Kyung-Jae Bae, Yoon-Seok Jeong, Woo-Sub Shim, Kyoung-Geun Oh, Ji-Hei Kang, Hye-Yeon, Seung-Jin Kwak // Proceedings of the World Library and Information Congress: 72nd IFLA General Conference and Council. Meeting 140: Science and Technology Libraries with Information Technology, 20-24 August 2006 y, Korea, Seoul.- Режим доступу: http://archive.ifla.org/IV/ifla72/papers/140-Bae-en.pdf 9. Rae Julie Breaking New Ground: a virtual global library service to widen access for people with print disabilities/ Julie Rae // Proceedings of the World Library and Information Congress: 75th IFLA General Conference and Council. Meeting 199: Libraries Serving Persons with Print Disabilities, 23-27 August 2009 y, Italy, Milan.-Режим доступу: http://www.ifla.org/files/hq/papers/ifla75/199-rae2-en.pdf. 10. Audiobooks and Access to Information for Canadians with Print Disabilities / Public library services.- Режим доступу: http://www.slais.ubc.ca/COURSES/libr500/06-07-t2/www2/S_LaBelle/toc.htm 11. Fineberg G. NLS pushes conversion to digital books // Libr. of Congr. inform, bull. – 2002. – Vol. 61, N 10. – P. 223 – 225.

УДК 004.89

В.В. Литвин, Н.Б. Шаховська, В.Я. Крайовський
Національний університет "Львівська політехніка"
кафедра інформаційних систем та мереж

РЕФЕРУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ НА ОСНОВІ ЗВАЖУВАННЯ МІРИ TF-IDF ОНТОЛОГІЄЮ ПРЕДМЕТНОЇ ГАЛУЗІ

Розглянуто підхід до автоматизації реферування текстових документів на основі модифікації міри TF-IDF онтологією предметної галузі, до якої належить реферований документ. Розроблено метод реферування, який ґрунтується на такому підході.

Ключові слова: онтологія, реферування, квазіреферат, оцінка якості.

This article discusses an approach to automate summarization of text documents based on the modified TF-IDF measure domain ontology, which is refereed paper. The method of abstracting, which is based on this approach.

Keywords: ontology, concept, abstracting, quality assessment.

Постановка проблеми у загальному вигляді

Переробка інформації, яка подана у вигляді текстів природною мовою, має багато аспектів. Це зокрема такі види інформаційних процесів, як розуміння текстів, їх переклад, стиснення семантичної інформації. Особливе значення має останній тип переробки; зокрема класифікація й індексування документів, їх анотування та реферування.

Задача автоматизації процесу реферування текстової інформації сьогодні залишається дуже актуальною, незважаючи на величезну кількість розробок, які з'явилися за останні роки у цьому напрямі. Це викликано насамперед необхідністю в умовах постійного зростання інформації знайомити спеціалістів та інших зацікавлених людей з необхідними їм документами, представленими в стислому вигляді, 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 супроводжують вторинним документом (рефератом або анотацією). Реферування використовується не тільки для економії часу при ознайомленні 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 процес видобування найважливішої інформації з одного або декількох джерел для складання їхньої скороченої версії для потреб певних користувачів або задач [1, 2].

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.

У нашій роботі розглядається новий метод автоматизованого реферування за допомогою онтологій. Тобто для представлення знань у системах автоматичного реферування (АР) використовуються онтології, які використовуються для оптимізації процедури автоматичного видобування знань із текстів природною мовою [3, 4]. Для розв'язання цієї задачі доцільно створювати декілька онтологій: онтології верхнього рівня і онтології предметних галузей. Онтологія верхнього рівня являє собою вироджену онтологію у вигляді словника метазначень (змістових категорій, характерних для рефератів – об'єкт, результат, ціль, засіб). Словник цих категорій визначається в процесі побудови моделі реферату.

