УДК 004.94:519.876.5

МОДЕЛЮВАННЯ СИСТЕМИ РОЗПІЗНАВАННЯ ТА АНАЛІЗУ ТЕКСТОВИХ ДАНИХ

І. А. Пількевич
Доктор технічних наук, професор, завідувач кафедри*
E-mail: igor.pilkevich@mail.ru
Н. М. Лобанчикова
Кандидат технічних наук, доцент**
E-mail: lobanchikovanm@rambler.ru
І. В. Шульга
Кандидат сільськогосподарських наук, доцент*
E-mail: eko_univer@i.ua
*Кафедра моніторингу навколишнього природного середовища
Житомирський національний агроекологічний університет
бульвар Старий, 7, м. Житомир, Україна, 10008
Р. С. Лазюта**
E-mail: roman.91@mail.ru
**Кафедра безпеки інформаційних і комунікаційних систем
Житомирський військовий інститут ім. С.П. Корольова Національного авіаційного університету
пр. Миру, 22, м. Житомир, Україна, 10004

Проведено аналіз нормативно-правового забезпечення захисту інформації, який дозволив виділити класи конфіденційності інформації. Розроблено автоматизовану систему виявлення конфіденційної інформації в текстових документах, а також створено програмний продукт, який дозволяє завантажувати текст або текстовий документ для подальшого аналізу та визначення ступеня його важливості

Ключові слова: інформаційно-комунікаційна система, база даних, конфіденційна інформація, класифікація документів, захист інформації

Проведен анализ нормативно-правового обеспечения защиты информации, позволяющий выделить классы конфиденциальности информации. Разработана автоматизированная система выявления конфиденциальности информации в текстовых документах, а также создан программный продукт, позволяющий загружать текст или текстовый документ для дальнейшего анализа и определения степени его важности

Ключевые слова: информационно-коммуникационная система, база данных, конфиденциальная информация, классификация документов, защита информации

1. Вступ

Сучасні новітні технології зумовили активний розвиток системи електронних інформаційних ресурсів. Створення великих обсягів сховищ електронних даних в інформаційно-комунікаційних системах призвели до необхідності їх аналізу та детальної обробки. Одним із напрямків обробки інформації є розпізнавання та виявлення конфіденційної інформації в текстових документах. Для визначення виду інформації необхідним є класифікація документів за їх важливістю, і, в залежності від встановленого ступеня важливості, визначення її доступності відповідним категоріям користувачів.

Системи розпізнавання текстових даних зайняли чільне місце в системах захисту інформаційно-комунікаційних систем. Численні роботи програмістів призвели до виникнення спеціальних програмних продуктів та засобів, що доступні на сьогодні людству [1 – 4].

Метою роботи є розробка автоматизованої системи виявлення конфіденційної інформації в текстових документах для автоматизації роботи експертів щодо розпізнавання та аналізу текстових даних.

Для досягнення зазначеної мети необхідним є вирішення наступних взаємозалежних технічних задач:

– аналіз нормативно-правового регулювання оцінки інформації з обмеженим доступом;

– аналіз технологій створення систем аналізу текстових даних;

– дослідження засобів створення інтелектуальних систем розпізнавання та аналізу тексту;

– реалізація системи.

Об'єктом дослідження виступають процеси побудови автоматизованих систем розпізнавання та аналізу текстових даних.

Предметом дослідження є моделі, методи та засоби побудови систем розпізнавання та аналізу текстових даних з метою виявлення конфіденційної інформації.

2. Аналіз нормативно-правового забезпечення захисту інформації та постановка задачі

Інформаційне законодавство України у 2011 році було оновлено з прийняттям Закону України „Про

доступ до публічної інформації" та нової редакції Закону України „Про інформацію", а також були підписані Укази Президента № 547 „Питання забезпечення органами виконавчої влади доступу до публічної інформації" та № 548 „Про першочергові заходи щодо забезпечення доступу до публічної інформації в допоміжних органах, створених Президентом України" [5 – 9].

Нова редакція Закону України „Про інформацію", як базового нормативно-правового акту в інформаційній сфері, надає нове визначення інформації та передбачає поділ за змістом інформації на такі види: інформація про фізичну особу; інформація довідково-енциклопедичного характеру; інформація про стан довкілля (екологічна інформація); інформація про товар (роботу, послугу); науково-технічна інформація; податкова інформація; правова інформація; статистична інформація; соціологічна інформація та інші види інформації.

Встановлено, що інформація про фізичну особу (персональні дані) – це відомості чи сукупність відомостей про фізичну особу, яка ідентифікована або може бути конкретно ідентифікована.

