УДК 004.89, 531

В. В. Литвин, Т. І. Черна, В. М. Ковалевич

МЕТОД КВАЗІРЕФЕРУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ НА ОСНОВІ ОНТОЛОГІЇ ПРЕДМЕТНОЇ ОБЛАСТІ

This paper deals with ontological approach to automatic summarization of text documents based on the weighting of TF-IDF measure by weights of concept and relation importance in the domain to which belongs the referenced document. Automatic text summarization method, which is based on this approach is developed.

Keywords: *information technologies, ontology, graph, knowledge base.*

Розглянуто онтологічний підхід до квазіреферування (автоматичне реферування) текстових документів на основі зважування міри TF-IDF вагами важливості понять та відношень предметної області, до якої належить реферований документ. Розроблено метод квазіреферування, який базується на такому підході.

Ключові слова: *інформаційні технології, онтологія, граф, база знань*.

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

Задача автоматизації процесу реферування текстової інформації сьогодні залишається актуальною, незважаючи на величезну кількість робіт, що зроблені за останні роки в цьому напрямі. Це зумовлено передусім необхідністю в умовах постійного зростання інформації ознайомлювати 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 випереджуються вторинним документом (рефератом або анотацією). Реферування використовують 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 [1, 2].

Мета процедури автоматизованого реферування – виділити з тексту документа найважливіші положення, які найповніше розкривають суть цього тексту. Як вхідний матеріал для такого реферату слугують речення, що складають текст документа. У результаті відбору деяких з них отримується скорочений варіант початкового документа, який, строго говорячи, не є рефератом у повному змісті цього слова. Цей скорочений таким чином текст прийнято називати квазірефератом. Тобто квазіреферат – сукупність розрізнених фраз, що зрозуміти зміст реферату можна тільки після додаткового опрацювання отриманого тексту людиною.

Задача опрацювання зв'язного тексту і генерації таких текстів є доволі складною, вона слабо піддається формалізації. Однак розроблено кілька методик, що дають змогу підвищити зв'язність тексту порівняно з простим відбором найбільш значущих речень. Одна з них полягає в тому, що найзв'язанішими вважають такі речення, які містять найбільшу кількість одних і тих ж значущих слів.

У нашій роботі розглядаємо новий метод автоматизованого квазіреферування за допомогою онтологій. Тобто для представлення знань у системах автоматичного квазіреферування (АкР) використовуються онтології, які використовують для оптимізації процедури автоматичного видобування знань із текстів природною мовою [3, 4]. Для розв'язання цієї задачі доцільне створення декількох онтологій: онтології верхнього рівня і онтологій предметних областей. Онтологія верхнього рівня являє собою вироджену онтологію у вигляді словника мета значень (змістових категорій, характерних для рефератів – об'єкт, результат, мета, засіб). Словник цих категорій визначається в процесі побудови моделі реферату.

Побудова онтології предметної області охоплювала видобування термінів із текстів і розподіл їх за категоріями онтології верхнього рівня, на базі якого будувалась концептуальна модель предметної області у вигляді таксономії понять певної області знань. Як мову опису онтології використовували 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 АкР потрібна текстова база знань. Вона складається із фактів і тверджень, пов'язаних із певною ситуацією (конкретним текстом). На відміну від онтології, яка містить незалежну 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. В результаті співставлення 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.

Проблема побудова квазіреферату залежить від правильної оцінки понять (ключових слів), словосполучень предметної області 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 області (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 φ для відбору речень та інформаційної новизни ψ для відсікання подібних речень, якщо реферат будується на основі колекції (множини) текстових документів.

Оцінка Луна. Одна з перших систем автоматизованого квазіреферування базувалася на ідеї, що для кожного документа специфічні слова, які часто трапляються у ньому, використовують для передачі основної ідеї, яка викладена у тексті. Використовували таку оцінку 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 відноситься документ, і для самого

цього документа; N_s – загальне число слів у реченні.

Згідно з 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 базується на визначенні кількості інформації, яка міститься у кожному з них. Для цього проводять частотний аналіз тексту з погляду подання в ньому важливих термінів. Згідно з гіпотезою автора цієї методики В. Пурто, чим важливішим є для деякого тексту той чи інший термін, тим частіше він трапляється в ньому. Тому для квазіреферату відбирають такі речення, які містять найбільшу кількість термінів, яка найчастіше повторюється у цьому документі.

Міра TF-IDF. TF-IDF (від англійського TF – term frequency, IDF – inverse document frequency) – статистична міра, яку використовують для оцінки важливості слова в контексті документа. Вага деякого слова пропорційна кількості вживання цього слова 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. Так оцінюється важливість слова a_i в межах окремого документа:

$$\mathrm{TF} = \frac{n_i}{\sum\limits_k n_k},$$

де n_i – число 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 трапляється 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 $n_i \neq 0$).

Отже, 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.