Побудова онтології предметної галузі передбачала видобування термінів із текстів і розподіл їх за категоріями онтології верхнього рівня, на основі якого будувалась концептуальна модель предметної галузі у вигляді таксономії понять певної галузі знань. Як мову опису онтології використовують OWL, в якому під онтологією розуміється сукупність тверджень, які задають відношення між поняттями та ті, які визначають логічні правила для суджень про них.

Поряд з онтологіями для роботи системи АР потрібна текстова база знань. Вона складається із фактів і тверджень, пов'язаних із певною ситуацією (конкретним текстом). На відміну від онтології, яка містить не залежну від ситуації інформацію, являє собою „інформаційне ядро", яке містить інформацію, яка залежить від ситуації. Для побудови текстової бази знань ми відштовхувались від понять, які містяться в заголовку документа, згідно з яким відшукуються відповідні їм іменні групи в тексті. У результаті зіставлення термінів з текстової бази знань з даними онтологіями формується набір понять, які необхідні для змістовного конструювання реферату, тобто формуються ланцюжки іменних груп для реферативних конструкцій.

Аналіз останніх досліджень та публікацій

Проблема побудови реферату залежить від правильної оцінки понять (ключових слів), словосполучень предметної галузі та вибору на основі їх ключових речень. Коефіцієнт важливості поняття (зв'язку) – це числова міра котра характеризує значимість цього поняття (зв'язку) у конкретній предметній галузі (ПО) і змінюється за визначеним алгоритмом (певними правилами) під час опрацювання текстових документів. Такий алгоритм належить розробити під час побудови моделі.

Відомими методами визначення значущості речень є оцінка, запропонована Г. Луном, гіпотеза В. Пурто, оцінка на основі міри TF-IDF [5, 6].

Тобто реферат буде тим кращий, чим точнішими будуть оцінки інформаційної значущості речення j для відбору речень та інформаційної новизни y для відсікання подібних речень, якщо реферат будується на основі колекції (множини) текстових документів.

Оцінка Луна. Одна з перших систем автоматизованого квазіреферування ґрунтувалася на ідеї, що для кожного документа специфічні слова, які часто зустрічаються в ньому, використовуються для передачі основної ідеї, яка викладена у тексті. Використовувалась така оцінка значущості кожного речення, що складають документ:

$$V_r = \frac{N_{z,s}^2}{N_s},$$

де V – значущість речення; $N_{z,s}$ – число значущих слів у цьому реченні, тобто таких слів, які є специфічними для ПО, до якої належать документ, і для самого цього документа; N_s – загальне число слів у реченні.

За такою методикою квазіреферат виглядає як сукупність розрізнених фраз, що зрозуміти зміст реферату можна тільки після додаткового опрацювання отриманого тексту людиною.

Гіпотеза Пурто. Інша методика оцінки семантичної значущості речень для відбору їх у квазіреферат ґрунтується на визначенні кількості інформації, яка міститься у кожному з них. Для цього проводять частотний аналіз тексту з погляду зустрічання в ньому важливих термінів. За гіпотезою автора цієї методики В. Пурто, чим важливішим є для деякого тексту той чи інший термін, тим частіше він зустрічається в ньому. Тому для квазіреферату відбираються такі речення, які містять найбільшу кількість термінів, які найчастіше повторюються у цьому документі.

Міра TF-IDF. TF-IDF (від англійського TF – term frequency, IDF – inverse document frequency) – статистична міра, що використовується для оцінювання важливості слова в контексті документа. Вага деякого слова пропорційна кількості вживання цього слова у документі, і оберненопропорційна 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. Так оцінюється важливість слова a_i у межах окремого документа:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k},$$

де n_i – число вживання слова у документі, а у знаменнику – загальна кількість слів у даному документі.

