УДК 681.518

О. В. ПОМОРОВА,
Л. О. КОВТУН,
В. Б. БЕЛЗА
Хмельницький національний університет

МОДУЛЬ ОБРОБКИ ТЕКСТОВИХ ДОКУМЕНТІВ АВТОМАТИЗОВАНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ ЗБЕРІГАННЯ ТА ОПРАЦЮВАННЯ НАУКОВИХ РОБІТ

У статті описана модель автоматизованої інформаційної системи із модулем обробки текстових документів, що дозволяє привести документи до єдиного стандартного вигляду за допомогою зовнішньої частини системи, проаналізувати і трансформувати їх для подальшого запису в базу даних. Результатом автоматичної обробки текстів є виділення з них наступних основних блоків: прізвище, ім'я та по батькові наукових діячів, назва статті, ключові слова, список використаних джерел та основний текст. На основі отриманих даних відбувається формування наукових колективів.

Ключові слова: інформаційна система, модуль обробки текстових документів, база даних.

O. POMOROVA,
L. KOVTUN,
V. BELZA
Khmelnytsky National University

MODULE OF PROCESSING TEXT DOCUMENTS OF THE AUTOMATED INFORMATION SYSTEM OF STORAGE AND WORKING OF SCIENTIFIC WORKS

The article describes a model of an automated information system with a text document processing module, which allows you to bring documents to a single standard appearance with the help of the external part of the system, analyze and transform them for further recording in the database. the result of the automatic processing of texts is the selection of the following main blocks: the surname, name and patronymic of the scholars, the title of the article, the keywords, the list of sources used and the main text. Implementation of the information system on the basis of the proposed model enables the creation of a repository of data on the activities of scientists, regardless of the place of residence and format of the publication of publications; creation of a model information system for the storage and processing of information and research staff; creating software for automatic text processing. The use of the proposed information system in the scientific community makes it possible to obtain the correct detailed information and needs of research staff; receiving lists of potential research groups; obtaining information about scientists on relevant topics.

Keywords: information system, text document processing module, database.

ВСТУП. На сьогоднішній день наукова спільнота стикається з такими проблемами, як відсутність актуальної інформації про наукові праці студентів та їх керівників; наявність великої кількості недостовірної інформації про студентів та наукових співробітників і їхні праці в мережі інтернет; великі трудовитрати, необхідні для перевірки істинності інформації; відсутність єдиного сховища даних про наукових діячів та автоматизованого введення; наявність великої кількості наукових статей, що зберігаються на різнорідних носіях в різних форматах; відсутність довіри наукових діячів до ресурсів подібного роду; відсутність спеціалізованого автоматизованого пошуку по науковим діячам; відсутність алгоритмів для вирішення поставлених завдань і підбору наукових колективів; територіальна розрізненість наукових установ [1–5].

ОСНОВНА ЧАСТИНА. Пропонована модель інформаційної системи зберігання та обробки наукових праць наукових діячів робить спробу вирішити більшість з представлених вище проблем [1–5]. Система, реалізована на основі запропонованої моделі, дозволяє в автоматизованому режимі здійснювати пошук опублікованих наукових робіт в мережі Інтернет, обробляти виявлену інформацію з наукових журналів інтелектуальним текстовим пошуком по документам, збирати дані про наукових діячів, акумулювати інформацію в базі даних, виконувати різні перевірки на достовірність наданої інформації.

Веб-інтерфейс створеної інформаційної системи містить вкладки з повним списком авторів наукових праць в алфавітному порядку, форму для підбору наукових колективів, контактну інформацію та загальну інформацію про ресурс. У вкладці, що містить список всіх наукових співробітників, реалізована можливість пошуку за прізвищем автора наукових статей. Вкладка з формою для підбору наукового колективу дозволяє створювати списки наукових діячів, відповідно до заданих користувачем параметрами. Ця функція дозволяє вибирати авторів, які займаються схожими завданнями, посилаються на однакову літературу, мають однакові ключові слова, мають спільні праці, що володіють обраним науковим ступенем або/і володіють обраним вченим званням.

Результатом вибірки за вказаними параметрами є список наукових діячів, які можуть скласти науковий колектив за рішенням деякої проблеми. Список може бути імпортований в окремий файл для ручної обробки та інших дій.

У даній моделі інформаційної системи здійснюється ручна подача документів на вхід програми з обробки та аналізу текстів. В алгоритмі роботи аналізатора враховуються основні використовувані формати текстових документів: .docx, .pdf, .txt, .html; і графічний формат .jpeg. Всі типи документів зводяться до стандартного вигляду за допомогою зовнішньої частини системи, а далі аналізуються і трансформуються для подальшого запису в базу даних.

