УДК 004:37

Н.Б. Шаховська, З.В. Стахів
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

АВТОМАТИЗОВАНА СИСТЕМА УКЛАДАННЯ РЕФЕРАТУ

Описано автоматизовану систему укладання рефератів. Спроектовано інформаційну модель такої системи. Визначені мета, завдання і сфера застосування системи.

Ключові слова: анотації, абстрагування, реферування.

The system of the automatic abstracting of documents is described in the article. The informative model of such system is projected. Goals, objectives and scope of such a system are defined.

Key words: annotation, abstracting, referencing.

Вступ. Постановка проблеми

З кожним роком зростає обсяг і потужність інформаційного потоку. Упорядкувати цей потік і тримати його в контрольованому руслі можна тільки за допомогою автоматизованих систем опрацювання інформації. В основі цих систем лежать процеси аналізу первинних документів – індексація і предметизація та синтезу – генерація вторинних документів, складання каталогів.

Суть наукового опрацювання документів полягає в процесі створення і перетворення документів. Цей процес призначений для полегшення користувачеві пошуку та виявлення необхідної інформації. Основними видами наукового опрацювання документів є: класифікація, сортування, перетворення, розміщення в базі даних і пошук. Результатом реферування документів є вторинні документи – реферати.

Ознайомлення з рефератами дає змогу оперативно одержати коротку інформацію про зміст первинних документів і завдяки цьому максимально правильно вирішити питання про необхідність використання їх. Інколи таке ознайомлення навіть замінює вивчення першоджерела, що особливо важливо, коли воно з якихось причин недоступне. Реферати також використовуються при формуванні бібліографічних і фактографічних пошукових масивів традиційних і автоматизованих інформаційних пошукових систем. Саме тому розроблення нових, ефективних систем автоматизованого укладання рефератів є актуальним.

Аналіз останніх досліджень та публікацій

Властивості реферату

Для орієнтації в потужних документно-інформаційних потоках, для проведення ефективного й оперативного пошуку інформації здійснюється наукове (або аналітико-синтетичне) опрацювання документів. Її суть полягає у згортанні інформації про первинні документи на основі застосування методів аналізу і синтезу.

Оскільки користувачі ставлять різні вимоги до згортання інформації про ознаки документів, існують різні види аналітико-синтетичного опрацювання. Безумовно до них можна зарахувати такі: складання бібліографічних описів, індексування, анотування, реферування.

Різноманітні види аналітико-синтетичної обробки документів використовуються всюди, де люди мають справу з документами, а найбільше у сфері документних комунікацій [1].

Реферат – це багатофункціональний вторинний документ. Він виконує безліч функцій: інформативну та науково-комунікативну, прогностичну, довідкову і адресну, індексування й індикативну. Відповідно до завдань реферат може надавати необхідну систематизовану фактографічну інформацію, оцінювати, узагальнювати, синтезувати її, рекомендувати найбільш нові, цінні та корисні повідомлення для конкретного користувача.

Особливе значення має інформативна функція. З усіх вторинних документів саме реферат розкриває зміст первинного документа найповніше, у цілісному, узагальненому вигляді. Вивчення першоджерел із залученням рефератів істотно економить час [2].

У системі наукової комунікації реферат є основною інформаційно-комунікативна одиницею, що зумовлено його споживчими властивостями:

1) серед усіх видів вторинних інформаційних документів реферат відрізняється найбільшою інформативністю в розкритті змісту першоджерела;

2) використання реферату для пошуку поточної або ретроспективної інформації дає змогу зекономити до 90 % часу, необхідного в разі звернення до первинних документів;

3) форма подання інформації у вигляді реферату зручніша для тривалого зберігання у фондах довідково-інформаційних служб й ІПС, полегшує та прискорює підготовку інформаційних видань і створення інформаційних масивів – БД рефератів (на їх підгрунті в подальшому можливе створення БД повнотекстових електронних документів);

4) у деяких випадках реферат може замінити першоджерело (коли необхідна користувачеві інформація стосується не основної теми роботи, а суміжних питань, або коли первинний документ недоступний унаслідок мовного або організаційного бар'єрів).

Центральне положення реферату в системі наукової комунікації не похитнулося й після широкого впровадження в інформаційну технологію засобів обчислювальної техніки. Під час вдосконалення інформаційних технологій відбувається лише перерозподіл його комунікативно-інформаційних функцій внаслідок зміни інформаційних потреб суспільства. Реферат міцно зайняв місце основного уніфікованого опису первинного документа в автоматизованих пошукових системах.

