УДК 681.3

В.В. Кот,
V. Kot
О.М. М‘якшило, канд. техн. наук
O. Myakshilo

Застосування нейронних мереж для автоматизованої класифікації літератури у науково-технічній бібліотеці

Application of neural networks for automated classification of literature in scientific and technical library

В статті розглянуто метод інтелектуального аналізу тексту за допомогою нейронних мереж. Запропонований метод використано для ідентифікації літературних джерел, що надходять у бібліотеку, з метою їх класифікації і каталогізації. На основі запропонованого методу розроблено і апробовано автоматизовану систему каталогізації літератури для науково-технічної бібліотеки Національного університету харчових технологій.

Ключові слова: *нейронні мережі, автоматизація, бібліотека, каталогізація.*

The article deals with predictive text analysis method using neural networks. The method of using neural networks to identify the literature that arrives in the library for their classification and cataloguing. Based on the method developed and tested an automated system for cataloguing books for Scientific and Technical Library National University of Food Technologies.

Keywords: *neural networks, automation, library, cataloguing.*

Одним з найбільш поширених видів інтелектуальної діяльності є розпізнавання образів, в процесі якого відбувається вирішення задачі класифікації – віднесення деякого об'єкту предметної області до певного класу.

Наявні системи класифікації використовувані для автоматичного зчитування інформації з анкет чи підрахунку голосів базуються на жорстко регламентованому розташуванні елементів, які підлягають класифікації. Якщо об‘єкти, представлені у деякому наборі документів, подібні, але мають довільне розташування, то можливостей діючих систем класифікації недостатньо. Такими об‘єктами можуть бути дані про друковані видання, що заносяться до каталогу бібліотеки. Довідки для ідентифікації певного джерела літератури розташовано, як правило, на другій сторінці книжки. Внесення їх до електронного каталогу бібліотеки Національного Університету харчових технологій (НТБ НУХТ) наразі відбувається в ручному режимі, що займає багато часу та витрат праці.

Зазначимо, що бібліотекар, дивлячись на опис літературного джерела визначає його складові: УДК, автора, видавництво, рік видання, кількість сторінок, тощо, тобто миттєво вирішує задачу розпізнавання елементів тексту. Для ефективного вирішення задачі автоматизованого розпізнавання елементів тексту було б доцільно використати методи на основі моделювання інтелектуальної діяльності людини. На сьогоднішній день таким інструментом є штучні нейронні мережі, системи, здатні навчатися на основі аналізу вхідних даних, та приймати рішення з застосуванням набутих знань.

Біологічний нейрон (нервова клітка) складається з тіла клітини – соми (soma), і двох типів зовнішніх деревоподібних відгалужень: аксона (axon) і дендритів (dendrites). Тіло клітини складається з ядра (nucleus), що містить інформацію про властивості нейрона, і плазми, яка продукує необхідні для нейрона матеріали. Нейрон отримує сигнали (імпульси) від інших нейронів через дендрити (приймачі) і передає сигнали, згенеровані тілом клітки, вздовж аксона (передавача), що наприкінці розгалужується на волокна (strands). На закінченнях волокон знаходяться синапси (synapses).