За порядком доступу інформація поділяється на відкриту та інформацію з обмеженим доступом. Відкритою вважається вся інформація, крім тієї, що віднесена законом до інформації з обмеженим доступом. Інформацією з обмеженим доступом є конфіденційна, таємна та службова інформації (рис. 1) [10, 11].

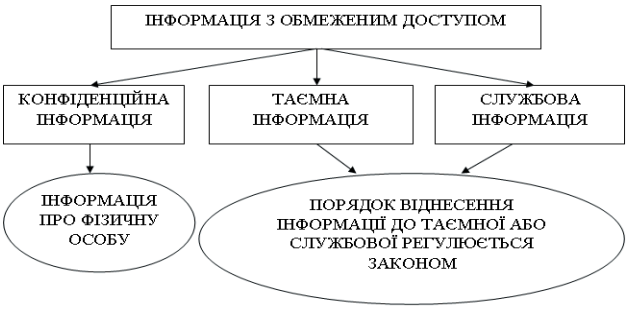


Рис. 1. Види інформації

Конфіденційною є інформація про фізичну особу, а також інформація, доступ до якої обмежено фізичною або юридичною особою, крім суб'єктів владних повноважень.

Конфіденційна інформація може поширюватися за бажанням (згодою) відповідної особи у визначеному нею порядку відповідно до передбачених нею умов, а також в інших випадках, визначених законом. Крім того, в новій редакції Закону чітко прописані права журналістів та питання їх акредитації [5].

Основною метою Закону України „Про доступ до публічної інформації" є створення механізму реалізації права кожного на доступ до публічної інформації. Він містить перелік гарантій дотримання прав на надання публічної інформації, поетапний порядок доступу до неї, надає визначення таким поняттям, як конфіденційна, таємна та службова інформації, регламентує порядок її отримання, визначає порядок і строки подачі та задоволення запиту на інформацію, а також процедуру оскарження рішень, дій чи бездіяльності розпорядників інформації.

Нижче наведено перелік кроків, які необхідно здійснити при побудові системи захисту інформації (в порядку пріоритету) [12]:

– створити багаторівневу систему доступу для роботи з корпоративною інформацією;

– забезпечити резервування інформації корпоративних файл-серверів;

– захистити від витоку інформацію з використанням віддаленого доступу, використовуючи систему паролів на кожному конкретному комп'ютері й в кожній локальній мережі;

– забезпечити безпеку серверного приміщення та резервних копій інформації файл-серверів;

– контролювати Інтернет-трафік, поштовий трафік;

– впровадити систему контролю за місцезнаходженням співробітників та відвідувачів у приміщеннях, де є доступ до інформаційних ресурсів, використовуючи при цьому складну та багаторівневу систему доступу до таких приміщень.

Послідовна (повна) реалізація вищевказаних кроків дозволить отримати на виході повноцінну систему захисту інформації [13].

Задачею роботи є дослідження процесів побудови та створення програмного продукту, який би дозволяв завантажувати текст або текстовий документ для подальшого аналізу та визначення ступеню його важливості. Виходячи з цього можна буде проаналізувати тематику документу, що обробляється, та присвоїти йому відповідний гриф важливості. Від цього залежатиме ступінь доступу до документа.

Середовищем розробки програмного продукту було обрано Microsoft Visual C# 2010. Такий вибір можна обґрунтувати тим, що дане програмне забезпечення (ПЗ) має такі переваги в порівнянні з іншими ПЗ [14]:

– велика сумісність з усіма версіями операційних систем Windows;

– наявність великої бібліотеки готових класів;

– дане ПЗ є складовою об'єктно-орієнтованого програмування.

3. Основна частина

Продукт, що представлений в роботі, по суті являє собою аналізатор тексту, який певним чином визначає ту чи іншу важливість документу/тексту.

На рис. 2 представлена типова структурна схема роботи аналізатора тексту.

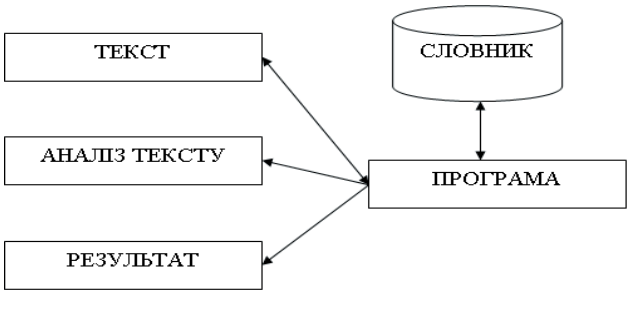


Рис. 2. Типова структура аналізатора тексту

Програмний продукт функціонує наступним чином.