IDF (inverse document frequency – зворотна частота документа) – інверсія частоти, з якою деяке слово зустрічається у документах колекції. Врахування IDF зменшує вагу широковживаних слів:

$$\mathrm{IDF} = \log \frac{|T|}{|T_j \supset a_i|},$$

де $|T|$ – кількість текстових документів в колекції; $|T_j \supset a_i|$ – кількість текстових документів, в яких зустрічається слово a_i (коли $n_i \neq 0$).

Тобто, міра TF–IDF є добутком двох множників: TF і IDF. Більшу вагу у TF–IDF отримають слова з високою частотою у межах конкретного документа і з низькою частотою вживання в інших документах.

Існують різні формули, що ґрунтуються на методі TF–IDF. Вони відрізняються коефіцієнтами, нормуванням, використанням логарифмічних шкал. Так пошукова система Яндекс протягом довгого часу використовувала нормування за самим частотним терміном в документі [5].

Міра TF–IDF часто використовується для подання документів колекції у вигляді числових векторів, які відображають важливість використання кожного слова з деякої множини слів (кількість слів множини визначає розмірність вектора) у кожному документі. Така модель називається векторною моделлю (VSM) і дає можливість порівнювати тексти, порівнюючи їхні вектори в якій-небудь метриці (евклідовий простір, косинусна міра, манхеттенська відстань, відстань Чебишова та ін.).

Формування цілей

Розробити метод реферування текстових документів на основі модифікації міри TF-IDF, який б враховував специфіку предметної галузі, до якої входить реферований документ. Така специфіка відображається в онтологіях.

Основний матеріал

Зважування понять на основі онтологій

Всі вищенаведені методи не враховують специфіки предметної галузі та окремих тем, які можуть в ній бути. Така специфіка враховується в онтологіях. Тому нами пропонується

використати адаптивні онтології, які містять коефіцієнти важливості понять W та зв'язків L [7, 8]. Ці коефіцієнти обчислюються за таким алгоритмом [9]:

1. 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, зв'язаних з даним класом не IS-A зв'язком):

$$W_j^i = Wo_j^i + Ws_j^i + Wn_j^i, \tag{1}$$

де $Ws_j^i = \sum_k Wc_k^{i+1} \cdot L_{j,k}$ − вага k підкласів j-го класу i-го рівня, причому для кореневого класу рівень $i = 0$; $Wc_k^{i+1} = Wo_k^{i+1} + Ws_k^{i+1}$ − вага класу C_k^{i+1}; $L_{j,k}$ − вага зв'язку між класами C_j^i та C_k^{i+1}.

Перерахунок окремих компонент повної ваги класу відображено на рис. 1.

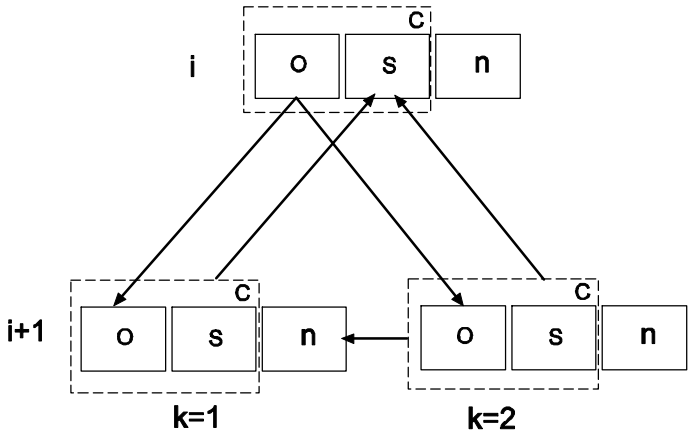

Рис. 1. Схема перерахунку окремих компонент повної ваги класу

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

$$Wo_j^{i+1} = \frac{1}{2} Wo_j^i \tag{2}$$

Вага класу Wc_j^i та усіх батьківських класів аж до кореневого збільшується на величину ваги новоствореного підкласу:

$$Wc_j^m = Wc_j^m + Wo_j^{i+1}, \forall m \le i. \tag{3}$$

3. Під час встановлення зв'язку між поняттями k_1 та k_2 між відповідними вершинами графу онтології 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 так, що:

$$Wn_j = \sum_k Wc_k \cdot L_{j,k} \tag{4}$$

Повторне встановлення зв'язків призводить до появи кратних ребер у графі.

