УДК 004:37

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

ІНФОРМАЦІЙНА СИСТЕМА РЕФЕРУВАННЯ МНОЖИНИ ДОКУМЕНТІВ, ПОДАНИХ У РІЗНИХ ФОРМАТАХ, БАЗОВАНА НА ОНТОЛОГІЇ

Описано систему автоматичного реферування множини документів. Показано, що серед існуючих систем практично немає таких, що займаються множинним реферуванням. Спроектовано інформаційну модель такої системи.

Ключові слова: онтологія, реферування, автоматизація, документ

The system of the automatic abstracting of great number of documents is described in the article. It is shown that among the existent systems practically is not such which engage in plural abstracting. The informative model of such system is projected.

Keywords: ontology, abstracting, automation, document.

Вступ

Задача автоматизації процесу реферування текстової інформації сьогодні залишається дуже актуальною, незважаючи на величезну кількість робіть, що зроблені за останні роки в 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-якому виданню – чи це монографія, підручник, аналітичний огляд тощо – завжди передує вторинний документ (реферат або анотація). Реферування використовується не тільки для економії часу при ознайомленні з великою кількістю джерел, але й з метою пришвидшення повнотекстового пошуку за множиною документів, оскільки обсяг реферату у декілька разів менший, ніж обсяг вхідного документу чи їх множини.

1. Аналіз останніх досліджень

Порівняємо комерційні системи реферування. Результати порівняння подано в табл. 1 [1].

Таблиця 1

Порівняльна характеристика систем автоматичного реферування

Системи автоматичного реферування\Властивості	Розміщення у тексті	Ключові фрази	Довжина речення	Машинне навчання	Дискурсивна модель	Один документ	Багато документів	Фрагментарний текст	Зв'язний текст	Загальний реферат	Спеціальний реферат	Налаштування стиснення	Багатомовний	Можливість налаштування реферування
Auto Sumarizer						+		+				+		
CONTEXT						+		+		+		+		
Data Hammer						+		+				+	+	+
DimSum				+		+						+	+	+
Extractor				+		+		+		+			+	+
GE Summarizer	+		+		+	+		+	+	+	+	+		
Intelligent Miner	+					+		+		+		+		
IntellScope	+					+		+		+	+	+	+	
InText						+		+			+	+		
InXihtSummarizerPlus	+	+	+	+		+		+		+		+	+	+
ProSum	+		+			+				+				
Search'2005 Developer Kit	+	+	+			+		+		+	+	+		
SMART						+			+		+			
SUMMARIST						+		+		+	+		+	
TexNet32						+		+		+	+	+		
TextAnalyst2.0				+		+	+				+			

Як бачимо, з найпопулярніших систем-реферувальників дуже мало складає реферат з декількох джерел. Тому **метою** реалізації буде розроблення архітектури системи реферування декількох джерел. Джерелами реферату виступатимуть текстові файли довільного формату:

- .txt,
- .doc,
- .rtf,
- веб-сторінки,
- xml-файли,

які зберігаються у локальній чи глобальній мережі.

Також проблемою сучасних систем реферування є те, що вони зазвичай орієнтуються на тексти англійською мовою. Для текстів українською мовою розроблено лише неповні онтології у деяких предметних областях. Областю, яка має достатньо багато усталених та відомих автору термінів і інформація про яку є доступна, є область інформаційних технологій. Саме її ми обрали для розроблення онтології та демонстрації роботи системи.

2. Основний матеріал.

Розроблення структури системи

Спроектуємо ER-модель системи реферування множини документів. Для цього побудуємо діаграми потоків даних та опишемо особливості опрацювання інформації всередині системи.

Насамперед опишемо зовнішні сутності, які водночас є і джерелами даних у задачі реферування. Концептуальну схему подано на рис. 1.