Аналіз методів автоматичного реферування

Процес реферування розпадається на три етапи: аналіз початкового тексту, визначення його характерних фраґментів і формування відповідного висновку. Більшість сучасних робіт концентруються навколо розробленої технології реферування одного документа.

Метод складання цитат припускає акцент на виділення характерних фраґментів (зазвичай, речень). Для цього методом зіставлення фразових шаблонів виділяються блоки найбільшої лексичної і статистичної релевантності. Створення підсумкового документу в такому випадку – просте об’єднання вибраних фраґментів.

У більшості методів застосовується модель лінійних вагових коефіцієнтів. Основу аналітичного етапу в цій моделі складає процедура призначення вагових коефіцієнтів для кожного блоку тексту відповідно до таких характеристик, як розташування цього блоку в оригіналі, частота появи в тексті, частота використання в ключових реченнях, а також показники статистичної значущості. Сума індивідуальних ваг, зазвичай, визначена після додаткової модифікації відповідно до спеціальних параметрів налаштування, пов’язаних з кожною вагою, дає загальну вагу всього блоку тексту *U* [3]:

$$Weight(U) = Location(U) + Cuephrase(U) + Statterm(U) + Addterm(U). \quad (1)$$

Ваговий коефіцієнт розташування (*Location*) в такій моделі залежить від того, де у всьому тексті або в окремо взятому параграфі з’являється певний фраґмент – на початку, всередині або в кінці, а також чи використовується він у ключових розділах, наприклад, у вступній частини чи в кінці. Ключовими фразами є лексичні або фразові підсумкові конструкції, такі як «на закінчення», «у цій статті», «згідно з результатами аналізу» і так далі. Ваговий коефіцієнт ключової фрази може залежати також і від прийнятого в певній предметній області оцінного терміну типу «відмінний» (найвищий коефіцієнт) або «малозначний» (значно менший коефіцієнт).

Крім того, при призначенні вагових коефіцієнтів в цій моделі враховується показник статистичної важливості (*Statterm*). Статистична важливість обчислюється на підставі даних, отриманих у результаті аналізу автоматичної індексації, при якому дослідники виявляють і

оцінюють цілий ряд метрик, що визначають вагові коефіцієнти терміну. Ці метрики дозволяють виділити документ із числа інших у певному наборі документів.

Одна група метрик, наприклад, метрика tf-idf, характеризує баланс між частотою появи терміну в документі і частотою його появи в наборі документів (як правило, використовується з іншими метриками частоти і засобами нормалізації довжини).

І, нарешті, ця модель припускає переглядання термінів у блоці тексту і визначення його вагового коефіцієнта відповідно до додаткової наявності термінів (*Addterm*) – чи з'являються вони також у заголовку, в колонтитулі, першому параграфі і в призначеному для користувача профілі запиту. Виділення пріоритетних термінів, що найточніше відображають інтереси користувача, – це один зі шляхів налаштувати реферат або анотацію на конкретну людину чи групу.

Почергово опрацьовується кожне речення початкового тексту. Вагові коефіцієнти ґрунтуються на вимірюваннях статистичної важливості (від частоти згадки терміну до операцій, наявності спеціальних термінів і розташування речення в тексті. Вагові коефіцієнти речення, отримані на етапі аналізу, передаються безпосередньо на вхід компоненту синтезу, на якому витягаються речення з найвищими коефіцієнтами, визначеними за ступенем стискання [3].

Аналіз систем реферування

Хоча деякі виробники вже нині пропонують інструменти для реферування, обсяг інформації в мережі росте і оперативно отримувати її коректні зведення стає все складніше. Такі інструменти, як функція Autosummarize в Microsoft Office, системи IBM Intelligent Text Miner, Oracle Context і Inxight Summarizer (компонент пошукового механізму Altavista), безумовно, корисні, але їх можливості обмежені виділенням і вибором оригінальних фрагментів з початкового документу і з'єднанням їх у короткий текст. Підготовка ж короткого викладу має на меті описати основний зміст тексту, і не обов'язково тими самими словами.

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 [4].