4. Кратність ребер відображає частоту зустрічання F пари семантично пов'язаних понять $L_{i+1} = F \cdot L_i$. Кратні ребра після перерахунку не збільшують валентності вершини.

5. Вага екземпляра БЗ дорівнює повній вазі його класу.

Отже, визначена модель онтології БЗ дає змогу розраховувати вагові коефіцієнти своїх компонентів у процесі їх додавання, вилучення і використання під час експлуатації системи, завдяки чому реалізує механізм адаптації до заданої користувачем ПО [8].

Очевидно, що в межах однієї онтології може описуватись кілька різних тем, що належать до визначеної цією онтологією ПО. Тому коефіцієнти важливості понять та зв'язків залежать від тематики. Нехай онтологія O описує m тем ПО – $\mathrm{Th}_1, \mathrm{Th}_2, ..., \mathrm{Th}_m$, тоді коефіцієнти ваг понять та

зв’язків насправді собою являють вектори, компонентами яких є відповідні значення згідно з темою, тобто $W=(W_1,W_2,...,W_m)$, $L=(L_1,L_2,...,L_m)$. А для процесу автоматизованого реферування наперед треба вибрати тему, до якої належатиме текстовий документ, що опрацьовується, щоб система використовувала правильні ваги.

Визначення основних понять та властивостей графу онтології бази знань

Ієрархічна багатозв'язкова структура семантичної мережі фреймів онтології БЗ інтелектуальної системи може бути подана як орієнтований зважений мультиграф. Графова модель онтології володіє такими властивостями:

1) всі вершини і ребра графу іменовані та зважені;

2) допускається існування паралельних ребер, циклів, петель, дублювання вершин з аналогічними параметрами та інших особливостей;

3) кожна вершина може мати зв'язок з іншими вершинами;

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

Оскільки база знань (БЗ) є семантичною мережею фреймів, в кожній вершині *C* графу мережі *G* міститься деяка множина елементів, що характеризують відповідний цій вершині об’єкт. Ребра графу, які відповідають зв’язкам (твердженням у самій БЗ), визначаються впорядкованими парами вершин <*i*, *j*>. *Шляхом* означимо послідовність дуг (орієнтованих ребер), така, що кінець однієї дуги є початком іншої дуги і використовуватимемо його для пошуку відстані між двома графами. Граф називатимемо *зв'язним*, якщо для довільної пари вершин існує шлях між ними. Зв’язність графу семантичної мережі онтології – властивість, яка означає, що усі елементи мережі знаходяться у межах досяжності інтелектуальної системи і можуть бути задіяні при генеруванні відгуку на звертання до неї.

Опишемо взаємозв’язок між структурою зв’язків онтології та механізмами реалізації міркувань. Модель повинна містити механізми міркувань, якими виступатимуть приєднані процедури фреймів, що використовують встановлені зв’язки (твердження) з метою вироблення необхідного рішення. Згідно з об’єктною парадигмою та фреймовою моделлю подання знань батьківський клас-фрейм містить приєднані процедури встановлення конкретних значень власних слотів-властивостей, та слотів нових екземплярів чи підкласів у процесі їх генерації. Поняття „містить” означає наявність у відповідних слотах фрейма адрес відповідних екземплярів класу приєднаних процедур (обробників подій). Ці приєднані процедури для новоствореного класу чи об’єкта генерує клас приєднаних процедур у відповідь на сигнал від батьківського класу, повертаючи при цьому адресу згенерованих екземплярів-процедур. Отже, кожен екземпляр деякого класу містить лише базову процедуру генерування звертань до інших екземплярів, всі інші процедури розміщені зовні як екземпляри класу процедур, а їх адреси розміщуються у слотах екземпляра, який може викликати дану процедуру. Процедура відгукується на виклик з відомими їй допустимими параметрами, обробляє їх і повертає результат, яким може бути, зокрема, адреса згенерованого цією процедурою нового класу чи екземпляра існуючого класу.