Після запуску головного вікна, програма повинна вибрати будь-який текстовий документ. Наступним етапом є підключення файлу словника, в якому зберігаються слова, що являють собою, скажімо так, „небезпеку".

Після того, як було обрано текст для обробки та словник, можна безпосередньо приступати до аналізу.

Текст аналізується доволі швидко, що дозволяє оперувати великими об'ємами даних. Як тільки аналіз буде завершено, програма автоматично видає ступінь конфіденційності текстових даних.

Програма складається з наступних основних елементів:

– головне вікно;

– функціональні кнопки;

– документ, що потребує аналізу;

– документ-словник, призначений для того, щоб звірити: чи є у тексті можливі небезпечні словоутворення.

Головне вікно програми містить у собі такі кнопки, як: „Підключити", „Відкрити", „Аналізувати". Кнопка „Підключити" відповідає безпосередньо за підключення (підвантаження) файлу-словника, згідно з яким буде звірений документ. Кнопка „Відкрити" текст призначена для відкриття текстового документу з можливістю його вибору у стандартному вікні Windows для вибору документів. При натисканні кнопки „Підключити" відбувається підключення до програми бази даних словника. При спрацьовуванні кнопки „Аналізувати" безпосередньо відбувається аналіз тексту або документу та визначення ступеня його важливості.

3.1. Алгоритм роботи програмного продукту

На рис. 3 наведено загальний алгоритм функціонування програмного продукту. На даному структурному алгоритмі зображений загальний принцип роботи програмного додатку. Розглянемо кожен з блоків детальніше.

Блок „Введення даних". Програма не буде працювати, якщо поле, що призначене для його заповнювання текстом, буде пустим. Отже, перед тим, як почати роботу, треба визначити текст, з яким буде працювати програма. Лише після того, як текст буде обрано, а поле буде заповненим, програма приступає до наступного кроку.

Блок „Вибір словника". Після обрання тексту, необхідно вибрати словник (базу слів), з якою буде працювати програма.

В даному випадку файлом-словником може бути текстовий файл під назвою Dictionary. Програма в автоматичному режимі перевіряє/порівнює слова, які є у відкритому тексті, та слова, які знаходяться у словнику.

Ті слова, що збігаються, прийнято вважати важливими елементами тексту і ця інформація буде застосована для розрахунку ступеня конфіденційності документу.

Наступним блоком є „Блок прийняття рішення". Цей блок проводить перевірку: чи всі елементи, що потрібні для початку роботи програми, підключені. Якщо хоча б один з них не задіяний, то програма видає помилку та не буде працювати поки всі відповідні поля не будуть заповнені.

Блок „Аналіз даних". В ньому відбувається порівняння слів тексту та словника, а також проходить аналіз тексту.

Блок „Відсоткове відношення важливості даних". Саме в цьому блоці визначається ступінь важливості тексту, яка у відсотковому відношенні розраховується за допомогою формули:

$$R = {}^{N_{\text{сп. слів}}}\!/_{N_{\text{слів}}} \cdot 100 ,$$

де $N_{\text{сп. слів}}$ – кількість слів, що співпали (слова є і у тексті і у словнику);

$N_{\text{слів}}$ – загальна кількість слів у документі;

R – результат у відсотках.