Інформаційна система зберігання і обробки властивостей наукових праць забезпечує кілька рівнів взаємодії з користувачем і адміністратором: внесення інформації в сховище, перегляд інформації, формування наукових колективів.

Внесення основного обсягу інформації про наукових співробітників та студентів здійснюється в автоматичному режимі за допомогою пошукової системи, спрямованої на пошук і обробку різних веб-ресурсів, що містять інформацію про наукових діячів. У разі якщо інформація є неповною або некоректною, автор має право звернутися до адміністратора за внесенням доповнень або виправлень в дані, що містяться в сховищі.

Для виправлення або доповнення інформації користувач повинен надіслати документ в будь-якому із запропонованих форматів, що підтверджує коректність виправлень. Документ повинен бути оформлений за ГОСТ і містити всю необхідну інформацію про наукового співробітника, в тому числі: прізвище, ім’я та по батькові; назва статті; ключові слова; використана література (оформлена за ГОСТ). Шаблон текстового документа, який вноситься у сховище інформації, наведено на рис. 1.

Інформаційна система дозволяє проглядати весь обсяг внесеної в сховище інформації за допомогою веб-ресурсу, доступного всім без винятку користувачам мережі Інтернет. Для перегляду інформації про конкретного наукового діяча необхідно мати мінімальні відомості про його діяльність: прізвище, тематика наукових досліджень або назва статті. За допомогою пошуку здійснюється запит до сховища даних, в якому проводиться порівняння на максимальну відповідність заявленим параметрам. Результати пошуку виводяться в таблиці на сторінці результатів пошуку.

В інформаційній системі зберігання і обробки властивостей наукових праць реалізована можливість формування наукових колективів. Для забезпечення максимально коректних списків наукових діячів, форма для формування списку має наступні додаткові параметри: наявність спільних праць; наявність посилань на однакову літературу; наявність перехресних посилань; схожі ключові слова; вчений ступінь; вчене звання.

Інформаційна система зберігання і обробки властивостей наукових праць дозволяє в автоматичному режимі обробляти документи різних форматів.

Спочатку документи можуть подаватися системі в форматах .HTML, .TXT, .DOCX, .PDF, .JPEG. Документи приводяться до єдиного формату за допомогою програми для обробки документів. Далі проводиться інтелектуальна обробка тексту, яка знаходить всю можливу інформацію про наукові праці та його автора і формує кортежі на вхід процесу ETL. ETL (від англ. Extract, Transform, Load – дослівно «вилучення, перетворення, завантаження») – один з основних процесів в управлінні сховищами даних, який включає в себе: вилучення даних із зовнішніх джерел; їх трансформація і очищення, щоб вони відповідали потребам відповідної моделі; завантаження їх в сховище даних.

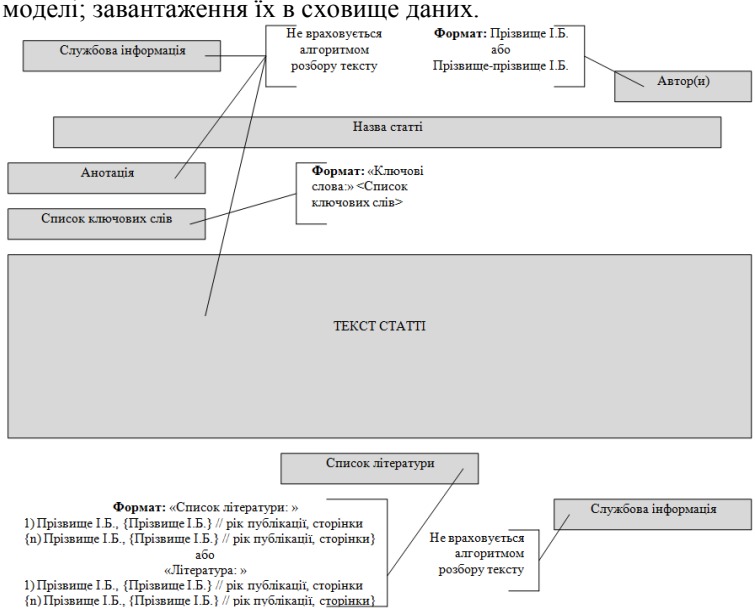

Рис. 1. Шаблон текстового документа, який вноситься у сховище інформації

Граничними умовами для документів є:

- файли повинні завантажуватися з перевірених джерел;
- файли повинні містити стандартну структуру, затверджену ГОСТ (рис. 1) .

