УДК 681.3.019

Д.Є. БАЛХОВСЬКИЙ, Т.В. ГРИЩУК, М.М. БИКОВ

ЕФЕКТИВНА ТЕХНОЛОГІЯ ЕЛЕКТРОНІЗАЦІЇ ДОКУМЕНТІВ В АВТОМАТИЗОВАНИХ ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ

Вінницький національний технічний університет,
Хмельницьке шосе, 95, Вінниця, 21010, Україна

Анотація. Запропонована технологія введення і оброблення текстових документів, що реалізує оптимальний розподіл процесу розпізнавання тексту між мікропроцесорним пристроєм введення і комп'ютером, вибрано критерій оцінки ефективності системи.

Аннотация. Предложена технология ввода и обработки текстовых документов, которая реализует оптимальное распределение процесса распознавания текста между микропроцессорным устройством ввода и компьютером, выбран критерий оценки эффективности системы.

Abstract. Technology of input and treatment of texts documents, which will realize the optimum distribution of the text recognition process between microprocessor device and computer, is offered, the criterion of the system efficiency estimation is chosen.

Ключові слова: автоматизація, електронізація документів, пошукова система.

ВСТУП

Автоматизація процесів наукових досліджень, прийняття рішень в різних областях організаційної і виробничої діяльності вимагає розробки нових технологій ефективної обробки текстових документів, що знаходяться в різноманітних сховищах інформації – бібліотеках, архівах підриємств і організацій, тощо. Автоматизований пошук і аналіз інформації з потрібної тематики передбачає збереження її в форматі електронних документів. На сьогоднішній день представлення текстів за допомогою графічних форматів (таких, наприклад, як *.pdf або *.djv) дозволяє розв'язати проблему підвищення швидкості їх електронізації, однак вимагає наявності людини для опрацювання з метою аналізу і розуміння. Автоматичний аналіз текстової інформації потребує представлення символів в ASCII кодах, що потребує їх розпізнавння в процесі введення. Автори пропонують нову інформаційну технологію електронізації текстів, яка передбачає їх введення і розпізнавання за допомогою розподіленої комп'ютерно-мікропроцесорної системи і дозволяє підвищити швидкість і надійність введення і розпізнавання, а також зменшити вартість системи.

АНАЛІЗ СТАНУ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧІ

Аналіз існуючих методів електронізації текстів в автоматизованих інформаційно-пошукових системах показує, що вони грунтуються на традиційній технології. Дана технологія передбачає введення графічного зображення тексту, а потім його посимвольне розпізнавання в комп'ютері за допомогою розроблених програмних продуктів [1]. Під час розпізнавання виконуються такі стандартні кроки: фільтрація зображення з метою його покращення; бінаризація чи приведення до вигляду, зручного для виділення вибраних інформативних ознак; сегментація на окремі символи; опис отриманих зображень символів в ознаковому просторі чи в лінгвістичному вигляді; розпізнавання за допомогою алгоритмів, що відповідають вибраному методу опису образів [2,3]. До переваг, властивих даному підходу, відносяться незначні затрати пам'яті, зумовлені невеликою кількістю еталонів символів. Основними недоліками є недостатня швидкодія, зумовлена значними обчислювальними затратами на процедури сегментації і розпізнавння кожного символа тексту, та недостатня надійність (точність) розпізнавання. Однак всякий тектстовий документ можна розглядати не тільки як графічне зображення, а і як деякий носій мовної інформації, що використовується для її передачі в тій чи іншій комунікативній системі [4]. З такої точки зору графіка тексту опосередкованим чином відображає різні інформаційні рівні, властиві

комунікативному акту: прагматичний, семантичний, лексичний, морфологічний, сигматичний і афективний [4]. Виникає питання – інформацію якого рівня і в якій послідовності потрібно використовувати в автоматизованому процесі введення і розпізнавання текстового документа, щоб отримати максимально можливу швидкість і мінімально можливі помилки і вартість. Для розв'язання цього питання автори в даній роботі пропонують нову технологію електронізації текстових документів, яка поряд з розпізнаванням графічних образів використовує часткове розуміння тексту. Вона передбачає використання на етапі введення і розпізнавання не тільки графічного зображення тексту, а й низки мовних складових інформації (лексичної, морфологічної, синтаксичної та інш.), що містяться в цьому зображенні і дозволяють здійснити його часткове розуміння, а також оптимально розподіляє процес обробки документа між пристроєм введення і комп'ютерною системою. В цьому дослідженні ставляться задачі оцінки інформативності окремих ознак, що характеризують той чи інший вид інформації, вибору критерію для оцінки ефективності процедури введення і оброблення тексту, а також розробки ефективної стратегії побудови запропонованої технології.