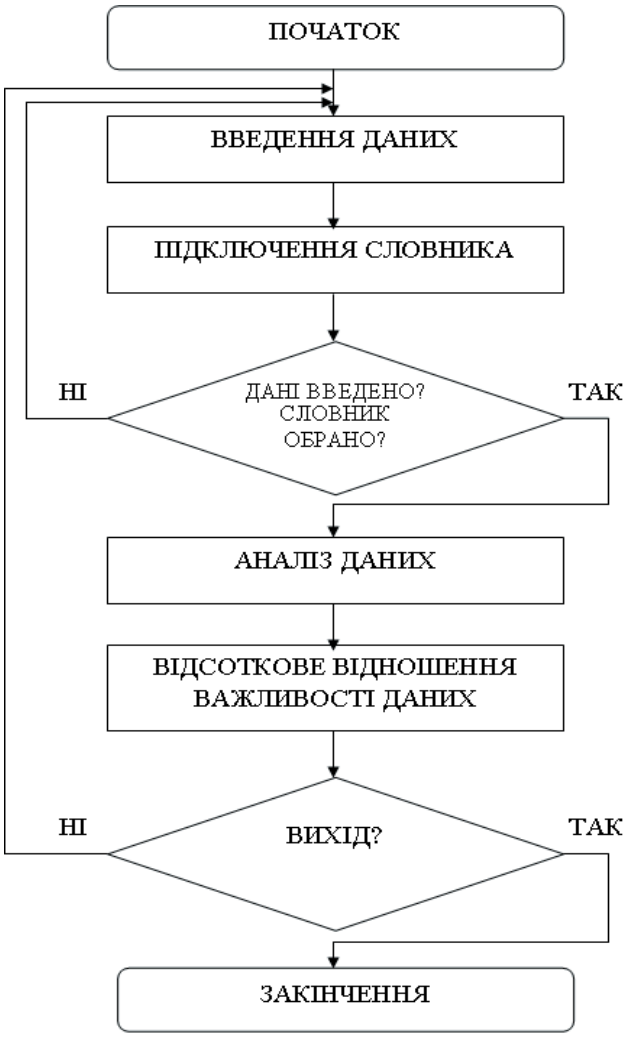


Рис. 3. Загальний алгоритм роботи програми

Після того, як було визначено ступінь конфіденційності, у користувача є вибір: продовжувати роботу з програмою або вийти з неї. Для спрощення роботи користувачу в програмі є функція очистки всіх полів, яка видаляє всю інформацію, що підверглася аналізу.

На рис. 4 представлено алгоритм, що розроблений в роботі, але у вигляді функціональних блоків.

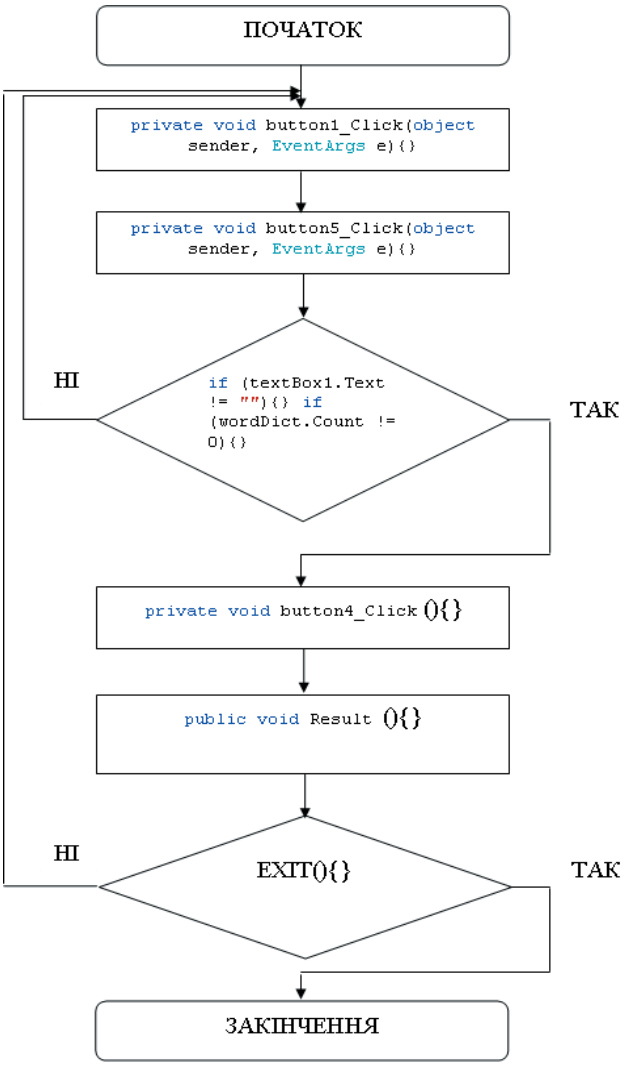
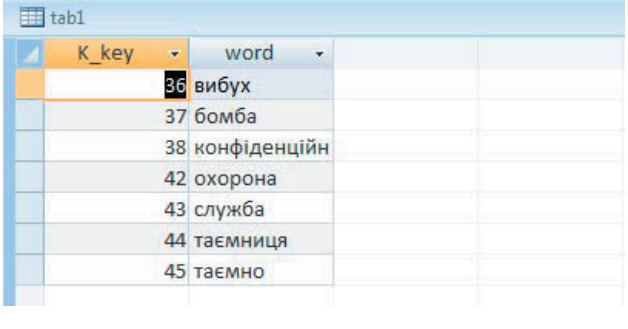


Рис. 4. Функціональний алгоритм роботи програми

Функціональний алгоритм, на відміну від структурного, відбиває структуру спрацьовування програмного коду, та послідовність дій, які відбуваються при роботі програми. По суті це той самий порядок, але з точки зору програмної реалізації.