Згідно алгоритму модуль автоматичної обробки текстів приймає на вхід стандартну статтю і виділяє дані для формування кортежів. У випадку наукового журналу, назва журналу зчитується і готується на вхід автоматичного довідника «Напрями дослідження». Дані про назву журналу не перевіряються на коректність введення в сховище. Надалі довідник може бути відредагований в ручному режимі в разі звернення користувача.

Система виявляє рядки, які не закінчуються крапкою, і вважає їх за назву статті. Система виконує перевірку на наявність односимвольних конструкцій. У разі входження даних конструкцій в назву, односимвольні змінні відрізаються разом з наступним словом або парою слів, розділених символом «-».

Конструкція з ініціалами автора наукової праці перевіряється на наявність будь-яких символів, крім букв, точок і дефісів. У разі входження інших символів, крім перерахованих, інформація вважається некоректною і повністю скидається.

Текст наукової статті зчитується в порожній пристрій до першого входження словосполучення «Ключові слова:».

Подальші слова і словосполучення списуються в окремий масив, пропускаючи пробіли і коми до першого входження символу «.». Всі слова перевіряються на входження сторонніх символів, цифр і невірного кодування. У разі появи цих символів весь текст статті вважається недостовірним і скидається.

Далі стаття зчитується до входження словосполучення «Література <новий рядок> 1».

У разі, якщо система не виявила даний абзац, дані вважаються некоректними і стаття повністю скидається. Список літератури формується в окремий багатовимірний масив, в якому окремо виділяються:

- ініціали автора;
- назва наукової праці;
- використовувані сторінки;
- дата.

Посилання зчитується після кортежу «<число>. <пробіл>» або«<Число>. <Пробіл>». Після оголошення нумерації по стандарту ГОСТ слідують прізвища авторів наукової праці.

Назва наукової роботи зчитується до подвійного входження символу «/» і записується в поле «Назва статті» кожному автору. Після подвійного символу інформація відсікається до входження числа в форматі «YYYY», яке порівнюється з нижньою межею «1800» і заноситься в масив з науковими діячами. Далі рядок зчитується з символу «новий рядок» до пропуску і перевіряється на формат «<число>» або «<число1> - <число2>». Числа в разі потрапляння у діапазон значень перевіряються на відповідність умові <число1> ≤ <число2>. У разі невдачі дані вважаються некоректними і стаття повністю скидається. Дані записуються в стовпець «Сторінки» з інформацією про наукових діячів.

Масив значень готовий для запису в таблиці «Посилання» і «Автори». У разі запуску транзакції запису в сховище, значення масиву перевіряються на дублікати, що містяться в базі даних. При знаходженні дубліката, рядок скидається, і система починає обробку наступного рядка масиву.

Текстові файли, отримані в одному з форматів (.TXT, .DOCX, .PDF, .JPEG, .HTML), обробляються за допомогою автоматичного алгоритму обробки тексту. Загальна схема роботи алгоритму вказана на рис. 2:

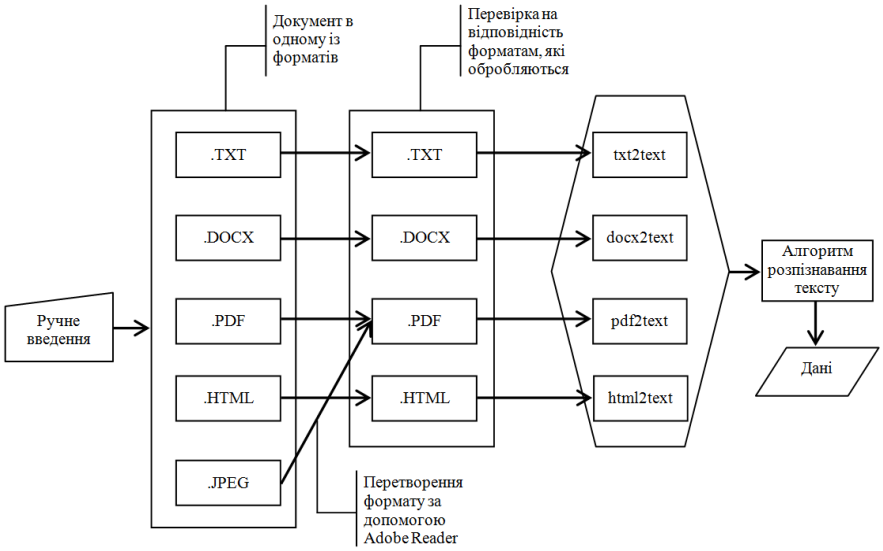

Рис. 2. Загальна схема роботи автоматичного алгоритму обробки тексту

Всі отримані документи перетворюються до єдиного формату .RTF за допомогою функцій <формат_документа> 2text. Отриманий файл з кодом декодується за допомогою стандартних функцій.

