., \hat{A}_l)$, $l \prec k$ і зберігає семантику початкового текстового документа T. Формально це запишемо як $\text{Summary}: T \to \hat{T}$, $S(T) = S(\hat{T})$.

Побудова реферату залежить від правильного оцінювання понять (ключових слів), словосполучень ПО та вибору на основі їх ключових речень.

Отже, оцінкою речень, що входять у текстовий документ, запропоновано вибрати добуток двох важливостей TF-IDF та важливості термінів W в онтології, що відповідає темі, якій належить запропонований до розгляду документ, тобто $\varphi = (\text{TF-IDF}) \cdot W$.

Така оцінка містить істотні переваги порівняно з іншими оцінками, оскільки у ній одночасно враховується як частотний аналіз вживання термінів у тексті (TF-IDF), так і специфіка ПО, до якої належить тематика цього тексту.

Для відбору речень для квазіреферату за основу взято відомий алгоритм просторового ранжування. Його модифіковано з врахуванням важливостей термінів тематики, які зберігаються в онтології ПО. Цей алгоритм ранжування зв'язних структур є універсальним алгоритмом ранжування об'єктів з врахуванням їх внутрішньої зв'язкової структури. Об'єкти зображені векторами у просторі Евкліда. У цьому разі вважається, що "близькість" двох об'єктів, зображених векторами, можна обчислити, як Евклідову міру або скалярний добуток векторів. Метою алгоритму є впорядкувати об'єкти з врахуванням внутрішніх зв'язків об'єктів між собою. Формально зв'язну структуру об'єктів зображають як деякий зважений граф, вершинами якого є самі об'єкти, а важливостями дуг задаються відстані Евкліда між об'єктами. У разі ранжування речень з метою відбору найзначущих з них для побудови квазіреферату алгоритм виглядатиме так:

1. Задається текст (набір речень) $T = (A_1, A_2, ..., A_k)$, тематика $T = (A_1, A_2, ..., A_k)$, до якої належить цей текст $T \in Th_l$. Згідно з тематикою з онтології ПО вибираються відповідні важливості понять та зв'язків $W_{l1}, W_{l2}, ..., W_{ln}$, $L_{l1}, L_{l2}, ..., L_{lm}$.

2. Вводиться $\varphi: T \to R$ – відображення, яке ставить у відповідність кожній точці A_i, $i = 1,2,..k$ значення рангу φ_i. Ми можемо розглядати φ_i як вектор $\varphi = (\varphi_1, \varphi_2, ..., \varphi_k)^T$.

3. Кожне речення (об'єкт) подають у векторному просторі так: $x_i = (\varphi_{i1}, \varphi_{i2}, ..., \varphi_{in_i})^T$, де $\varphi_{ij} = (\text{TF-IDF})_{ij} \cdot W_{ij}$ – міра відносної важливості терма a_{ij}.

4. Набір речень являє собою зважений граф з матрицею важливостей $X = (x_{ij})$. Для кожної пари x_i та x_j речень обчислюється важливість їх "лексичної близькості" за допомогою стандартної Евклідової міри: $x_{ij} = \dfrac{(x_i, x_j)}{|x_i| \cdot |x_j|}$.

Зауважимо, що діагональні елементи матриці $x_{ii} = 0$, щоб отриманий граф не містив циклів. Потрібно зазначити, що отримана матриця важливостей є симетричною відносно своєї головної діагоналі.

Матриця важливостей піддається симетричній нормалізації $S = D^{-\frac{1}{2}} X D^{\frac{1}{2}}$, де $D = (d_{ij})$ – діагональна матриця, де її діагональні елементи d_{ii} дорівнюють сумі елементів *i*-го рядка

матриці X. Нормалізація матриці необхідна для того, щоб ітеративний алгоритм збігався. Значення φ обчислюється як результат ітеративного процесу: $\bar{\varphi}(t+1) = \alpha \cdot S \cdot \bar{\varphi}(t) + (1-\alpha) \cdot \bar{y}$, де $\bar{y}$ – одиничний вектор.

Згідно з теоремою, наведеною в [10], такий ітеративний процес збігається з φ^*. Отже, φ_i^* – отриманий ранг речення A_i. Алгоритм полягає в поступовому розповсюдженні об'єктами свого рангу на суміжні об'єкти-вершини. Отже, ранг φ^* кожного речення A_i обчислюється не лише з врахуванням "близькості" його до еталонного об'єкта (важливостей тематики Th в онтології O), але й із врахуванням зв'язної структури тексту, тобто ранг "поширюється" по графу з врахуванням важливостей зв'язків структур.

Розроблено систему квазіреферування на основі розробленого методу. Квазіреферування – це перший з двох етапів реферування, який полягає у відборі значущих речень із всього текстового документа. Другий етап полягає у зв'язуванні цих речень (у цій роботі даний етап не розглядається, ця задача належить до компетенції прикладної лінгвістики).