3.2. Розробка словника

В програмному продукті було використано та застосовано електронний словник. Він представляє з себе файл бази даних, в якому містяться певні вирази, що необхідно виявити при аналізі текстового документу (рис. 5).


tab1

K_key	word
36	вибух
37	бомба
38	конфіденційн
42	охорона
43	служба
44	таємниця
45	таємно

Рис. 5. Слова, що зберігаються у словнику

Простіше кажучи, словник – це сукупність певних слів, що містяться у базі даних, яка створена за допомогою програмного забезпечення Microsoft Office Access 2007. В даний файл можна вільно заносити нові слова або вилучати ті, що вже не потрібні. Якщо база даних буде пуста (в ній не буде слів), то програма обов'язково повідомить про це користувача.

3.3. Інтерфейс програмного продукту

Даний програмний продукт було створено за допомогою програмного забезпечення Microsoft Visual Studio. Мовою програмування було обрано C# [15]. Головна форма інтерфейсу системи розпізнавання та виявлення конфіденційної інформації у текстових документах представлена на рис. 6.

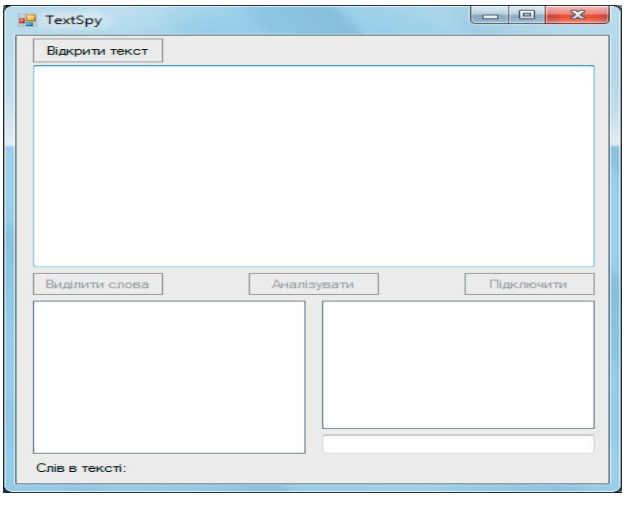


Рис. 6. Головне вікно програми

Аналіз рис. 6 показує, що на початку роботи користувачу доступна лише одна кнопка „Відкрити текст". Якщо за допомогою даної кнопки текст не буде відкритий, то решта кнопок залишаються неактивними (ніяких дій не можна буде зробити).

При натисканні на кнопку „Відкрити текст" спрацьовує наступний програмний код, який відкриває діалогове вікно:

```
private void button1_Click(object sender, EventArgs e)
        {
            if (openFileDialog1.ShowDialog() ==
System.Windows.Forms.DialogResult.OK)
            {
                FileStream fs = new FileStream(openFileDialog1.FileName,
FileMode.Open, FileAccess.Read);
                StreamReader sr = new StreamReader(fs, Encoding.Default);
                textBox1.Text = sr.ReadToEnd();
                text = textBox1.Text;
                sr.Close();
                fs.Close();
            }
            button2.Enabled = true;
            button3.Enabled = true;
            button4.Enabled = true;
            button5.Enabled = true;
            button6.Enabled = true;
    }
```

Отже, до запуску програми всі кнопки, окрім „Відкрити текст" – неактивні. Після того, як документ буде відкрито, кнопки стають активними і можна продовжувати роботу (рис. 7).

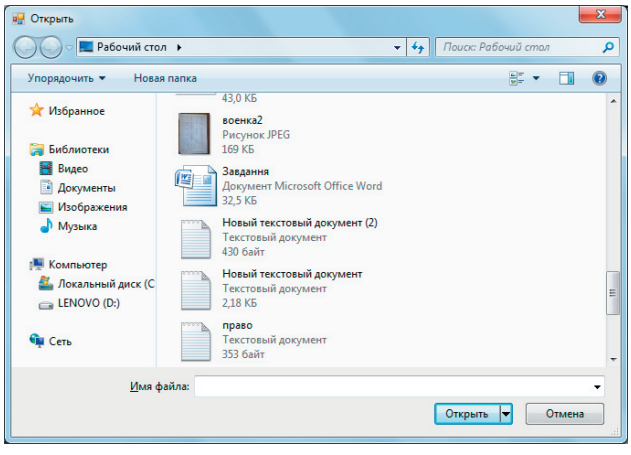


Рис. 7. Вікно вибору текстового документу, що належить аналізу

За допомогою даного вікна можна вибрати документ, що підлягає аналізу та виявленню конфіденційної інформації.

3.4. Тестування програмного продукту

Наступний крок – аналіз документу. Після того, як файл обрано, текст документу, що аналізується, автоматично завантажується у середовище програми для подальшої обробки. На рис. 8 наведено вигляд тексту в текстовому редакторі, а на рис. 9 – у вікні створеної програми.