Формулювання цілей статті

Метою роботи розроблення автоматизованої системи укладання реферату, яка дає змогу значно скоротити тимчасові витрати на складання реферату в порівнянні з іншими системами реферування. Система призначена для роботи переважно з текстами української та російської мов, що дає значну перевагу, оскільки більшість сучасних систем все ж орієнтуються на англомовні тексти.

Аналіз отриманих результатів

Алгоритм реферування

Наукова стаття, структурована текстова інформація, розбивається на слова, при цьому зразу відкидаються слова, що містять менше трьох символів, формується загальний список слів у документі, при цьому зберігається інформація про їх форматування та місце в тексті. Далі слова класифікуються, шляхом видалення з загального списку слів, які містяться в базі даних «Стоп-слова» та неінформативних слів і словосполучень. До бази даних «Стоп-слова» входять службові частини мови. Таким чином поповнюється множина «Ключові слова тексту». Вона модифікується в процесі стеммінгу, тобто відкидаючи закінчення слів, ми також видаляємо однакові слова з бази даних, але збільшуємо значення, що відповідає за кількість вживань цього слва в тексті, а ваги, що були попередньо присвоєні цим словам, додаються. Користувач може вносити свої ключові слова і вищначати їх вагу, таким чином спрямовуючи систему на виділення інформації, яка пов'язана з введеними ключовими словами.

Алгоритм реферування, описаний в статті, також базується на понятті ваги речення, розрахований на опрацюваня наукових статей.

Наукова стаття – це документ, який складається з таких елементів: назва *T*, ключові слова *K*, автор *A*, основна частина *M*, література *L*:

$$D = \{T, K, A, M, L\} \quad (2)$$

Визначення елементів документа здійснюється на основі виділення таких ознак тексту:

- місце розміщення в документів,
- місце розміщення абзацу (вліво, вправо, центрування),
- тип написання (жирний, курсив, підкреслення, звичайне накреслення),
- символи розпізнавання.

На основі вказаних ознак формують базу правил розпізнавання елементів документа (табл. 1). Для формування реферату виділяються речення з основної частини. Основна частина своєю чергою ділиться на фрагменти за підрозділами та розділами, введеними авторами. Вважається, що речення, що з'являються у вступній частині та висновках, мають вище інформативне значення, ніж речення зі середини тексту [3].

Таблиця 1

База правил розпізнавання елементів

id	type	place	paragraph	alpha	symbols
1	title	BEGIN	{Center;Right}	{Bold}	
2	author	BEGIN	{Center; Left}		{By;©: (C) }
3	keyword	BEGIN			{Keyword;Keywords;Ключові слова;Ключевые слова}
4	main	CENTER			
5	literature	END		{Typical, Italic}	

Передусім введемо поняття ваги речення. Для цього формалізуємо елементи формули (1).

Коефіцієнт розташування визначається як

$$Location = \frac{1}{n \cdot m}, \quad (3)$$

де $n = \overline{1..3}, m = \overline{1..3}$ – місця заходження в основні частині та абзаці відповідно. Почоток та кінець тексту або абзацу оцінюються значенням 1, середина – 3.

Коефіцієнт ключової фрази визначається на основі входження в речення *U* елементів з множини значущих фраз *A* на основі функції належності:

$$Cuephrase = m_A(U), \quad (4)$$

A={«На закінчення», «У підсумку», «Отже»…}.

Показник статистичної важливості формується на осноі появи в реченні ключових слів, зазначених автором статті:

$$Statterm = m_K(U). \quad (5)$$

Показник додаткової наявності термінів визначається як відношення слів речення, що зустрічаються у назві статті (*word*) до загальної кількості слів у реченні (*words*) за винятком слів, довжина яких менша за три символи:

$$Addterm = \frac{word}{words}. \quad (6)$$

Алгоритм відбору речень передбачає такі кроки:

1 крок: визначення загального обсягу завантаженого до опрацюваного тексту;

2 крок: визначення обсягу реферату, дозволеного користувачем;

3 крок: всі речення максимально скорочуються шляхом відкидання підрядних частин;

4 крок: визначення оцінки значущості речень за формулою 1;

5 крок: сортування речень за спаданням значущості;

6 крок: внесення в кінцевий реферат найінформативніших речень до тих пір, поки не досягнуто поріг, дозволений користувачем.