Отже, зв'язки у семантичній мережі фреймів реалізуються через обмін повідомленнями між їх приєднаними процедурами.

Наш підхід до подання знань у формі зваженої семантичної мережі (концептуальних графів) полягає у тому, що будь-яке можливе узагальнення, тобто комплексне, складене поняття завжди явним чином артикульоване, назване і як окреме поняття фігурує в БЗ. Тому якщо деяке узагальнення має спільні властивості чи способи функціонування, вони фізично можуть бути реалізовані через властивості та обробники подій відповідного узагальнюючого поняття, згідно з принципом наслідування [9].

Підходи до семантичного зважування понять онтології

За останні декілька років можна спостерігати посилення уваги фахівців у галузі інформаційного пошуку та інженерії знань до об’єктної парадигми, що ґрунтується на фреймовій

моделі подання знань та моделі семантичних мереж. Доцільність такого використання пов'язана з високою ефективністю процедур семантичного аналізу таких структур, а також існуванням відповідних стандартів відображення даних зі складною ієрархічною структурою (XML, RDF, OWL). При цьому механізм оцінювання семантичної ваги знань покращує результати порівняння текстових документів за їх релевантністю до запиту або до деякого еталонного документа [2].

У роботі [5] онтологію використовують для визначення подібності між атомарними та складеними поняттями, які утворюють метазнання. Автори пропонують подавати таксономічну структуру онтології зваженим орієнтованим графом, зв'язки якого мають парну структуру (якщо існує зв'язок Vi→Vj , то існує також Vj→Vi), а кожному типу зв'язку присвоєні певні коефіцієнти подібності. Наприклад, для зв'язку типу спеціалізації (″IS-A″) – σ=0.9, для узагальнення (″KIND-OF″) – γ=0.4, для причинного зв'язку (″CAUSED-BY″) – ρ_{CBY}=0.3, для характеризуючого зв'язку (″CHARACTERIZED-BY″) – ρ_{WRT}=0.2. Цей підхід створює умови для семантичного порівняння подібності різних понять онтології, оцінюючи шлях семантичних зв'язків між ними, виражений як добуток усіх ланок на цьому шляху. Його застосовують для конструювання запитів на основі онтології, проте одним із істотних його недоліків є постійність значення ваги зв'язків між поняттями і, відповідно, відсутність механізмів адаптації системи до ПО під час її експлуатації.

Ще один підхід, який передбачає зважування семантичних зв'язків, використовують для автоматичного поділу великих онтологій на менші модулі на основі структури ієрархії класів [10]. Тут визначення сили залежності між поняттями ґрунтується на теорії соціальної мережі через обчислення пропорційної сили мережі для залежного графу. Пропорційна сила між двома вершинами описує важливість з'єднання однієї вершини з рештою на основі числа наявних у вершині зв'язків.

У роботі [2] за допомогою онтології автоматично створюють профілі, котрі дають змогу ефективніше відображати інформаційні інтереси користувача. Профіль користувача поданий зваженою ієрархію понять на основі векторів ключових слів.

Узагальнюючи, можна зробити висновок про наявність ряду підходів до семантичного зважування зв'язків між поняттями в онтологіях, проте у всіх цих методиках відсутні процедури автоматичного зважування понять онтології, що є основним їхнім недоліком. На думку автора, статично визначені вагові коефіцієнти понять та зв'язків онтології не забезпечують оцінювання актуальної інформаційної цінності досліджуваних текстових документів. Крім того, недоліком фіксованих вагових коефіцієнтів є неможливість самонавчання системи на основі налаштування її онтології до заданої ПО, а також неможливість здійснювати пошук та вилучення надлишкових елементів онтології за їх семантичною вагою.