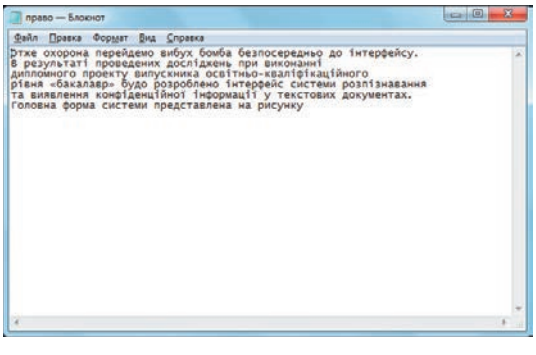


Рис. 8. Текст у текстовому редакторі Блокнот

Якщо текст завантажено вірно, то програма приступає до його обробки. Для цього, за допомогою кнопки „Виділити слова”, необхідно виділити всі слова, що є у тексті.

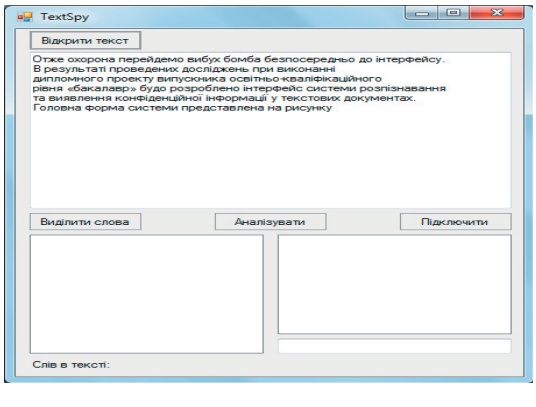


Рис. 9. Відкрити текст у створеному програмному продукті

При натисканні кнопки „Виділити слова” відбуваються наступні дії:

```
private void button3_Click(object sender, EventArgs e)
    {
      listBox2.Items.Clear();
      wordText.Clear();
      string[] masText = text.Split(new Char[] { ' ','.',',',';',':','-' });
      foreach (string res in masText)
      {
        if (res != "")
        {
          wordText.Add(res);
        }
      }
      for (int i = 0; i < wordText.Count; i++)
      {
        listBox2.Items.Add(wordText[i]);
      }
      label1.Text = "Слів в тексті: " + wordText.Count.ToString();
    }
```

Поле для заповнення словами попередньо очищується.

Весь текст, що міститься у документі, представляється у вигляді масиву: кожне слово це окремий елемент масиву; а такі знаки, як: пробіли, крапки, коми та інші знаки пунктуації – це певного роду границі кожного елементу.

Тобто, програма сприймає текст як набір слів/елементів масиву і заповнює поле цими словами, а також автоматично веде їх підрахунок. Наступним кроком є підключення словника створеної бази даних.

Для цього необхідно натиснути на кнопку „Підключити”.

Як видно з рис. 10, всі слова, що виділені, розташовані у вигляді списку. Нижче вказана загальна кількість слів у тексті.

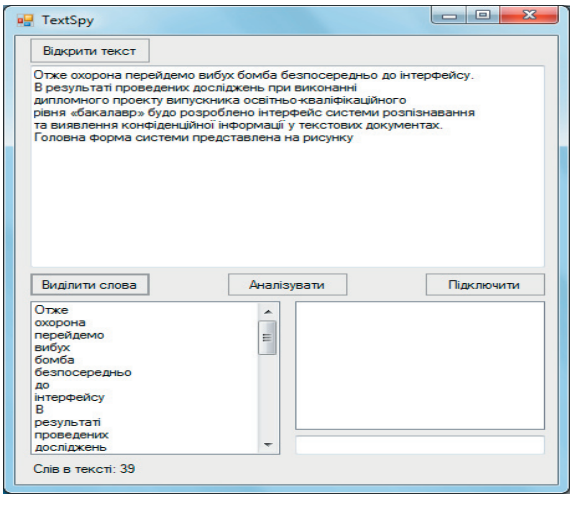


Рис. 10. Вилучення слів з текстового документа

При натисканні на кнопку „Підключити”, в програмі спрацьовує функція openDict() , яка видаляє старі дані з поля та попереднє з’єднання з базою даних, а також встановлює нове з’єднання з базою і вилучає з неї всі слова, що в ній знаходяться (рис. 11).