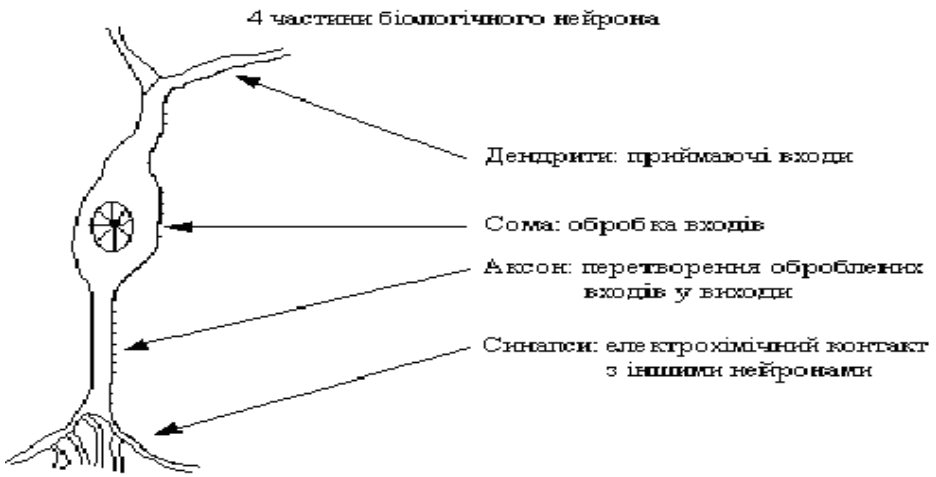

Рис. 1. Схема біологічного нейрона

Синапс є функціональним вузлом між двома нейронами (волокно аксона одного нейрона і дендрит іншого). Коли імпульс досягає синаптичного закінчення, продукуються хімічні речовини, названі нейротрансмітерами. Нейротрансмітери проходять через синаптичну щілину, збуджуючи або гальмуючи, у залежності від типу синапсу, здатність нейрона-приймача генерувати електричні імпульси.

Базовий модуль нейронних мереж штучний нейрон моделює основні функції природного нейрона (рис. 2).

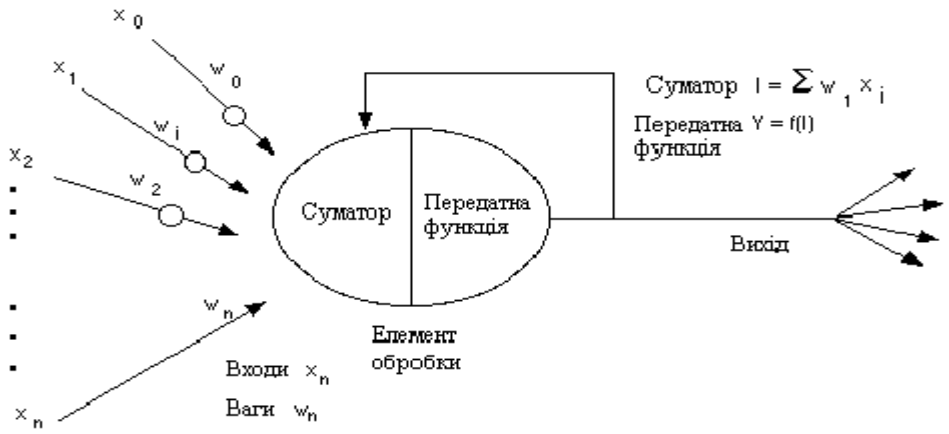

Рис. 2. Базовий штучний нейрон

Вхідні сигнали xn мають вагові коефіцієнти wn. У найпростішому випадку функція суматора виглядає як Σ wi*xi, де і змінюється від 1 до n. Вхідні дані проходять через передатну функцію, генерують результат і виводяться.

Здатність до навчання є фундаментальною властивістю мозку. Процес навчання може розглядатися як визначення архітектури штучної нейронної мережі і налаштування ваг зв'язків для ефективного виконання спеціальної задачі. Нейромережа налаштовує ваги зв'язків по наявній навчальній множині. Властивість нейронних мереж навчатися на прикладах робить їх більш привабливими в порівнянні із системами, які функціонують згідно визначеній системі правил, сформульованої експертами.

Для процесу навчання необхідно мати модель зовнішнього середовища, у якій функціонує нейронна мережа – потрібну для вирішення задачі інформацію. По-друге, необхідно визначити, як модифікувати вагові параметри мережі. Алгоритм навчання означає процедуру, в якій використовуються правила навчання для налаштування ваг.

В даній статті розглядається метод розпізнавання та класифікації ключових елементів текстової інформації для ефективного вирішення задач каталогізації бібліотечних документів. Досягнення мети полягає у швидкій та ефективній класифікації ключових елементів зворотної сторони титульного листа бібліотечного документу та структурованому представленні вихідного текстового блоку у вигляді картки бібліотечного каталогу. В якості критерію ефективності виступає точність класифікації.