Реферування одного текстового документа

Отже, нами в якості оцінки речень, що входять до текстового документа, запропоновано взяти добуток двох ваг TF-IDF та ваги термінів W в онтології, що відповідає темі, якій належить запропонований до розгляду документ. Тобто

$$j = (\text{TF-IDF}) \cdot W . \qquad (5)$$

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

Для відбору речень для квазіреферату за основу нами взято відомий алгоритм просторового ранжування. Він нами модифікований з врахуванням ваг 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 представляються векторами у просторі Евкліда. У цьому випадку вважається, що «близькість» двох об'єктів, представлених векторами, може бути обчислена як Евклідова міра або скалярний добуток векторів. Метою алгоритму є впорядкувати об'єкти з врахуванням внутрішніх зв'язків об'єктів між собою. Формально зв'язна структура об'єктів представляється як деякий зважений граф, вершинами якого

є самі об'єкти, а як ваги дуг задаються відстані Евкліда між об'єктами. У випадку ранжування речень з метою відбору найбільш значущих з них для побудови квазіреферату алгоритм виглядатиме так:

1. Задається текст (набір речень) $T=(A_1,A_2,...,A_k)$, тематика Th_l до якої належить цей текст $T\in \mathrm{Th}_l$. Згідно з тематикою з онтології ПО вибираються відповідні ваги понять та зв'язків $W_{l1},W_{l2},...,W_{\ln}$, $L_{l1},L_{l2},...,L_{lm}$.
2. Вводиться $j:T\to R$ – відображення, яке ставить у відповідність кожній точці A_i, $i=1,2,..k$ значення рангу j_i. Ми можемо розглядати j як вектор $j=(j_1,j_2,...,j_k)^T$.
3. Кожне речення (об'єкт) представляється у векторному просторі таким чином: $x_i=(j_{i1},j_{i2},...,j_{in_i})^T$, де $j_{ij}=(\text{TF-IDF})_{ij}\cdot W_{ij}$ – міра відносної важливості терма a_{ij}.
4. Набір речень являє собою зважений граф з матрицею ваг $X=(x_{ij})$. Для кожної пари x_i та x_j речень обчислюється вага їх «лексичної близькості» за допомогою стандартної Евклідової міри:

$$x_{ij}=Sim(x_i,x_j), \tag{6}$$

де $Sim(x_i,x_j)=\dfrac{(x_i,x_j)}{|x_i|\cdot|x_j|}$.

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

Матриця ваг піддається симетричній нормалізації

$$S=D^{-\frac{1}{2}}XD^{\frac{1}{2}}, \tag{7}$$

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 j обчислюється як результат ітеративного процесу:

$$\bar{j}(t+1)=a\cdot S\cdot\bar{j}(t)+(1-a)\cdot\bar{y}, \tag{8}$$

де $\bar{y}$ – одиничний вектор.

Згідно з теоремою, наведеною в [5], ітеративний процес збігається до j^*. Отже, j_i^* – отриманий ранг речення A_i. Алгоритм полягає в поступовому розповсюджені об'єктами свого рангу на суміжні об'єкти-вершини. Тобто, ранг j^* кожного речення A_i обчислюється не лише з врахуванням «близькості» його до еталонного об'єкта (ваг тематики Th в онтології O), але й із врахуванням зв'язної структури тексту, тобто ранг «поширюється» по графу з врахуванням вагів зв'язків структур.