Система шукає у вхідному тексті головне речення і формує квазіреферат з указанням смислових класів. Система використовує морфологічний і гіперсинтаксичний засоби "розуміння" тексту. Перевірка гіпотези здійснювалася на масиві 20 довільно відібраних статей за тематикою інформаційних технологій. Були введені такі якісні характеристики квазірефератів: а) повнота передавання основного змісту документа; б) точність – відсутність у квазірефераті речень, надлишкових для передавання основного змісту документа; в) зв'язність (у звичайному розумінні цього слова). Були також введені такі кількісні оцінки кожної з перелічених характеристик квазірефератів: 1 – дуже погано; 2 – погано; 3 – задовільно; 4 – добре; 5 – відмінно. Квазірефератиоцінював автор, тобто людина, яка знає мову, але не обізнана зі змістом тексту, що реферується. Оцінки виставляли винятково з погляду майбутнього користувача системи, в припущенні, що квазіреферат в ідеалі повинен мати статут самостійного документа, тобто давати користувачеві чітке уявлення про тему вхідного документа, інформувати про його основний зміст, але не містити при цьому надлишкової інформації, відрізняючись тим самим від повного документа. Документи, що опрацьовувалися, були поділені на два класи: (а) які піддаються інтелектуальному реферуванню і (б) які не піддаються інтелектуальному реферуванню (наприклад, таблиця порівнянь швидкостей процесорів).

Обсяг одержаних квазірефератів – від трьох до шести речень; у двох випадках обсяг становив 7 речень: це були документи, котрі не підлягають інтелектуальному реферуванню. Отже, експеримент дав змогу зробити такі висновки. Одержані квазірефератимістять мало надлишкової інформації, а її наявність спричинена переважно помилками, не пов'язаними з якістю нашої моделі. Речення, що входять у квазіреферат, містять, як правило, основну інформацію вхідного тексту, тобто відповідають визначенню головного речення. Кількість головних речень, як правило, становить не більше 25 % всіх речень цього тексту: коефіцієнт стиснення, менший, ніж 4, одержаний тільки для дуже коротких текстів. Припущення про те, що з головних речень може бути складений новий текст, який має власну гіперсинтаксичну структуру, частково спростовують результати експерименту: 3 реферати з 20 одержали низьку оцінку за параметром "зв'язність", тобто ці реферати мають вигляд скоріше штучних об'єднань речень, які належать до однієї теми, ніж до тексту. З іншого боку, основною причиною цього були зовнішні для нашої моделі чинники, тому треба вважати одержаний результат попереднім і таким, що потребує додаткової перевірки.

Висновки

Розроблено метод квазіреферування текстових документів на основі зважування міри TF-IDF вагами важливості елементів адаптивної онтології ПО тематики, до якої належить текстовий документ. Для цього запропоновано зважувати терміни ПО та зв'язки між ними в межах онтології. Відповідне програмне забезпечення, яке реалізує розроблений метод, написано на мові програмування C#. Побудований на основі такого підходу квазіреферат показав задовільну якість. Перспективою подальших досліджень є квазіанотування та повноцінне реферування на основі розробленого квазіреферату шляхом зв'язування речень в єдину семантику за допомогою використання методів прикладної лінгвістики.

Список літератури: 1. *Литвин В.В.* Бази знань інтелектуальних систем підтримки прийняття рішень / В.В.Литвин. Львів: Видавництво Львівської політехніки, 2011. 240 с. **2.** *Інтелектуальні* системи, базовані на онтологіях // *Д.Г. Досин, В.В. Литвин, Ю.В. Нікольський, В.В. Пасічник*. Львів: "Цивілізація", 2009. 414 с. **3.** *Lytvyn V.* Design of intelligent decision support systems using ontological approach / V.Lytvyn // An international quarterly journal on economics in technology, new technologies and modelling processes. Lublin. 2013. Vol. II, No 1. P. 31-38. **4.** *Литвин В.В.* Підхід до побудови інтелектуальних систем підтримки прийняття рішень на основі онтологій // Проблеми програмування : наук. журн. / Національна академія наук України; Інститут програмних систем. Київ, 2013. №4. С. 43-52. **5.** *Литвин В.В.* Метод моделювання процесу підтримки прийняття рішень у конкурентному середовищі / В.В.Литвин, О.В.Оборська, Р.В. Вовнянка // Математичні машини й системи : наук. журн. Київ, 2014. №1. С. 50-57. **6.** *Lytvyn V.* Definition of the semantic metrics on the basis of thesaurus of subject area / V.Lytvyn, O.Semotuyk, O.Moroz // An international quarterly journal on economics in technology, new technologies and modelling processes. Lublin. 2013. Vol. II, No 4. P. 47-51. **7.** *Досин Д.Г.* Архітектура інтелектуальної системи інформаційного пошуку в мережі Інтернет/ Д.Г. Досин, В.М. Ковалевич //Штучний інтелект. 2012. №3. С. 241-252. **8.** *Даревич* Р. Р. Оцінка подібності текстових документів на основі визначення інформаційної ваги елементів бази знань / Р. Р. Даревич, Д. Г. Досин, В. В. Литвин, З. Т. Назарчук // Штучний інтелект: наук.-техн. журн. / Національна академія наук України; Інститут проблем штучного інтелекту. Донецьк, 2006. № 3. С. 500–509. **9.** *Белоногов Г.Г.* Компьютерная лингвистика и перспективные информационные технологии / Г.Г. Белоногов, Ю.П. Калинин, А.А. Хорошилов. М.: Русский мир, 2004. 246 с. **10.** *Zhou J.* Ranking on data manifolds / J. Zhou, A. Weston O. Gretton, B. Schlkopf // In Proceedings of NIPS. 2003. P. 234–237.

Надійшла до редколегії 13.02.2014

Черна Тарас Ігорович, аспірант кафедри інформаційних систем Національного університету „Львівська політехніка". Наукові інтереси: побудова інтелектуальних систем. Адреса: Україна, 79000, Львів, вул. С. Бандери, 12, тел. (032) 258-25-38.