Реалізація методу складається з трьох етапів:

1. Підготовчий етап:

- Збирання та підготовка комплекту текстів для навчальної вибірки;
- Підготовка навчальної вибірки для структурних класів тексту.

2. Етап навчання:

- Навчання нейронів штучної нейронної мережі на основі навчальної вибірки;
- Коригування результатів навчання.

3. Етап використання навченої штучної нейронної мережі.

На першому етапі здійснюється підготовка текстових файлів для навчальної вибірки шляхом сканування зворотних сторінок титульних листів певного набору літератури.

Наступним кроком є складання навчальних вибірок для кожного класу, який буде розпізнаватись. Навчальна вибірка для класу – це масив текстових рядків з визначеними коефіцієнтами належності до класу, що коливаються у діапазоні від 0.1 до 0.9. Етап навчання полягає у навчанні нейронів за алгоритмом коригування вагових коефіцієнтів методом зворотнього розповсюдження помилки для визначення класів об‘єктів тексту на основі відповідної навчальної вибірки. Проводиться оцінка результатів навчання кожного нейрона та здійснюються відповідні коригування параметрів навчання, а саме коефіцієнту швидкості навчання, функцій зсувів вхідних значень, активаційної функції. Навчання вважається завершеним, коли всі нейрони в переважній більшості випадків правильно класифікують відповідні структурні елементи тексту.

Етап використання навченої штучної нейронної мережі полягає у аналізі вхідного текстового блоку у такій послідовності:

- визначення інформаційного рядка за ключовими характеристиками класу;
- визначення інших структурних елементів за набором ключових характеристик класу та функції відстані від інформаційного рядка;
- коригування неправильних результатів шляхом вибору вірних значень та проведення додаткового навчання;
- формування файлу результатів класифікації;

Метод було застосовано для вирішення задачі автоматизованої каталогізації літератури науково-технічної бібліотеки Національного університету харчових технологій (НТБ НУХТ). До програмних модулів системи включено метод розпізнавання структурних елементів тексту, об'єктну модель класу „Штучний нейрон" та алгоритм розбиття тексту на елементарні структурні одиниці.

Навчання нейронів на основі навчальних вибірок. Функція надає можлівість обрати нейрон, задати кількість епох навчання, та обрати текстовий файл з навчальною вибіркою.