Якість квазіреферування розробленим методом

Онтологія зберігається в пам'яті комп'ютера і в будь-який момент може бути доповнена новими об'єктами або зв'язками. Для правильного і ефективного функціонування онтології необхідне виконання декількох умов:

1. Реалізація програмної бібліотеки, яка б могла представляти онтологію в пам'яті, зчитувати опис онтології з файла та збереження її у файл;
2. Бібліотека повинна використовувати ефективні алгоритми, швидкодія яких дає змогу працювати із онтологіями великих розмірів;
3. Файли, у яких зберігається опис онтології, повинні бути у форматі, що використовується й іншими програмними засобами для роботи з онтологіями: для забезпечення можливості роботи із сторонніми онтологіями [4].

Ефективним методом вважається встановлення зв’язків із сторонніми онтологіями за допомогою мережі – підхід, що пропагується парадигмою Semantic Web.

Отже, модуль роботи із онтологією повинен складатися із таких підсистем:

- підсистема перетворення онтології з загальноприйнятого формату у внутрішній та навпаки;
- об’єктна модель для представлення сутностей онтології у пам’яті комп’ютера;
- підсистема пошуку об’єктів та зв’язків;
- підсистема додавання об’єктів та зв’язків;
- підсистема контролю цілісності онтології;
- механізм об’єднання онтологій з декількох вихідних файлів та зовнішніх онтологій з мережі;
- інтерфейс для роботи із користувачем, в даному випадку системою автоматизованого реферування.

Оскільки формати, у яких зберігаються та передаються описи онтологій, повинні відповідати певним стандартам, то для роботи з ними їх потрібно опрацювати і подати у форматі, передбаченому архітектурою системи. Забігаючи наперед, скажемо, що буде використовуватися стандарт OWL [5] як один із найпоширеніших і найефективніших, а отже, ця підсистема являтиме собою дещо вдосконалений механізм для роботи з XML.

Об’єктна модель для подання онтології у пам’яті – це бібліотека класів об’єктно-орієнтовною мовою програмування, яка дає змогу повною мірою відобразити всі елементи онтології в пам’яті комп’ютера.

Підсистема пошуку об’єктів та зв’язків – це компонент, що забезпечує навігацію по онтології у відносно швидкому режимі.

Підсистема додавання об’єктів та зв’язків – в онтологію в будь-який момент можна додати об’єкт чи зв'язок. Видалення елементів стандартом не передбачене.

Підсистема контролю цілісності виявляє колізії, що з’являються в онтології, коли додаються нові елементи. Згідно з принципами, що декларуються стандартом OWL, протиріччя, що утворюються при цьому, повинні опрацьовуватися системою-користувачем онтології згідно з підходом, який передбачений принципами її функціонування.

Оцінка якості квазіреферату

Назва файла	Коефіцієнт стиснення	Оцінка повноти	Оцінка точності	Оцінка зв'язності
address munging.doc	3	4	4	3
audioblog.rtf	3	4	4	4
barfmail.txt	4	5	4	4
blog.hml	4	5	4	4
blogosphere.xml	3	5	4	3
clickstream.doc	3	4	5	3
collaboratory.doc	3	4	4	2
e-signature.doc	3	5	5	3
nooksurfer.doc	4	5	4	4

Нами розроблено систему квазіреферування на основі розробленого методу. Система шукає у вхідному тексті головне речення і формує квазіреферат із зазначенням смислових класів. Система використовує морфологічний і гіперсинтаксичний засоби “розуміння” тексту. Перевірка гіпотези здійснювалася на масиві 20 довільно відібраних статей за тематикою інформаційних технологій. Було введено такі якісні характеристики квазірефератів: а) повнота передачі основного змісту документа; б) точність – відсутність у квазірефераті речень, надлишкових для передачі основного змісту документу; в) зв’язність (у звичайному розумінні цього слова). Було також введено такі кількісні оцінки кожної з перелічених характеристик квазірефератів: 1 – дуже погано; 2 – погано;

3 – задовільно; 4 – добре; 5 – відмінно. Квазірефераты оцінювалися автором, тобто людиною, яка знає мову, але не обізнана зі змістом тексту, що реферується. Оцінки виставлялися виключно з погляду майбутнього користувача системи, припускаючи, 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, було поділено на два класи: (а) які піддаються інтелектуальному реферуванню; (б) які не піддаються інтелектуальному реферуванню (наприклад, таблиця порівнянь швидкостей процесорів). Оцінку якості окремих квазірефератів текстів обох класів наведено в таблиці.