Є різні формули, що базуються на методі TF-IDF. Вони відрізняються коефіцієнтами, нормуванням, використанням логарифмічних шкал. Так пошукова система Яндекс впродовж тривалого часу використовувала нормування за найчастотнішим терміном у 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. Таку модель називають векторною моделлю (VSM), вона дає можливість порівнювати тексти, порівнюючи їх вектори в якій-небудь метриці (евклідовий простір, косинусна міра, манхеттенська відстань, відстань Чебишова та ін.). Запропоновано для визначення ваги речення зважувати міру TF-IDF онтологією ПО. Тобто $v = (\mathrm{TF\text{-}IDF}) \cdot W$. Така оцінка має значні переваги

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

Постановка задачі. Розробити метод квазіреферування текстових документів на основі зважування міри TF-IDF, який б враховував специфіку предметної області до якої входить реферований документ. Така специфіка відображається в онтології предметної області.

Викладення основного матеріалу. ***Зважування понять на основі онтологій.*** Всі вищенаведені методи не враховують специфіки предметної області та окремих тем, які можуть в ній бути. Така специфіка враховується в онтологіях. Тому ми пропонуємо використати адаптивні онтології, які містять коефіцієнти важливості понять W та зв'язків L [7, 8]. Ці коефіцієнти обчислюють 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 рисунку.

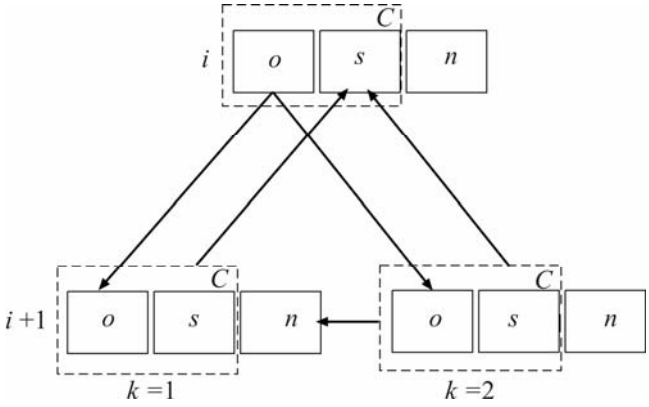

Схема перерахунку окремих компонент повної ваги класу.

2. У момент внесення на i+1-й рівень нового підкласу йому присвоюється власна вага Wo_j^{i+1}, рівна 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 визначення зв'язку між поняттями k_1 та k_2 між відповідними вершинами графа онтології з'являється ребро, а до ваги суміжних класів Wn_1 дода-

ється вага Wc_2 і навпаки – до Wn_2 додається вага нового суміжного до нього класу Wc_1 так, що:

$$Wn_j = \sum_k Wc_k \cdot L_{j,k} \ . \quad (4)$$

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 валентність вершини.

5. Вага екземпляра БЗ дорівнює повній вазі його класу.

Так 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 відносяться до визначеної цією онтологією ПО. Тому коефіцієнти важливості понять та зв'язків залежать від тематики. Нехай онтологія O описує m тем ПО – Th_1, Th_2, …, Th_m, тоді коефіцієнти ваг понять та зв'язків насправді собою представляють 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 квазіреферування наперед треба вибрати тему, до якої належатиме текстовий документ, що опрацьовується, щоб система використовувала правильні ваги.

Визначення основних понять і властивостей графа 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 графа 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 графа мережі G міститься деяка множина елементів, що характеризують відповідний цій вершині об'єкт. Ребра графа, 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 графа семантичної мережі онтології – властивість, яка означає, що усі елементи мережі знаходяться у межах досяжності інтелектуальної системи і можуть бути задіяні під час генерування відгуку на звертання до неї.

Опишемо взаємозв'язок між структурою зв'язків онтології та механізмами реалізації міркувань. Модель повинна містити механізми міркувань, в якості яких виступатимуть приєднані процедури фреймів, що використовують визначені зв'язки (твердження) з метою вироблення необхідного рішення. Згідно з об'єктною парадигмою та фреймовою моделлю подання знань батьківський клас-фрейм містить приєднані процедури визначення 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. Таким чином, кожен екземпляр деякого класу містить в собі 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 зумовлювати цю 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 в області 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). 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 $Vi \rightarrow Vj$, то існує також $Vj \rightarrow Vi$), а кожному типу зв'язку присвоєні певні коефіцієнти подібності. Наприклад, для зв'язку типу спеціалізації ("IS-A") – $\sigma = 0{,}9$, для узагальнення ("KIND-OF") – $\gamma = 0{,}4$, для причинного зв'язку ("CAUSED-BY") – $\rho_{CBY} = 0{,}3$, для характеризуючого зв'язку ("CHARACTERIZED-BY") – $\rho_{WRT} = 0{,}2$. 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 суттєвих 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 графа. Пропорційна сила між двома вершинами описує важливість з'єднання одної 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 ме-

тодиках 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. Для задання ваг важливості відношень використано дослідження, які провели датські вчені Кнаппе, Бульшков та Андреасен. Вони визначили такі значення ваг відношень: ієрархія – $L_1 = 0{,}9$; агрегація – $L_2 = 0{,}8$; функціональні – $L_3 = 0{,}3$; семіотичні – $L_4 = 0{,}2$. Для відношення тотожності прийнято, що $L_5 = 1$ [5].