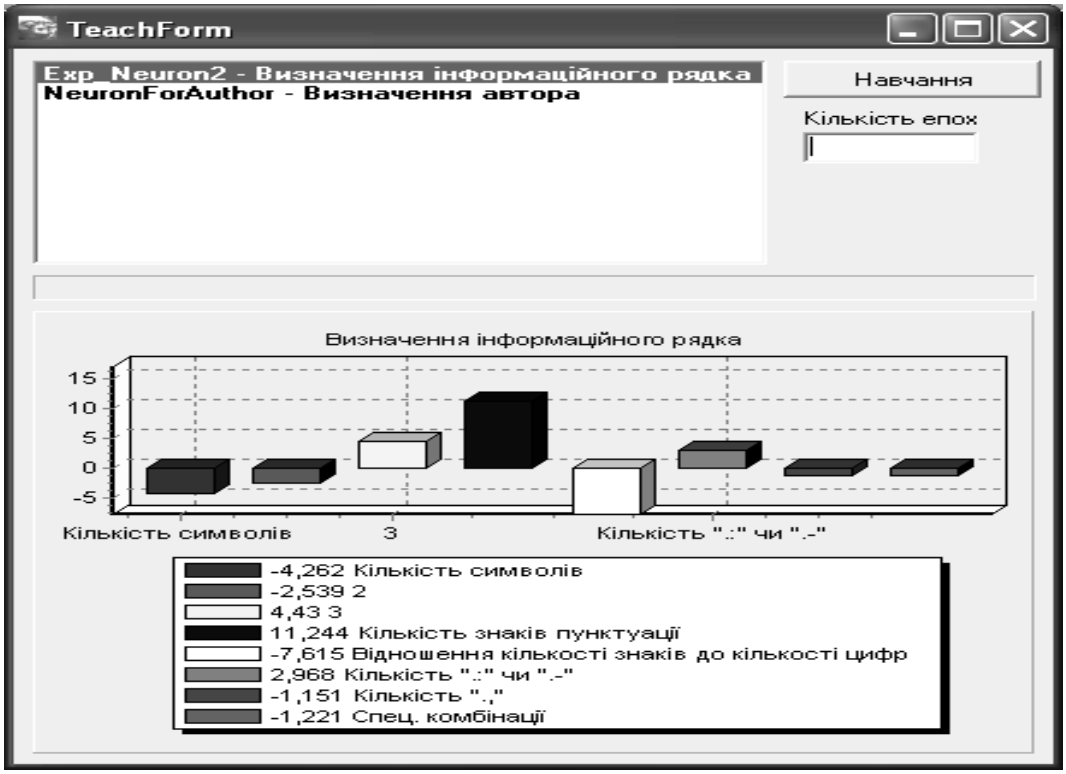

Рис.3 Процес визначення параметрів навчання нейрону

Процес навчання супроводжується візуалізацією змін вагових коефіцієнтів за допомогою діаграми поточного значення вагових коефіцієнтів та діаграми зміни вагових коефіцієнтів у часі.