Розроблення структури системи

Для успішної реалізації проекту об'єкт проектування (ІС) повинен бути насамперед адекватно описаний, повинні бути побудовані повні і несуперечливі функціональні й інформаційні моделі ІС. Крім того, у процесі створення і функціонування ІС інформаційні потреби користувачів можуть змінюватися чи уточнюватися, що ще більш ускладнює розробку і супровід таких систем.

На головній контекстній діаграмі (рис. 1) показано зовнішні зв'язки автоматизованої системи укладання реферату. Головний процес (робота), який виконує система – це укадання реферату. Основну масу вхідних даних система бере з інформаційних систем електронних бібліотек, бібліотек та наукових організацій. Дані, які змінюються під час роботи – це публікації, результати наукових досліджень, наукові статті, на діаграмі позначені стрілками, що входять у ліву грань роботи, до системи посупають внаслідок дій користувача. Робота виконується, керуючись правилами оцінки значущості речень, алгоритмом відбору речень та алгоритмом АВТТ (автоматизованого визначення тематики тексту) [5], на діаграмі позначені стрілками, що входять у верхню грань роботи.

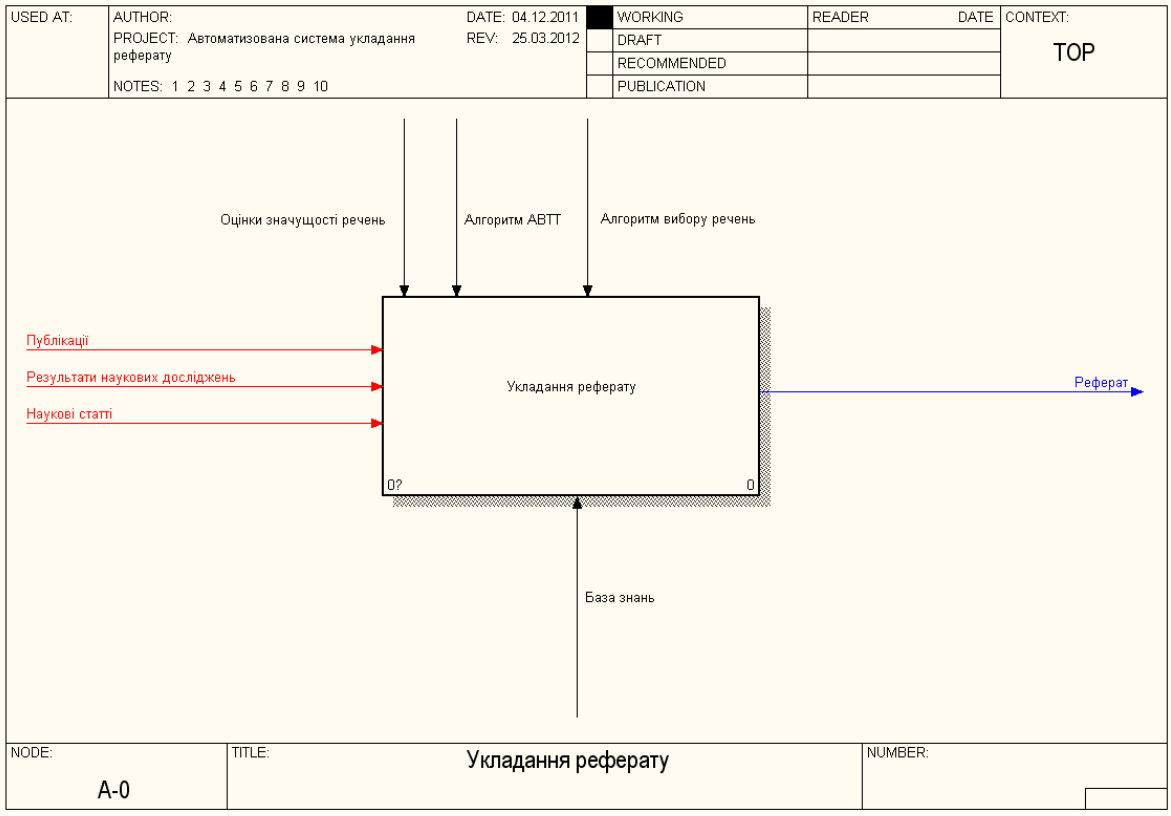

Рис. 1. Контекстна діаграма потоків даних IDEF0

Ресурсом, який є необхідним для виконання роботи виступає база знань, на діаграмі позначена стрілкою, що входить у нижню грань роботи. База знань включає в себе перелік стоп-слів (службові частини мови, вставні слова). У результаті виконання роботи отримуємо реферат, на діаграмі позначений стрілкою, що виходить з правої грані роботи.