Обсяг одержаних квазірефератів – від 3 до 6 речень; у двох випадках обсяг становив 7 речень: це були документи, котрі не підлягають інтелектуальному реферуванню. Отже, експеримент дав змогу зробити такі висновки. Одержані квазірефераты містять мало надлишкової інформації, а її наявність викликана переважно помилками, не пов'язаними з якістю нашої моделі. Включені в квазіреферат речення містять, як правило, основну інформацію вхідного тексту, тобто відповідають визначенню головного речення. Кількість головних речень переважно становить не більше 25% всіх речень цього тексту: коефіцієнт стиснення, менший за 4, одержано тільки для дуже коротких текстів. Припущення про те, що з головних речень може бути складений новий текст, що має власну гіперсинтаксичну структуру, частково спростовується результатами експерименту: 3 рефератів з 20 одержали низьку оцінку за параметром "зв'язність", тобто ці реферати мають вигляд скоріше штучних об'єднань речень, які стосуються однієї теми, ніж 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 модифікації міри TF-IDF. Суть такої модифікації полягає у зважуванні термінів, понять предметної галузі та зв'язків між ними. Зважування відбувається за рахунок ваг важливості концептів предметної області, які зберігаються в її онтології. Побудований на основі такого підходу квазіреферат показав задовільну якість.

1. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии / Г.Г. Белоногов, Ю.П. Калинин, А.А. Хорошилов. – М.: Русский мир, 2004. – 246 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. Литвин В.В. Метод автоматизованого реферування текстових документів з використанням онтологій / В.В. Литвин, В.А. Гайдін, О.Ю. Пшеничний // Складні системи і процеси. – Запоріжжя, 2009. – №1, – С. 81–87. 4. Крайовський В.Я Основні підходи до розроблення программного комплексу автоматичного реферування текстових документів / В.Я. Крайовський, В.В. Литвин, Н.Б. Шаховська // Інститут проблем моделювання в енергетиці. – К., 2009. – Вип. 51. – С.178–186. 5.Zhou J. Ranking on data manifolds / J.Zhou, A.Weston O.Gretton, B.Schlkopf // In Proceedings of NIPS. – 2003. – P. 234–237. 6. Ланде Д.В. Интернетика: Навигация в сложных системах: модели и алгоритмы / Д.В. Ланде, А.А. Снарский, И.В. Безсуднов. – М.: Книжный дом «ЛИБРОКОМ», 2009. – 264 с. 7. Крайовський В.Я. Використання адаптивних онтологій в інтелектуальних системах прийняття рішень / В.Я. Крайовський, В.В. Литвин, Н.Б.Шаховська // Східноєвропейський журнал передових технологій. – Харків, 2009. – №4/3(40). – С.7–12. 8. Литвин В.В. Мультиагентні системи підтримки прийняття рішень, що базуються на прецедентах та використовують адаптивні онтології / В.В. Литвин // Радіоелектроніка, Інформатика, Управління. – Запоріжжя, 2009. – №2(21). – С. 120–126. 9. Даревич Р.Р. Метод автоматичного визначення інформаційної ваги понять в онтології бази знань / Р.Р. Даревич, Д.Г. Досин, В.В. Литвин // Відбір та обробка інформації. – 2005. – Вип. 22(98). – С.105–111. 10.Застосування інформаційних технологій для координації наукових досліджень / Р.Р. Даревич, Д.Г. Досин, В.В. Литвин, Л.С. Мельничок. – Львів: „СПОЛОМ", 2008. – 240 с.