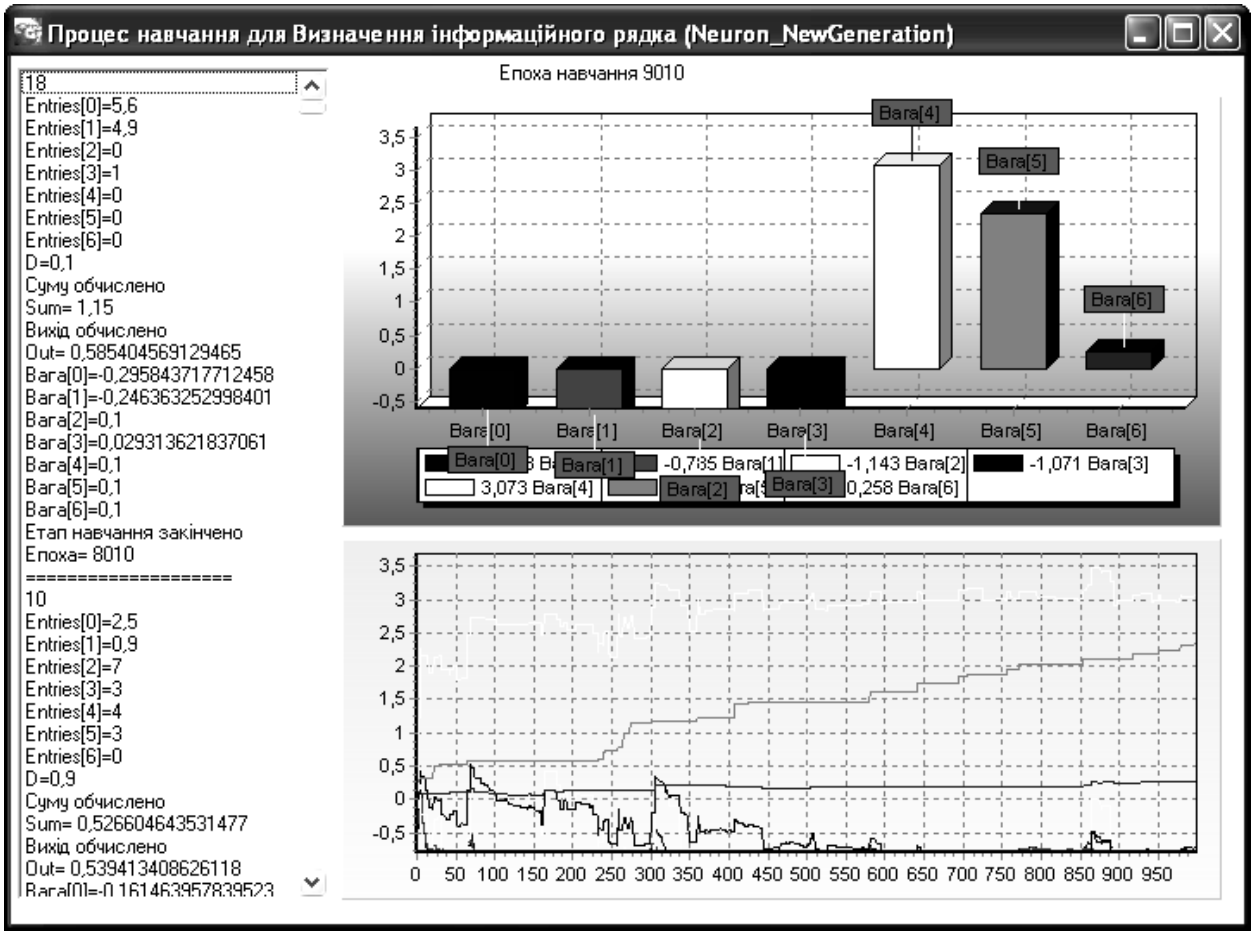

Рис.4 Візуальне відображення процесу навчання нейрону

Налагодження. Функція використовується для створення нових екземплярів класу „Нейрон” , визначення та збереження їх параметрів у файл.