Робота «Укладання реферату» розбивається на 5 робіт: «Декодування інформації з зовнішнього формату», «Розбиття тексту на структурні підрозділи», «Формування ключових слів», «Присвоєння ваг», «Формування результату».

Ці роботи виконуються в системі послідовно, одна за одною. Для виконання роботи «Декодування інформації з зовнішнього формату» на вхід до неї подаються дані, які змінюються в ході роботи – це

публікації, результати наукових досліджень, наукові статті, поступають в систему внаслідок дій користувача. А результатом виконання цієї роботи і відповідно вхідними даними для роботи «Розбиття тексту на структурні підрозділи» виступає текстова інформація, тобто дані, що надійшли в систему після виконання роботи «Декодування інформації з зовнішнього формату» перетворюються в дані відповідного для системи формату, в якому вони є готові для подальшого опрацювання.

Внаслідок роботи «Розбиття тексту на структурні підрозділи» текстова інформація розбивається на частини, речення. Робота відбувається керуючись алгоритмом АВТТ [5]. Результатом виконання цієї роботи є набір речень, який подається на вхід роботи «Формування ключових слів» для подальшого опрацювання та на вхід роботи «Присвоєння ваг». Ці роботи керуються правилами оцінки значущості речень та алгоритмом АВТТ, а також тут використовується база знань. Результатом роботи «Присвоєння ваг» є зважені речення. Вони надходять на вхід для роботи «Формування результату», яка враховуючи результати отримані внаслідок всіх попередніх робіт та керуючись алгоритмами АВТТ і відбору речень, формує остаточний результат – реферат.