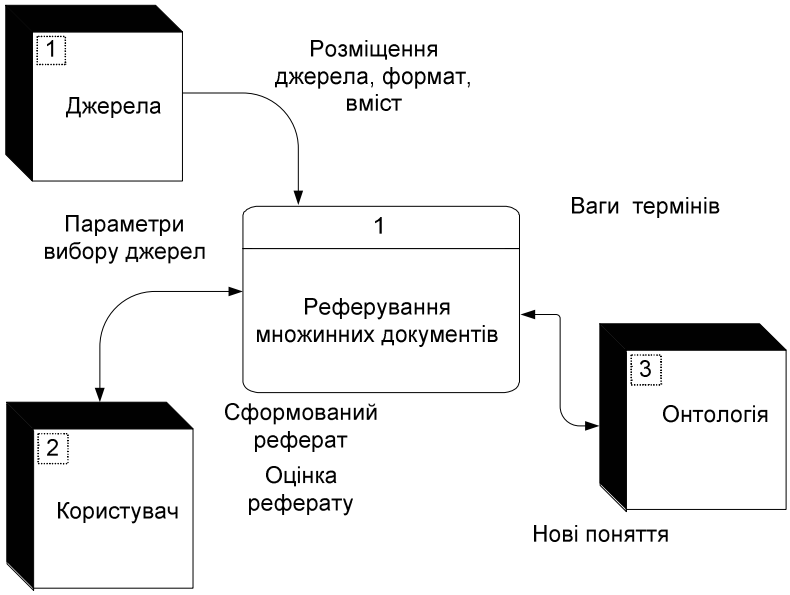


Рис. 1. Концептуальна схема системи реферування множини документів

Зовнішніми сутностями є Джерела даних, Користувач та Онтологія. Джерела даних – множини файлів різних типів для реферування. Користувач задає параметри пошуку джерел даних для реферування та переглядає отриманий реферат і оцінює його якість. Зовнішня сутність Онтологія містить перелік тем документів, термінів та їх ваг стосовно тем.

Процес 1. Реферування множинних документів складається з 5 підпроцесів, поданих на рис. 2.

Також використано три сховища даних:

– Сховище джерел – містить опис джерела (шлях до документа, його власника, тип документа, процедури, що використовуються для доступу до нього тощо);

– Сховище термінів – онтологія проблемної області (інформаційних технологій);

– Сховище вимог користувачів – вимоги до реферату, що будується (коефіцієнт стискання, максимальна кількість речень тощо).