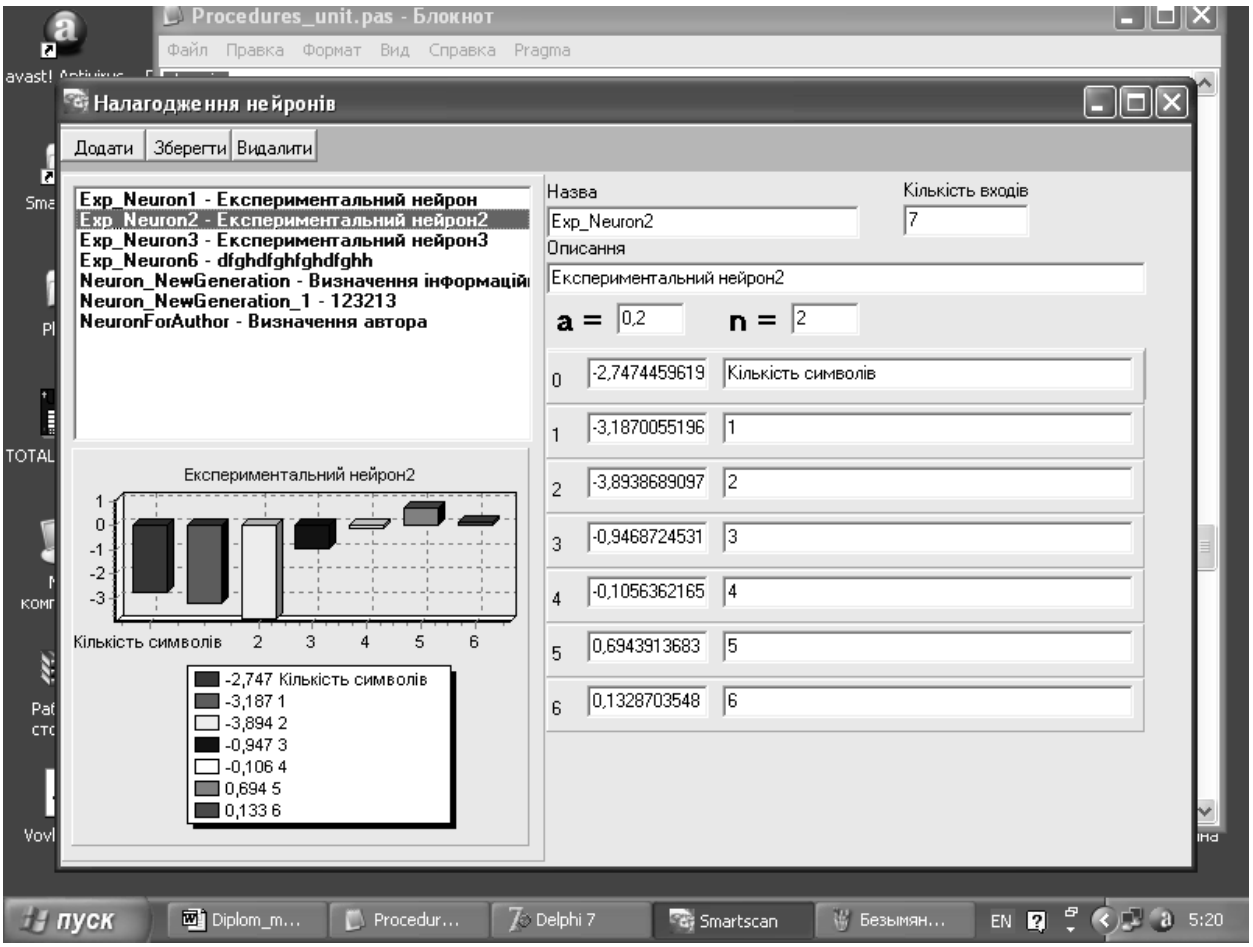

Рис.5 Налагодження нейрону

Розпізнавання з файлу. Функція застосовується для вибору текстового файлу, який буде проаналізовано, з метою розпізнавання ключових структурних елементів тексту.