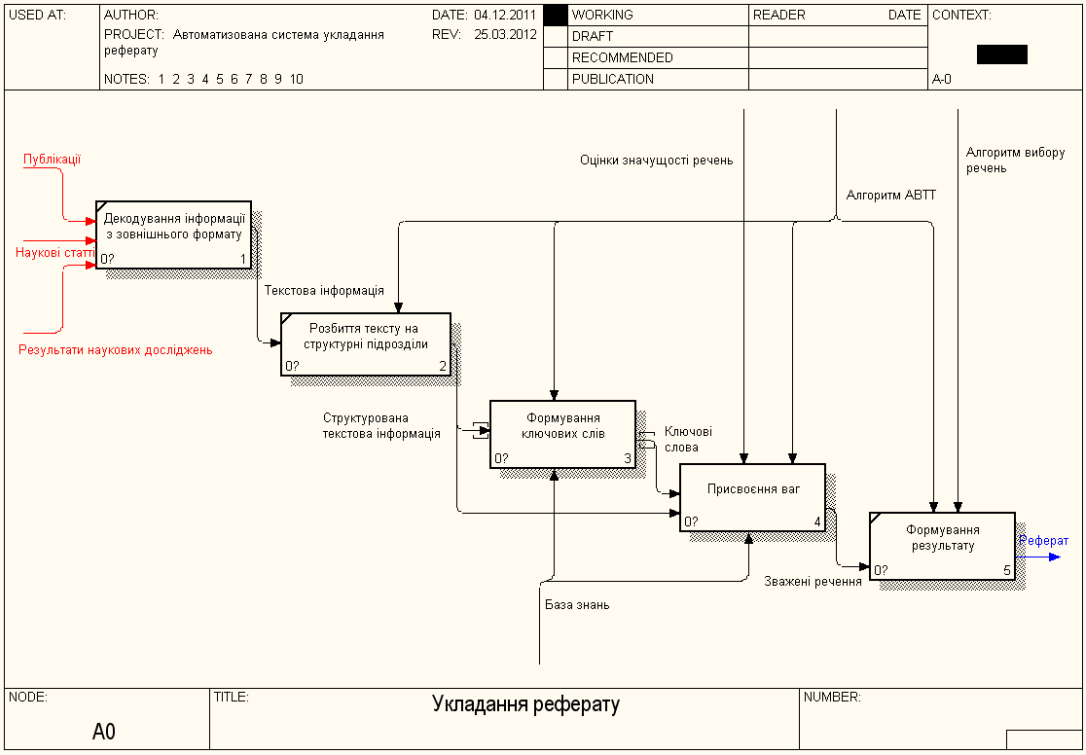

Рис. 2. IDEF0 діаграма для підзавдань головного бізнес процесу

У DFD-діаграмі блока «Формування ключових слів» (рис. 3) представлені такі роботи: «Виділення слів», «Класифікування слів», «Стемінг». Зовнішнім посиланням виступає «Користувач», який може в ході виконання цієї роботи надавати інформацію про додаткові ключові слова. Використовуються 3 бази даних: «Ключові слова тексту», «Стоп-слова», «Загальний список слів». При цьому «Загальний список слів» та «Ключові слова тексту» формується в ході виконання роботи, в той час як «Стоп-слова» тільки використовуються і не можуть бути змінені.

Розроблена система автоматичного укладання рефератів

Для написання програми реферування документів використано мову програмування C# та середовище Microsoft Visual Studio. Головне вікно програми, зображене на рис. 4, містить 4 кнопки: «Вибрати», «Додати ключове слово», «Задати параметри стиснення», «Скласти реферат».

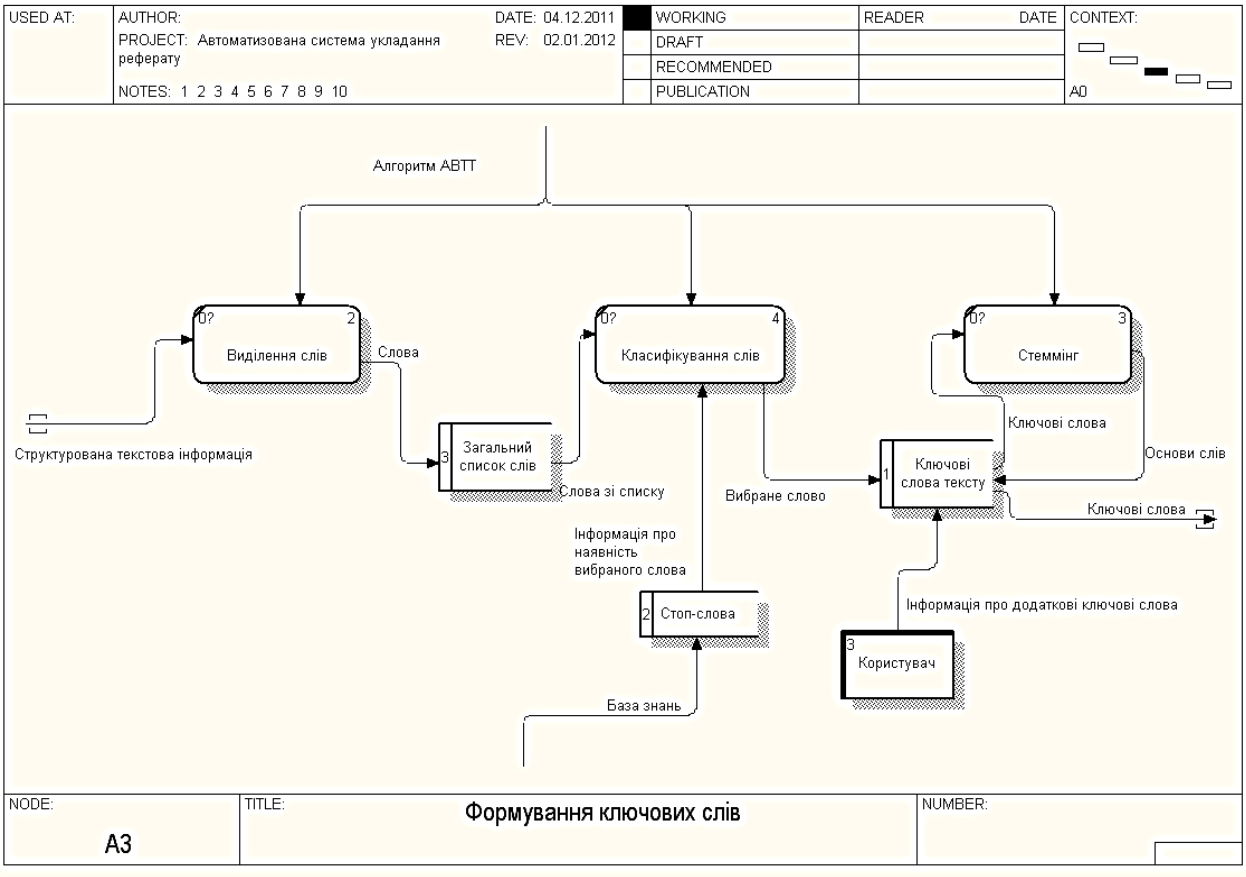

Рис. 3. DFD-діаграма блока " Формування ключових слів "