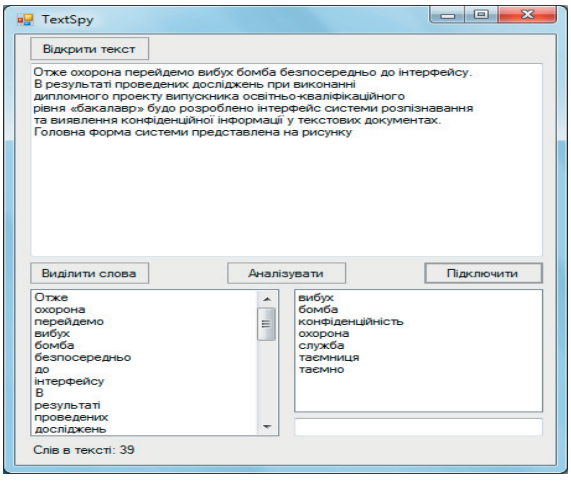


Рис. 11. Підключення бази даних

Для того, щоб проаналізувати текст, необхідно натиснути кнопку „Аналізувати”. Аналіз приведеного коду показує, що на початку аналізу програма перевіряє: чи правильно встановлено з'єднання зі базою даних; наявність відкритого текстового документа та виділених з нього слів. Якщо всі ці вимоги дотримуються, то програма починає порівнювати слова, що виділені, зі словами у словнику, які по суті таким же чином відхилені з бази даних. Тому слова порівнюються як елементи масиву. Якщо певні елементи співпадають, то дане слово слід віднести до роду небезпечних.

Якщо відсоткове відношення небезпечних слів у тексті не перевищує 5%, то з'являється повідомлення „Небезпечних слів 4%, текст безпечний, маловажливий”.

Далі все іде за аналогією: 5-15% – текст середньої важливості, 15-30% – текст важливий; більше 30% – текст представляє небезпеку, а його зміст має дуже важливе значення.

На рис. 12 представлено зовнішній вигляд вікна результату аналізу тексту.

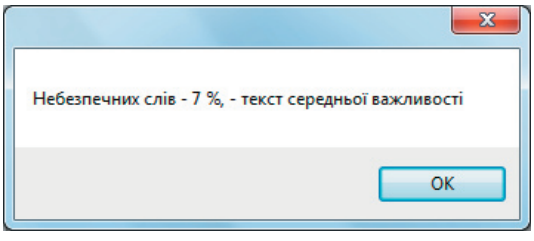


Рис. 12. Вікно „Результат аналізу тексту”

Неважко перевірити даний результат „вручну”. У відкритому тексті „приховано” три слова, що знаходяться у словнику, а це відповідає відсотковому відношенню важливості даних: $\frac{3}{39} \cdot 100 \approx 7\%$.

4. Висновки

Аналіз нормативно-правового забезпечення дозволив виділити такі класи конфіденційності інформації: конфіденційна; таємна; службова. Кожний з цих класів окремо представляє різні сфери діяльності, завдяки чому і відбувається поділ інформації.

Основою функціонування глобальної інформаційної пошукової системи є: гіпертекстові зв'язки; ключові слова; критерії смислової відповідності; активна роль людини в управлінні їх формою і в оцінці релевантності результатів пошуку запитом.

В запропонованій програмі здійснено процедури загального підрахунку всіх та небезпечних (ключових) слів, підключення бази даних та аналізу самого тексту з подальшою оцінкою його конфіденційності.

Подальшим напрямком дослідження є використання семантичних мереж у блоці аналізу та прийняття рішень.

Література

1. Chen, P. P. The Entity-Relationship Model: Toward a Unified View of Data [Text] / P. P. Chen // ACM Trans. On Database Syst. – 1976. – V.1, №1. – pp. 9-36.
2. Codd, E. F. A relational model of data large shared data banks [Text] / E. F. Codd // Comm. ACM. – 1970. – V.13, №6. – pp. 377-387.
3. Коннолли, Т. Базы данных. Проектирование, реализация и сопровождение. Теория и практика [Текст] : пер. с англ. / Т. Коннолли, К. Бегг. – 3-е изд. – М. : Изд. дом „Вильямс”, 2003. – 1440 с.
4. Дейт К. Дж. Введение в системы баз данных [Текст] : пер. с англ. / К. Дж. Дейт. – 8-е изд. – М. : Изд. дом „Вильямс”, 2005. – 1328 с.
5. Закон України „Про внесення змін до Закону України „Про інформацію” : станом на 13.01.2011 року [Текст] / Верховна Рада України. – Офіц. вид. – К. : Відомості Верховної Ради України. – 2011. – №10. – С. 21, стаття 356. – (Бібліотека офіційних видань).
6. Закон України „Про інформацію” : станом на 09.05.2011 року [Текст] / Верховна Рада України. – Офіц. вид. – К. : Відомості Верховної Ради України. – 1992. – №48. – Стаття 650. – (Бібліотека офіційних видань).
7. Закон України „Про захист інформації в інформаційно-комунікаційних системах” : станом на 30.04.2009 року [Текст] / Верховна Рада України. – Офіц. вид. – К. : Відомості Верховної Ради України. – 1994. – №31. – Стаття 286. – (Бібліотека офіційних видань).
8. Закон України „Про державну таємницю” : станом на 24.02.2011 року [Текст] / Верховна Рада України. – Офіц. вид. – К. : Відомості Верховної Ради України. – 1994. – №16. – С. 422, стаття 93. – (Бібліотека офіційних видань).