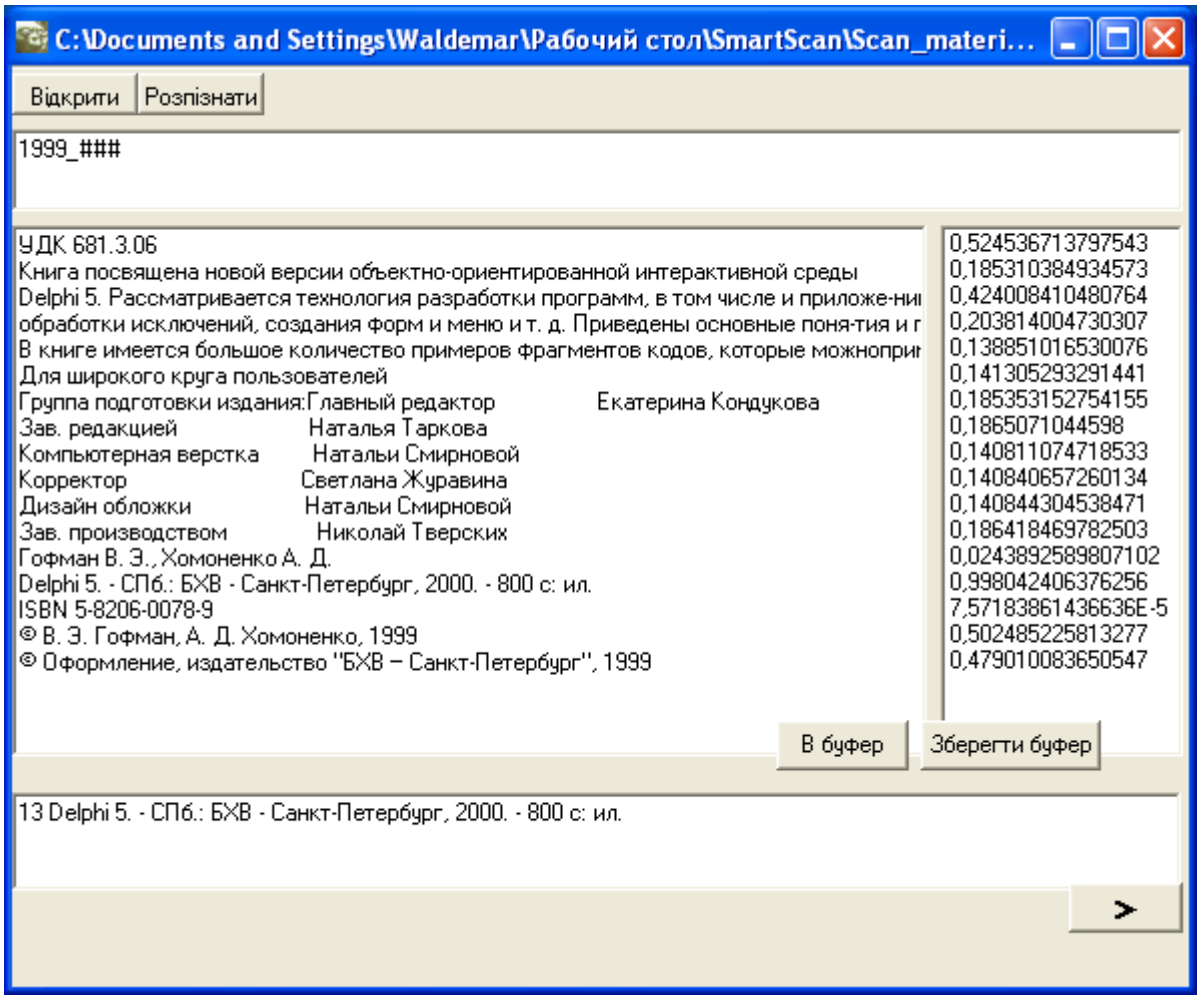

Рис. 4. Розпізнавання інформаційного рядка

Текстовий файл, що містить відскановану другу сторінку підручника, проходить автоматичну перевірку кожного рядка на належність до класу інформаційних рядків. За результатами перевірки обирається елемент, характеристики якого максимально відповідають класу, що визначається.

В даному випадку інформаційним рядком, що відображує назву підручника, видавництво та рік видання буде 13 рядок (нумерація починається з нульового рядка) – коефіцієнт співпадання більше ніж 0,99.

Для класифікації інформаційного рядка використовують наступні характеристики:

1. довжина рядка

2. кількість символів у рядку
3. кількість цифр
4. кількість знаків пунктуації
5. Відношення кількості цифр до кількості символів
6. Кількість спеціальних комбінацій знаків пунктуації.
7. Кількість спеціальних символів.

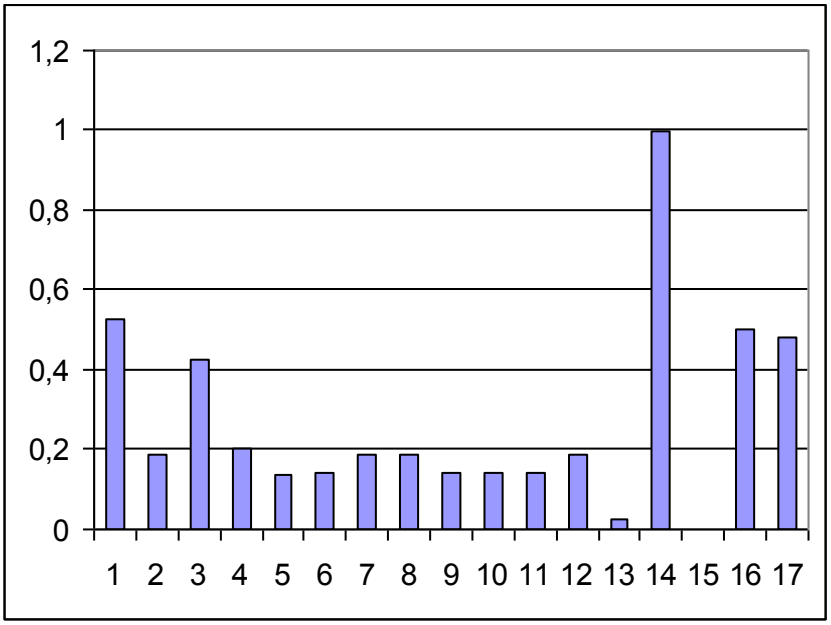

Рис. 5 Діаграма результатів класифікації елементів текстового блоку.