У разі натиснення кнопки «Вибрати» відкривається діалогове вікно «Відкрити файл» (рис. 5), у якому ми задаємо шлях до файлу, який буде джерелом для складання реферату. Обраний шлях до файлу, з'явиться в полі «Шлях до файлу» (рис. 4). При натисненні кнопки «Додати ключове слово» користувач задає ключове слово і його вагу. При натисненні кнопки «Задати параметри стиснення» користувач задає коефіцієнт стиснення у відсотках, тобто на скільки він хоче зменшити реферат.

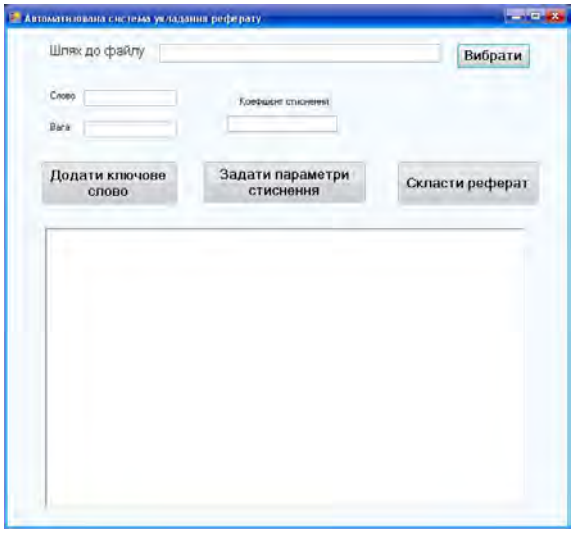

Рис. 4. Головне вікно програми «Автоматизована система укладання реферату»

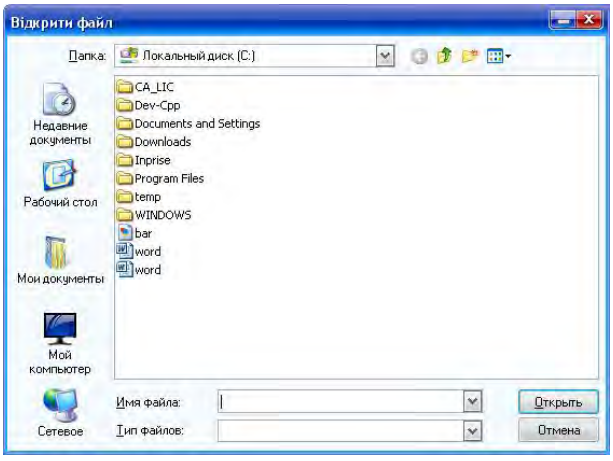

Рис. 5. Діалогове вікно «Відкрити файл»

Для розробленої системи була проведена експертна оцінка. Її проводили лише для текстів українською мовою. Методика оцінки полягала у наступному. П’яти експертам було представлено документ-джерело та отриманий на його основі реферат. Експерти відповідали на наступні питання, вибравши відповідь згідно такої шкали оцінки:

1. Наскільки повно реферат відображає зміст документів? (1 – не відображає, 2 – не досить повно, 3 – задовільно).
2. Чи присутня надмірність у рефераті? (1– так, забагато, 2 – так, не надто багато, 3 – ні).
3. Чи задовольняє реферат властивості зв'язності тексту? (1 – ні, 2 – зустрічаються не зв'язані речення, 3 – так).
4. Оцініть довжину реферату (1 – дуже довгий, 2 – дуже короткий, 3 – оптимальний).

У разі натиснення кнопки «Скласти реферат» користувач може переглянути отриманий реферат.