В інформаційній системі зберігання і обробки властивостей наукових праць формуються кортежі даних на запис в інформаційне сховище. Інформація повинна записуватися в строго визначеному порядку для запобігання порушенню цілісності даних. Схема алгоритму обробки тексту вказана на рис. 3.

Декодування тексту за наведеним алгоритмом працює наступним чином:

1) на вхід системи подається документ;

2) система формує масив @array_main:string(1000);

3) система обробляє перші 1000 символів;

4) система віднаходить перше входження <A'.'A'.'A{a}{-A{a}}> і записує в блоки AUTHOR_PREV.SURNAME, AUTHOR_PREV.NAME, AUTHOR_PREV.SECONDNAME;

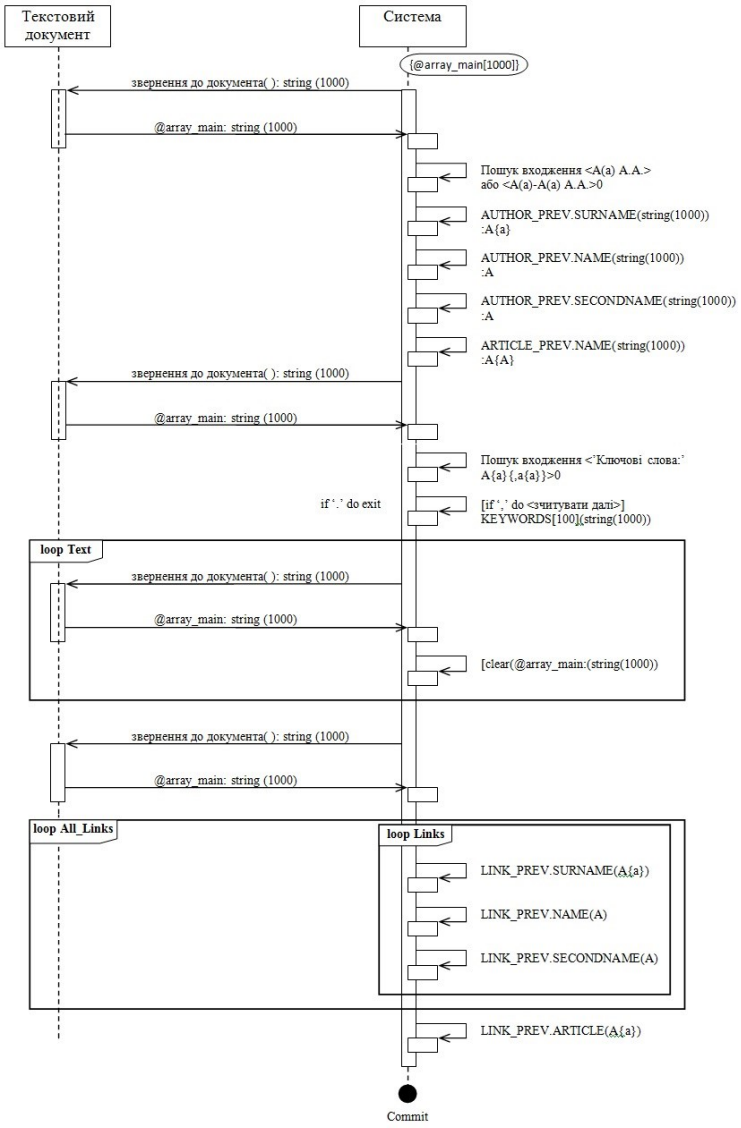

Рис. 3. Алгоритм декодування тексту

5) система віднаходить попередні входження типу <A{A}> і записує в блок ARTICLE_PREV.NAME, замість символів переводу рядка символами ' ';

6) останнім символом в масиві є ' '. Він буде видалений функціями LTRIM(RTRIM(ARTICLE_PREV.NAME)) при запису у сховище даних.

7) система віднаходить перше входження <'Ключові слова: ' a{a}{,a{a}}.> і записує у рядковий масив KEYWORDS[100]. Передбачається, що наукова робота не може містити більше 100 ключових слів.

8) слова зчитуються:

- від символу ':' до символу ',' - в разі першого з багатьох ключових слів;
- від символу ':' до символу '.' - в разі першого з одного ключових слів;

- від символу ',' до символу ',' - в разі центрового елемента з багатьох ключових слів;
- від символу ',' до символу '.' - в разі останнього з багатьох ключових слів.