Для визначення ефективності роботи розробленої інформаційної системи, проведемо тестування на масиві тексту, що не приймав участі у формуванні навчальної вибірки.

УКД 802.0-5 (075.3)
ББК 81.2 Англ-2-922
В-31

Серія
«Учням та абітурієнтам»
(заснована в 1997 році)

В-31 Верба Л.Г., [Верба Г.вТ

Граматика сучасної англійської мови. Довідник;

Київ, *Логос», 2000 р. — 352 с.

Рекомендовано Головним управлінням шкіл
Міністерства освіти України

У довіднику висвітлюються основні граматичні явища сучасної англійської мови, які потрібно засвоїти для ведення бесіди і розуміння тестів англійською мовою.

Теоретичний матеріал закріплюється розширеною системою вправ.

Довідник призначено для школярів, студентів, всіх хто вивчає англійську мову.

УКД 802.0-5 (075.3)
ББК 81.2 Англ-2-922

ISBN 996-509-001-1

Рис.6 Відсканований текст для тестування розробки

В результаті тестового аналізу тексту отримали наступні дані:

Інформаційний рядок: Граматика сучасної англійської мови. Довідник;Київ, *Логос», 2000 р. — 352 с.

Автор: Верба Л.Г.

Місто видання: Київ

Видавництво: *Логос

Кількість сторінок: 352

Код УДК: 802.0-5 (075.3)

Код ББК: 81.2 Англ-2-922

Код ISBN: 996-509-001-1

Підсумовуючи вищесказане зазначимо, що в результаті проведених досліджень розроблено метод, модель та алгоритм, які дозволяють більш ефективно вирішувати задачі розпізнавання та класифікації структурних елементів тексту, в тому числі:

- Метод розпізнавання структурних елементів тексту з використанням штучних нейронних мереж, який відрізняється від існуючих методів здатністю оброблювати слабко структурований текст.

- Об’єктна модель штучної нейронної мережі, яка може бути швидко та ефективно підключеною до інформаційної системи з використанням методів моделювання штучного інтелекту.
- Алгоритм розбиття тексту на елементарні структурні одиниці, як допоміжний метод для попередньої обробки тексту.

Розробка виконувалась у рамках наукової магістерської роботи на кафедрі інформаційних систем НУХТ, пройшла дослідну експлуатацію та впроваджена в НТБ НУХТ.

Висновок: використання штучних нейронних мереж для вирішення задач класифікації елементів тексту, розвинене слабко, не має чіткої методології та реалізації, що багато в чому пов'язано з відсутністю опрацьованої теорії і практики рішення подібних задач. Тому вирішення

практичної задачі класифікації елементів тексту з використанням штучних нейронних мереж та проведення досліджень в цьому напрямку є актуальним в контексті розвитку наукових знань з моделювання інтелектуальної діяльності людини.

Список використаних джерел.

1. Kohonen T. Self-organized formation of topologically correct feature maps. // Biological Cybernetics. — 1992.— N 43. — P. 59—69.
2. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688 с.
3. Минский М., Пейперт С. Персептроны. — М.: Мир, 1991. —261 с.
4. Роберт Каллан. Основные концепции нейронних сетей.: Пер. с англ. – М. :Изд. дом „Вильямс”, 2003. – 288 с.
5. Сэлтон Г. Автоматическая обработка, хранение и поиск информации:Пер. с англ. / Под ред. А.И. Китова. – М.: Советское радио, 1993. – 560 с
6. Шульце К.-П., Реберг К.-Ю. Инженерный анализ адаптивных систем: Пер. с нем. — М.: Мир, 1992. — 280 с.