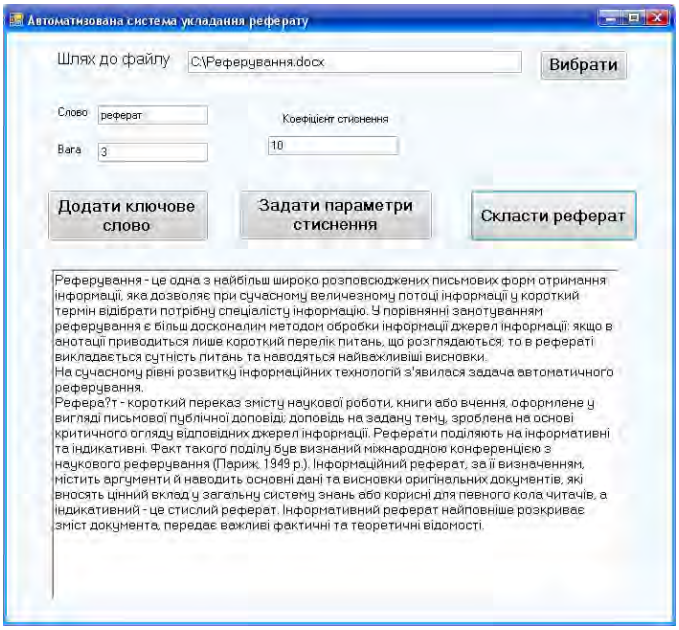

Рис. 6. Результат виконання програми

Результати експертних оцінок наведені в табл. 2.

Таблиця.2

Експертні оцінки якості рефератів

Метод	№ експерта	Повнота	Надмірність	Оцінка зв'язності	Оцінка довжини
Автоматизована система укладання реферату	1	2	2	2	3
	2	3	2	1	1
	3	3	3	3	2
	4	2	1	2	3
	5	2	3	2	3
	Агреговане твердження	2,35	2,05	1,89	2,22
Autosummarize в Microsoft Office	1	2	1	2	1
	2	3	3	1	2
	3	3	2	3	1
	4	1	1	1	1
	5	2	1	2	3
	Агреговане твердження	2,05	1,43	1,64	1,43

Найчастіше експерти знижували оцінки через те, що зустрічаються речення, які порушують картину зв'язності тексту. Розроблена система укладання реферату має вищі агреговані оцінки по всіх критеріях. Найвищі оцінки експерти поставили при оцінюванні повноти, а це означає, що система досить точно передає суть вихідного документа.

Розроблена автоматизована система укладання реферату дозволяє значно скоротити тимчасові витрати на складання реферату порівняно з іншими системами реферування. Її алгоритм роботи є доволі простим, але він має такі переваги:

- використання вагових коефіцієнтів значно підвищує якість отриманого реферату;
- користувач сам може визначати вагу деяких термінів, залежно від того, на яку тему орієнтований реферат він хоче отримати;
- система розроблена для роботи переважно з текстами української та російської мов, що дає значну перевагу, оскільки більшість сучасних систем все ж орієнтуються на англомовні тексти.

Висновки

Результатом роботи автоматизованої системи укладання реферату є вторинний документ, який прискорює відбір документів; забезпечує підвищення ступеня точності, повноти інформації; дає можливість оперативного інформування споживачів; полегшує процес індексації та класифікації документів; є засобом поточного інформування щодо нових досягнень науки і техніки; дає змогу здійснювати ретроспективний пошук. Автоматичне квазіреферування можна вважати також одним із перших кроків до автоматичної оцінки науково-технічної інформації, до автоматичного створення високоефективних фактографічних систем.

1. Сорока М.Б. Національна система реферування української наукової літератури / НАН України, Нац. б-ка України імені В.І. Вернадського. – К.: НБУВ, 2002. – 209 с. 2. Колодяжная Ж.А. Основные понятия об аннотировании и реферировании научных документов // Источники науч.-техн. информации и их аналитико-синтетическая обработка. – М., 1974. – С. 25–45. 3. Hahn U., Mani I. The Challenges of Automatic Summarization // Computer. – 2000. – Vol. 33. – № 11. – P. 29–36. 4. Шаховська Н. Інформаційна система реферування множини документів, поданих у різних форматах, базована на онтології / Н. Шаховська, В. Литвин, В. Крайовський // Вісник Нац. ун-ту "Львівська політехніка". – 2010. – № 672: Комп'ютерні науки та інформаційні технології. – С. 63–71. 5. Данилюк І.Г. Технологія автоматичного визначення тематики тексту // Лінгвістичні студії. – Вип. 17. – С. 290–293.