Квазіреферування одного текстового документа. Отже, ми як оцінку речень, що входять у текстовий документ, запропонували взяти добуток двох ваг TF-IDF та ваги термінів W в онтології, що відповідає темі, якій належить запропонований до розгляду документ. Тобто

$$\varphi = (\text{TF-IDF}) \cdot W \,. \tag{5}$$

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

Для відбору речень для квазіреферату за основу ми брали 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 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 буде виглядати так.

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

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 як вектор $\varphi = (\varphi_1, \varphi_2, ..., \varphi_k)^T$.

Кожне речення (об'єкт) подається 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}.

Набір речень – це зважений граф з матрицею ваг $X = (x_{ij})$. Для кожної пари x_i та x_j речень обчислюється вага їх "лексичній близькості" за допомогою стандартної Евклідової міри:

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

де

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

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

$$S = D^{-1/2} X D^{1/2}, \tag{7}$$

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 φ обчислюють як результат ітеративного процесу:

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

де $\bar{y}$ – одиничний вектор.

Згідно з теоремою, наведеною в [5], ітеративний процес збігається до φ^*. Таким чином φ_i^* – отриманий ранг речення A_i . Алгоритм полягає в поступовому розповсюджені 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 (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. Ми розробили систему квазіреферування на основі розробленого методу. Система шукає у вхідному тексті головне речення і формує квазіреферат з вказанням 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 передачі основного змісту документу; б) точність – відсутність у квазірефераті речень, надлишкових для передачі 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 виключно 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 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). Оцінка якості окремих квазірефератів текстів обох класів наведена в таблиці.

Обсяг одержаних квазірефератів – від 3 до 6 речень; у двох випадках обсяг склав 7 речень. Це були документи, котрі не підлягають інтелектуальному реферуванню. Отже, експеримент дав змогу зробити такі висновки. Одержані квазірефератиістять мало надлишкової інформації, а її наявність зумовлена в основному помилками, не пов'язаними з якістю нашої моделі. Введені в квазіреферат речення містять, як правило, основну інформацію вхідного тексту, тобто відповідають визначенню головного речення. Кількість головних речень, як правило, складає не більше 25% всіх речень цього тексту: коефіцієнт стиснення (менше 4) одержаний 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 реферати з 20 одержали низьку оцінку по параметру “зв’язність”, тобто ці реферати мають вигляд радше штучних об’єднань речень, які відносяться до однієї теми, ніж тексту. Основною причиною цього були зовнішні для нашої моделі чинники, тому треба вважати одержаний результат попереднім і таким, що потребує додаткової перевірки.

Оцінка якості квазіреферату

Назва файлу	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
porn sifter.doc	4	5	4	4

ВИСНОВКИ

Розроблено метод квазіреферування текстових документів на основі модифікації міри TF-IDF. Суть такої модифікації полягає у зважуванні термінів, понять предметної області та зв’язків між ними. Зважування відбувається завдяки вагам важливості концептів предметної області, які зберігаються в її онтології. Побудований на основі такого підходу квазіреферат показав задовільну якість.

1. *Белоногов Г. Г., Калинин Ю. П., Хорошилов А. А.* Компьютерная лингвистика и перспективные информационные технологии. – 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., Гайдін В. А., Пшеничний О. Ю.* Метод автоматизованого реферування текстових документів з використанням онтологій // Складні системи і процеси. – Запоріжжя. – 2009. – № 1. – С. 81–87.
4. *Крайовський В. Я., Литвин В. В., Шаховська Н. Б.* Основні підходи до розроблення програмного комплексу автоматичного реферування текстових документів. – К.: Ін-т проблем моделювання в енергетиці, 2009. – Вип. 51. – С. 178–186.
5. *Knappe R., Bulskov H., Andreasen T.* Perspectives on Ontology-based Querying [Електронний ресурс] // Int. J. Intelligent Systems. – 2004. – Режим доступу: http://akira.ruc.dk/~knappe/publications/ijis2004.pdf.
6. *Ланде Д. В., Снарский А. А., Безсуднов И. В.* Интернетика: Навигация в сложных системах: модели и алгоритмы. – М.: Книжный дом “Либроком”, 2009. – 264 с.
7. *Lytvyn V.* Design of intelligent decision support systems using ontological approach // An international quarterly journal on economics in technology, new technologies and modelling processes. – Lublin. – 2013. – **II**, № 1. – P. 31–38.
8. *Литвин В. В.* Підхід до побудови інтелектуальних систем підтримки прийняття рішень на основі онтологій // Проблеми програмування. – К.: Ін-т програмних систем НАН України. – 2013. – № 4. – С. 43–52.
9. *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). – С. 105–111.
10. *Литвин В. В.* Бази знань інтелектуальних систем підтримки прийняття рішень. – Львів: Вид-во Львівської політехніки, 2011. – 240 с.

Національний університет “Львівська політехніка”
Фізико-механічний інститут ім. Г. В. Карпенка НАН України, Львів

Одержано 03.03.2014