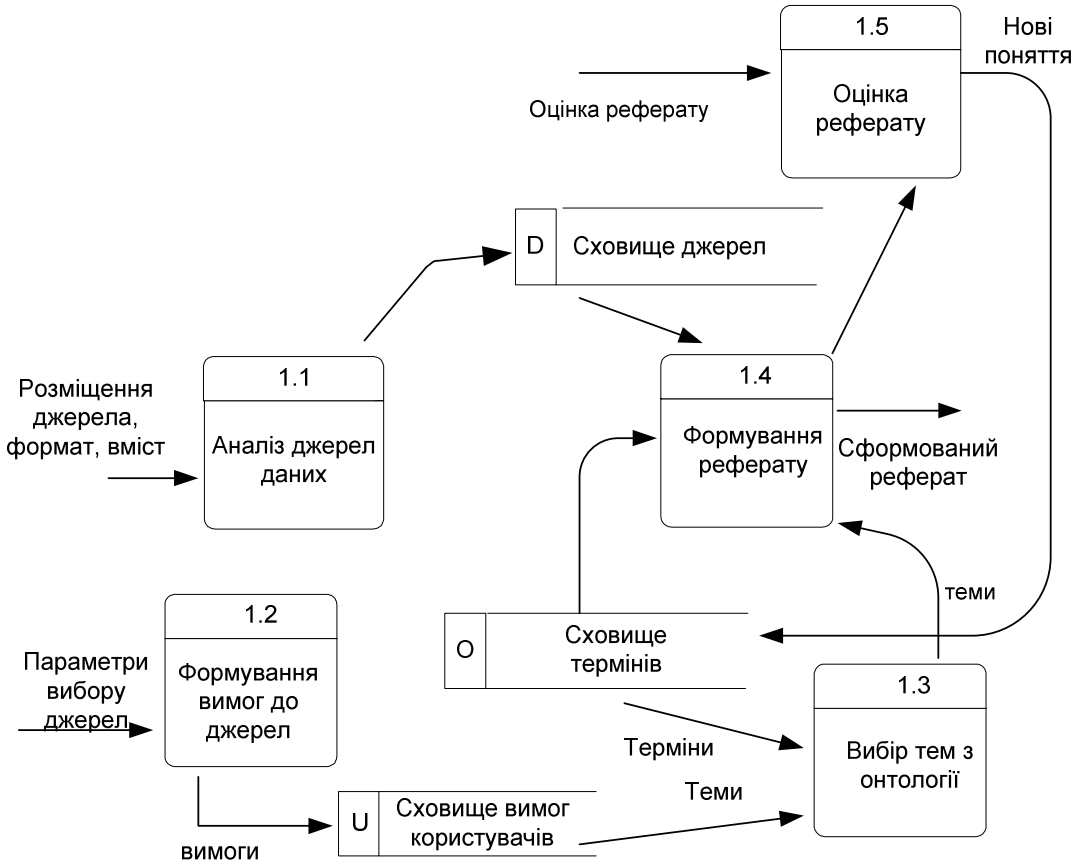


Рис. 2. Діаграма розгортання потоків даних Реферування множинних документів

Підпроцес 1.1. Аналіз джерел даних виконує визначення формату джерела. Підпроцес 1.2. Формування вимог до джерела визначає теми та основні вимоги до джерел, за якими здійснюватимуть реферування. Підпроцес 1.3. Вибір тем з онтології. Вибирають терміни, які шукатимуть у джерелах. Підпроцес 1.4. Формування реферату. Власне реферування за вибраними термінами та з вибраних джерел. Реферат повертається зовнішній сутності Користувач. Підпроцес 1.5. Оцінка реферату. Користувач оцінює реферат та додає нові терміни до онтології.

Деталізація підпроцесу «Аналіз джерел даних» подана на рис. 3.

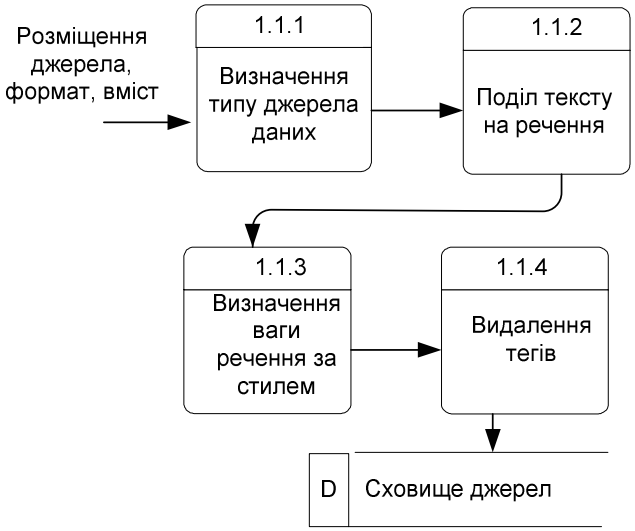
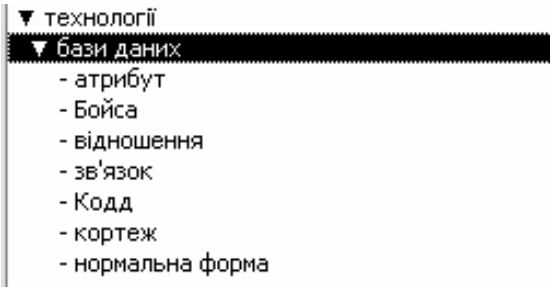


Рис. 3. Діаграма розгортання потоків даних Реферування множинних документів

Розроблення засобів опису онтології предметної області

Для написання програми реферування документів та опису онтології використано мову програмування Visual Basic. Для виділення речень у файлах з розширенням .doc та .rtf використано мову програмування Visual Basic for Application (VBA).

Онтологія предметної області описується ключовими словами, поданими на рис. 4.