9) система видаляє символи '-\n' і замінює символ кінця рядка символом ' ':

10) у випадку, якщо входження <'Ключові слова: '> не знайдено, масив @array_main очищається і повторюються пункти 3, 7-9.

11) система очищує масив @array_main. Передбачається, що наукова публікація не може містити менше 2000 символів;

12) система зчитує наступні 1000 символів у масив;

13) система віднаходить входження типу <'Література'\n> або <'Список літератури'\n>.

14) система записує входження <{число$_1$'. '{A{a}{-A{a}}' 'A'.'A'.'}' '{A{a}}'//'символи'. 'число$_2$'. 'число$_3$'. 'число$_4$'-'число$_5$'.'}> в блоки:

```
WHILE {A{a}' 'A'.'A'.'} DO
FOR i = 1 TO n DO
LINK_PREV.SURNAME[n] = <A{a}>;
LINK_PREV.NAME[n] = <A>;
LINK_PREV.SECONDNAME[n] = <A>;
WHILE NOT <'//'> DO
FOR i = 1 TO n DO
LINK_PREV.ARTICLE[n] = <A{a}>;
LINK_PREV.DATE[n] = <число3>;
LINK_PREV.PAGES[n] = <число4'-'число5>;
```

15) система повторює пункт 14 доки не віднайде кінець документа або наступне входження, яке не відповідає шаблону;

16) в результаті виконання пункту 15 алгоритму, система формує наступні масиви значень;

17) система пропонує адміністратору перевірити внесені зміни і підтвердити завантаження в сховище даних.

18) адміністратор вносить необхідні правки і підтверджує процес завантаження.

19) далі система формує кортежі зі значень масивів і відбувається підсумковий запит до бази даних.

ВИСНОВКИ. На основі досліджених функціональних вимог була створена модельна інформаційна система зберігання і обробки властивостей наукових праць. Інформаційна система володіє наступними перевагами: створено алгоритм автоматичної обробки текстів та підбору наукових колективів; розроблено основні функції інформаційної системи; існує можливість збору статистики; створено алгоритм автоматичного пошуку по ресурсів мережі Інтернет. Реалізація інформаційної системи на основі запропонованої моделі дає можливість створення сховища даних про діяльність наукових діячів незалежно від місця проживання і формату зберігання публікацій; створення модельної інформаційної системи зберігання і обробки інформації про наукових діячів; створення програмного забезпечення для автоматичної обробки тексту.

Використання запропонованої інформаційної системи в науковому співтоваристві дає можливість отримання коректної детальної інформації про потреби наукових колективів; отримання списків потенційних наукових колективів; отримання інформації про наукових діячів за відповідною тематикою.

Література

1. Береза А. М. Основи створення інформаційних систем : навч. посібник / А. М. Береза. – 2 видання, перероблене і доп. – К. : КНЕУ, 2001. – 134 с.
2. Соколов В. Ю. Інформаційні системи і технології : навч. Посіб. / В. Ю. Соколов. – К. : ДУІКТ, 2010. – 138 с.
3. Грицунов О. В. Інформаційні системи та технології : навч. посібник / О. В. Грицунов. – Х. : ХНАМГ, 2010. – 222 с.
4. Басюк Т. М. Методи та засоби мультимедійних інформаційних систем : навч. посіб. / Т. М. Басюк, П. І. Жежнич. – Львів. : Львів. політехніка, 2015. – 426 с.
5. Гайдамакин Н. А. Автоматизированные системы, базы и банки данных. Вводный курс : учебное пособие / Н. А. Гайдамакин. – М. : Гелиос АРВ, 2002. – 368 с.

References

1. Bereza A. M. Osnovy stvorennia informatsiinykh system : navch. posibnyk / A. M. Bereza. – 2 vydannia, pereroblene i dop. – K. : KNEU, 2001. – 134 s.
2. Sokolov V. Yu. Informatsiini systemy i tekhnolohii : navch. Posib. / V. Yu. Sokolov. – K. : DUIKT, 2010. – 138 s.
3. Hrytsunov O. V. Informatsiini systemy ta tekhnolohii : navch. posibnyk / O. V. Hrytsunov. – Kh. : KhNAMH, 2010. – 222 s.
4. Basiuk T. M. Metody ta zasoby multymediinykh informatsiinykh system : navch. posib. / T. M. Basiuk, P. I. Zhezhnych. – Lviv. : Lviv. politekhnika, 2015. – 426 s.
5. Gaydamakin N. A. Avtomatizirovannyie sistemyi, bazyi i banki dannyih. Vvodnyiy kurs : uchebnoe posobie / N. A. Gaydamakin. – M. : Gelios ARV, 2002. – 368 s.