9. Кавун, С. В. Інформаційна безпека [Текст] : підручник / С. В. Кавун. – Харків : Вид. ХНЕУ, 2009. – 368 с.
10. Кавун, С. В. Механизм оценивания экономической эффективности системы экономической безопасности [Текст] / С. В. Кавун // Бизнес-информ. – 2009. – № 8. – С. 58-64.
11. Кавун, С. В. Классификатор видов информации и форм документов [Текст] / С.В. Кавун // Науковий вісник Полтавського університету споживчої кооперації України. Сер. Економічні науки: наук. журнал. – Полтава : РВВ ПУСКУ, 2009. – № 5(36). – С. 69-75.
12. Грибунин, В. Г. Комплексная система зашиты информации на предприятии [Текст] : учеб. пособие для студ. высш. учеб. заведений / В. Г. Грибунин, В. В. Чудовский. – М. : Изд. центр „Академия", 2009. – 416 с.
13. Корченко, О. Г. Системи захисту інформації [Текст] : монографія / О. Г. Корченко. – К. : НАУ, 2004. – 264 с.
14. Фролов, Л. В. Базы данных в Интернете [Текст] : практическое руководство по созданию Web-приложений с базами данных / Л. В. Фролов, Г. В. Фролов. – 2-ое изд., испр. – М. : Издательско-торговый дом „Русская редакция", 2000. – 448 с.
15. Троелсен, Э. С. и платформа. NET 3.0 [Текст] / Э. Троелсен. – 1-ое изд. – СПб. : Издательский дом „Питер", 2008. – 1456 с.

Запропоновано алгоритм цифрової фільтрації корисних сигналів різних форм від шумових складових з використанням прямого та зворотного FFT- перетворення. На основі запропонованого алгоритму розроблено програмний компонент для обробки цифрових сигналів «iFFT- Noise Gate»; наведено результати моделювання процесу фільтрації шумів та порівняльну характеристику шумознижуючих властивостей нового алгоритму та відомих аналогів за найбільш суттєвими параметрами: THD, SINAD, SNR

Ключові слова: подавлення шуму, фільтрація, цифрова обробка сигналів, спектр, алгоритм, програмне забезпечення

Предложен алгоритм цифровой фильтрации полезных сигналов разной формы от шумовых составляющих с использованием прямого и обратного FFT–преобразования. На основе предложенного алгоритма разработан программный компонент для обработки цифровых сигналов «iFFT- Noise Gate»; приведены результаты моделирования процесса фильтрации шумов и сравнительную характеристику шумоподавляющих свойств нового алгоритма и известных аналогов по наиболее существенным параметрам: THD, SINAD, SNR

Ключевые слова: подавление шума, фильтрация, цифровая обработка сигналов, спектр, алгоритм, программное обеспечение

УДК 519.688

ПРОГРАММНЫЙ КОМПОНЕНТ ДЛЯ ОБРАБОТКИ ЦИФРОВЫХ СИГНАЛОВ IFFT-NOISE GATE

Е. В. Семенов
Аспирант*
E-mail: j.semenov@mail.ru
В. В. Шведова
Кандидат технических наук, доцент*
E-mail: shvedova_victoria@ukr.net
*Кафедра информационно-измерительной техники
Национальный технический университет Украины
«Киевский политехнический институт»
пр. Победы, 37, г. Киев, Украина, 03056

1. Введение

В наши дни цифровая фильтрация приобрела широкую популярность в связи с широким использованием микроконтроллеров в разных сферах человеческой деятельности.

Очевидно и преимущество применения цифровой обработки сигнала наряду с аналоговым: улучшается помехозащищенность канала связи, бесконечные возможности кодирования информации [1]. Применение микропроцессоров в радиотехнических системах существенно улучшает их массогабаритные, технические и экономические показатели, открывает широкие возможности реализации сложных алгоритмов цифровой обработки сигналов. В состав современных вычислительных устройств, обрабатывающих информацию технологических процессов, часто входят блоки программно или аппаратно реализованных цифровых фильтров (ЦФ). По сравнению с аналоговыми фильтрами они предпочтительны во множестве областей (например, сжатие данных, биомедицинская обработка сигналов, обработка речи, обработка изображений, передача данных, цифровое аудио, телефонное эхо подавление) [2], так как обладают рядом преимуществ и недостатков, часть из которых описана ниже.