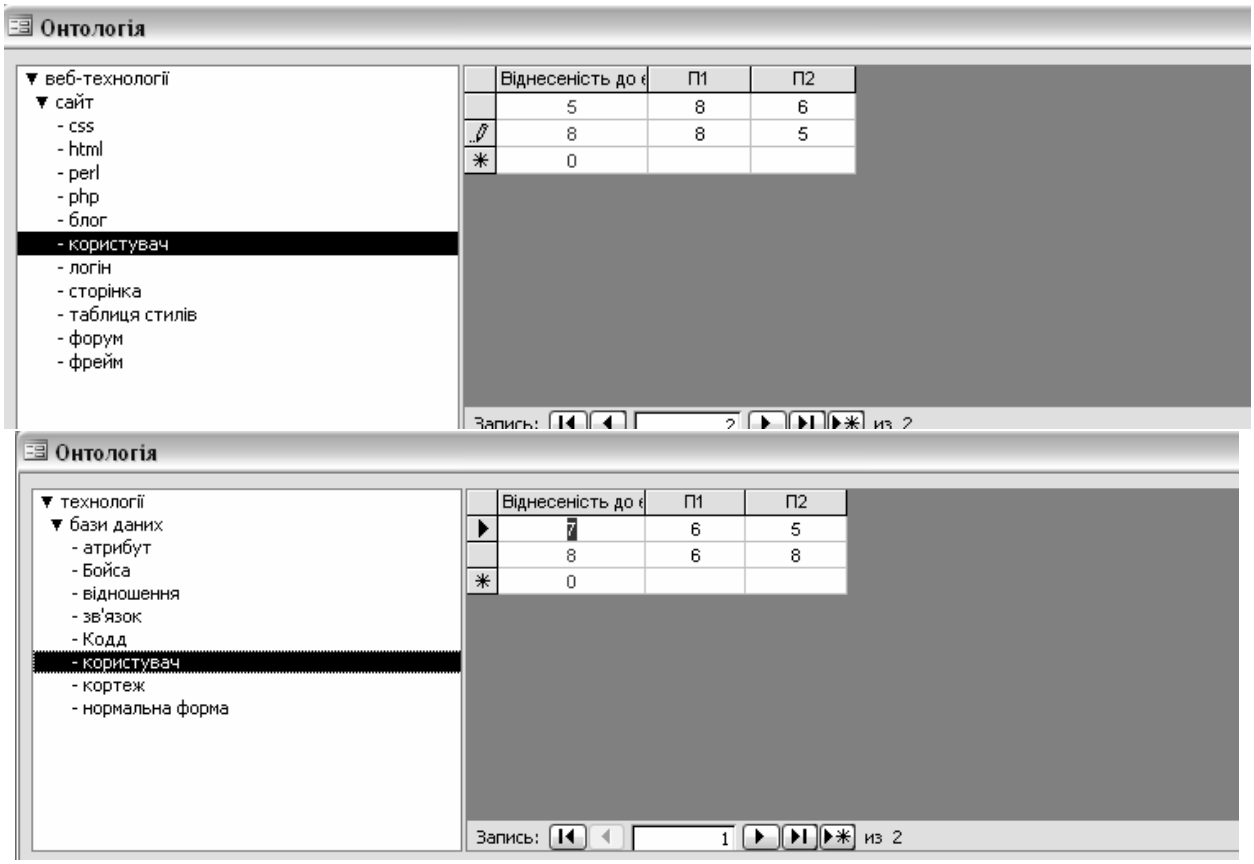


Рис. 4. Онтологія предметної області веб-сервісів

Ліва частина форми містить дерево онтології. Кожне з понять онтології має свій внутрішній код (рис. 5) Права частина форми описує важливість терміна до заданої гілки дерева. Оскільки термін може зустрічатися у різних предметних областях, то важливим є показник важливості терміна для заданої ПО (поле П1) та до ПО з кодом *Віднесення до ел онтології* (поле П2). Таблиці 1 – 3 використовуються для побудови дерева онтології. Таблиця 4 використовується для задання важливості поняття для заданої гілки.

tab_1	
Kod_1	**Name_1**
5	технології
6	аналітика
7	програмування
8	веб-технології
9	графіка

tab_2		
Kod_2	**Kod_1**	**Name_2**
14	5	бази даних
15	5	сховища даних
16	5	інтеграція
17	5	адміністрування
…	…	…
26	8	агент
27	8	форум
…	…	…….

tab_3		
Kod_3	**Kod_2**	**Name_3**
16	16	BAI
17	14	зв'язок
…	…	….
4	15	Інмон
43	25	користувач
79	14	користувач
…	…	…
59	25	логін
14	15	метадані
…	…	….

tab_4		
Kod_3	**П1**	**П2**
43	8	6
79	6	5
79	6	8
43	8	5
…	…	…

Рис. 5. Фрагмент табличного подання онтології із зазначенням кодів термінів та їх важливостей

Додавання джерел та визначення їх структури здійснюється у формі, поданій на рис. 6.

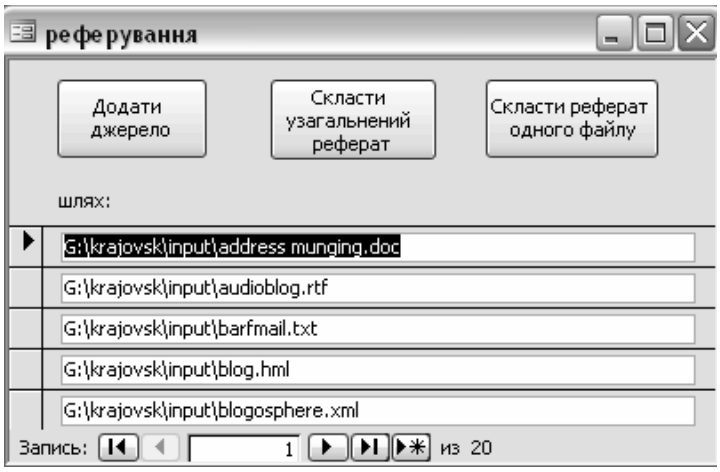


Рис. 2. Опис джерел для реферування

Опис об’єкта онтології подається у формі на рис. 7.

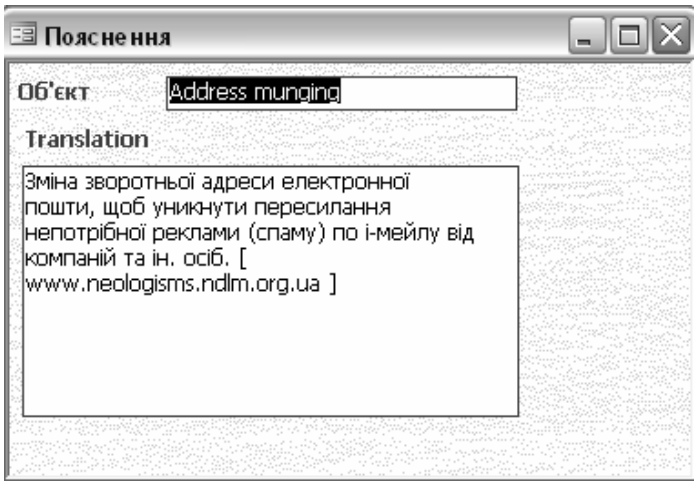


Рис. 3. Опис терміна онтології

Перейдемо до аналізу роботи алгоритмів реферування.
Насамперед проаналізуємо результат побудова одиничного реферату.

Аналіз роботи алгоритму реферування одиничного документа

Реферат складатиметься з речень, взятих з тексту оригінального повідомлення. Для того, щоб речення потрапило у реферат, воно повинно містити слова з онтології.

Перш за все виділимо критерії, за якими виділялися речення (назвемо їх головними):

- Речення – заголовок певного рівня.
- Вага речення не нижча за вагу, встановлену експертом.

Після виділення головних речень до них висувається ряд умов:

- Кількість головних речень тексту становить не більше 25% всіх речень цього тексту.
- З головних речень може бути складений такий новий текст, що утворить гіперсинтаксичну структуру.

Автори розробили подібну описаній вище систему квазіреферування, що витягає зі вхідного тексту головне речення і формує квазіреферат із зазначенням смислових класів. Система використовує морфологічний і гіперсинтаксичний засоби “розуміння” тексту. Перевіряли гіпотезу на масиві 20 довільно відібраних статей за тематикою інформаційних технологій (рис. 6).

Було введено такі якісні характеристики квазірефератів:

а) повнота передачі основного змісту документа;

б) точність – відсутність у квазіреферате речень, надлишкових для передачі основного змісту документа;

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, в припущенні, що квазіреферат в ідеалі повинен мати статут самостійного документа, тобто надавати 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.

Опрацьовувані документи було поділено на два класи:

а) які піддаються інтелектуальному реферуванню,

б) які не піддаються інтелектуальному реферуванню (наприклад, таблиця порівнянь швидкостей процесорів).

Оцінка якості квазірефератів текстів обох класів наведена в табл. 2.

Таблиця 2

Оцінка якості реферату

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
crowdfunding.doc	4	5	4	3
Cyber Monday.rtf	4	4	5	4
cyberbalkanization.doc	3	3	4	2
cyberpiracy.htm	3	4	5	4
cybersquatting.doc	3	5	5	4
dotbam.txt	3	5	5	4
e-commerce.htm	3	5	4	3
e-cruitment.htm	4	4	4	3
e-mail bankruptcy.doc	3	4	4	2
e-mail fatigue.doc	4	4	4	4
e-signature.doc	3	5	5	3
nooksurfer.doc	4	5	4	4
porn sifter.doc	4	5	4	4

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 квазіреферати містять мало надлишкової інформації, а її наявність викликана 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, становить не більше 25% всіх речень цього тексту (табл. 2): коефіцієнт стиснення, менший за 4, одержано тільки для дуже коротких текстів.

Припущення про те, що з головних речень може бути складений новий текст, який 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 належать до 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.

3. Аналіз роботи алгоритму узагальненого реферування

Формування узагальненого реферату ґрунтується на формуванні проміжних рефератів. Результат формування проміжних рефератів подано у табл. 2.

На основі проміжних рефератів знову запускається процес реферування. У результаті повторного реферування отриманих квазірефератів отримується один документ (рис. 4).

World Wide Web (WWW) – один з видів сервіса Internet. WWW надає можливість роботи з документами, в яких об"єднані текст, графічні ілюстрації, звукові фрагменти і анімації.

Web-система представляє собою розподілену інформаційну систему з уніфікованим інтерфейсом доступу до даних, базованому на використанні ряду відкритих стандартів. Це в першу чергу стандарт представлення текстової та гіпертекстової інформації - мова HTML.

WWW-сервер (World Wide Web) – це така частина глобальної чи корпоративної мережі, яка дає можливість користувачам мережі одержувати доступ до гіпертекстових документів, розміщених на даному сервері. Для взаємодії з WWW сервером користувач мережі повинен використовувати спеціалізоване програмне забезпечення – броузер (від англ. browser), інша назва – програма перегляду.

На сьогоднішній день існує три варіанти розміщення готового Web-сервера. У даному контексті під Web-сервером розуміється набір текстових, графічних і інших файлів, що містять усі необхідні елементи для нормального функціонування серверу.

Рис. 4. Кінцевий реферат

Після цього відсікають неінформативні речення. Кінцевий результат реферування подано на рис. 5.

World Wide Web (WWW) – один з видів сервіса Internet. WWW надає можливість роботи з документами, в яких об"єднані текст, графічні ілюстрації, звукові фрагменти і анімації.

Web-система представляє собою розподілену інформаційну систему з уніфікованим інтерфейсом доступу до даних, базованому на використанні ряду відкритих стандартів. Це в першу чергу стандарт представлення текстової та гіпертекстової інформації - мова HTML.

Для взаємодії з WWW сервером користувач мережі повинен використовувати спеціалізоване програмне забезпечення – броузер (від англ. browser), інша назва – програма перегляду.

На сьогоднішній день існує три варіанти розміщення готового Web-сервера. У даному контексті під Web-сервером розуміється набір текстових, графічних і інших файлів, що містять усі необхідні елементи для нормального функціонування серверу.

Рис.5. Кінцевий реферат

Експертна оцінка проводилася лише для текстів українською мовою. Методика оцінки полягала у такому. Шести експертам було запропоновано 10 текстів документів і реферат, побудований на їх основі. Експерти відповідали на такі запитання, вибравши відповідь за такою шкалою оцінювання:

1. Повнота відображення зміст документів? (0 – не відображає, 1 – не достатньо повно, 2 – задовільно).
2. Надмірність у рефераті? (0 – так, забагато, 1 – так, не надто багато, 2 – ні).
3. Задоволення властивості зв'язності тексту? (0 – ні, 1 – зустрічаються не зв'язані речення, 2 – так).
4. Довжина реферату (0 – дуже довгий, 1 – дуже короткий, 2 – оптимальний).

Результати експертних оцінок наводяться в табл. 3.

Таблиця 3

Експертні оцінки якості рефератів

Метод	№ експерта	Повнота	Надмірність	Оцінка зв'язності	Оцінка довжини
На основі онтології	1	2	1	1	2
	2	1	0	1	0
	3	2	2	2	0
	4	1	1	0	2
	5	2	1	2	0
	6	2	2	2	0
Г.Г. Бєлоногова	1	1	0	0	0
	2	0	1	1	0
	3	2	2	2	2
	4	0	1	0	0
	5	1	0	2	0
	6	2	2	2	0

Найчастіше експерти знижували оцінки через довжину реферату (дуже довгий) і через те, що зустрічаються речення, які порушують картину зв'язності тексту. Від довжини текстів, за якими складали узагальнений реферат, експертні оцінки практично не залежали. Порівняльне оцінювання цього методу з іншими поширеними методами реферування не проводили на великій вибірці тестових завдань. Проте ми провели експеримент з реферування на ряді україномовних текстів за цим методом і за методом, запропонованим Г.Г. Бєлоноговим [3].

Поповнення онтології

Поповнюють онтологію (додають нові терміни) у тому випадку, якщо середня інформативність речень отриманого реферату є нижча за встановлену експертно. Це означає, що у заданому тексті знайдено або мало термінів з онтології, або їх вага порівняно мала. Розроблено програмні засоби додавання терміна до онтології (рис.10). Також цю форму використовують для експертного наповнення онтології.

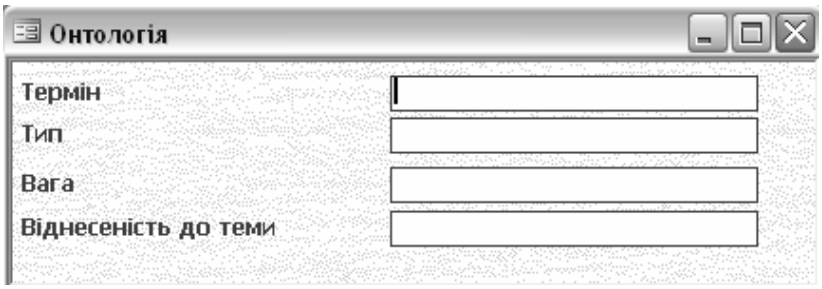


Рис. 6. Кінцевий реферат

Висновки

Реферування множини документів є актуальним та важливим завданням, вирішити яке практично неможливо сучасними засобами реферування. Ресурси, що підлягають реферуванню, мають різноманітну природу, традиційно по-різному обліковуються і потребують попереднього опрацювання.

Використання вагових коефіцієнтів значно підвищує якість отриманого реферату.

Запропонований метод складання рефератів, що розглядається у цій роботі, може бути успішно застосований в сучасних інформаційних системах під час опрацювання великих інформаційних потоків, коли проводиться автоматичне рубрикування документів, тобто віднесення їх до певної тематики. Порівняно з системами реферування, в яких проводиться повний цикл опрацювання документів, цей метод дає змогу значно скоротити тимчасові витрати на складання реферату. Проведена незалежними експертами оцінка якості реферування показала, що метод загалом дає задовільні результати. Зазначена експертами у ряді рефератів дуже велика їх довжина є недоліком, який можна усунути, скорочуючи довгі речення. У роботі [1] описується метод виділення по одному фрагменту з кожного речення реферату аж до вичерпання заданого ліміту (у символах). Проте на наш погляд, за такого підходу неможливо сформувати гладкий зв'язний текст. Тут потрібний інший підхід, близький до запропонованого в [2], заснований на синтаксичному аналізі текстових документів (хоча ця робота присвячена реферуванню російськомовних текстів, вона легко адаптується до україномовних текстів). Відштовхуючись від результатів синтаксичного аналізу речень, можна відсікати поширені доповнення, дієприкметникові і дієприслівникові звороти. У подальших дослідженнях ми збираємося випробувати алгоритм синтаксичного аналізу для реферування. У ряді випадків експерти зазначили незв'язність тексту реферату. Для вирішення проблеми зв'язності можна використовувати методи розпізнавання анафоричних зв'язків. Ця складна задача може бути предметом окремого дослідження, проте з метою поліпшення вигляду реферату на основі емпіричних спостережень можна розробити алгоритм для розпізнавання анафор. Програми, де реалізований подібний алгоритм, вже існують [3].

1. Браславский П.И., Колычев И.С. Автоматическое реферирование веб-документов с учетом запроса // Интернет-математика, 2005. – М. : Яндекс, 2005. – С. 485–501. 2. Емашова О.А., Мальковский М.Г. Функциональные стили русского языка и их влияние на задачу автоматического реферирования текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2007» (Бекасово, 30 мая). 3. Абрамова Н.Н., Абрамов В.Е. Автоматическое составление обзорных рефератов новостных сюжетов // Интернет-математика 2007. – Екатеринбург, 2007. – С. 1–11.