ВИБІР КРИТЕРІЯ ЕФЕКТИВНОСТІ І РОЗРОБКА ЕФЕКТИВНОЇ СТРАТЕГІЇ ПОБУДОВИ СИСТЕМИ ВВЕДЕННЯ І РОЗПІЗНАВАННЯ ТЕКСТІВ

Задача побудови оптимальної стратегії введення і розпізнавання тексту в нетривіальній постановці повинна пов'язуватися з ефективністю роботи системи в цілому, а не тільки з умовою досягнення заданої точності розпізнавання. Формалізація процедури побудови ефективної стратегії розпізнавання можлива тільки в тому випадку, коли попередньо неформальними методами вибрано апріорний алфавіт ознак образів, а також вибрано критерії для оцінки ефективності системи розпізнавання. Тоді формальна постановка задачі побудови оптимальної стратегії розпізнавання може бути сформульована як задача пошуку оптимального по загальносистемному критерію дерева рішень, в якому на кожному кроці класифікації з апріорного алфавіту вибирається підмножина ознак, що максимально зменшує на досягнутому кроці ентропію про образ і збільшує швидкість класифікації.

Тому в роботі як загальносистемний критерій використовується узагальнений функціонально-статистичний критерій, модифікований належним чином до системи введення і розпізнавання тексту шляхом належного вибору потенціальної і реальної систем [5], [6]:

$$Э = \frac{Э_{\text{Р}}}{Э_{\text{П}}}\Bigg| E = E_{¶} \,, \tag{1}$$

де $Э_{\text{Р}}$ і $Э_{\text{П}}$ - функціонально-статистичні критерії для реальної і потенціальної систем відповідно;

$$Э_{\text{Р}} = \frac{\text{I}_{\text{Р}}}{C_{\text{Р}}}, \tag{2}$$

$$Э_{\text{П}} = \frac{I_P}{C_P}, \tag{3}$$

E_{∂} - задана в технічному завданні точність розпізнавання; $\text{I}_{\text{Р}}, \text{I}_{\text{П}}$ - кількість інформації, яку дістає реальна і потенціальна системи відповідно, визначається з урахуванням ентропійних властивостей текстових образів; $C_{\text{Р}}, C_{\text{П}}$ - вартість реальної і потенціальної систем відповідно;

$$C_{\text{Р}} = C_{\text{Х}} + C_{\text{К}} \,, \tag{4}$$

де $C_{\text{Х}}$ - складність обчислення ознакового опису образів; $C_{\text{К}}$ - складність обчислень класифікації образів.

Критерій (1) задовольняє всім вимогам, що висуваються до показників якості роботи систем.

Формальна постановка задачі побудови оптимальної стратегії розпізнавання може бути подана в наступному вигляді:

$$\tilde{S}_{Gopt} = \arg u \max Э(\tilde{S}_{Gi})\Big|_{\{\tilde{S}_{Gi} \in \tilde{S}_G, W_d, r_d, E_d\}}, \quad (5)$$

де $\widetilde{S}_{Gi}$ - одна із стратегій розпізнавання із замкнутої відносно доступної інформації множини стратегій розпізнавання $\tilde{S}_G$; W_d, r_d, E_d - задані умовою задачі розпізнавання алфавіт образів, рівень завад і точність розпізнавання відповідно.

Складність рішення задачі оптимізації, сформульованої у вигляді (5), полягає в тому, що обчислювальна складність C_p критерію ефективності (2) є функцією двох змінних - обчислювальних затрат на процедуру класифікації C_k та затрат на обчислення ознакового опису C_x. Декомпозицію цих змінних в критерії (2) можна здійснити шляхом зображення стратегії розпізнавання в вигляді покрокової процедури класифікації на дереві рішень.

Використання послідовно-паралельної стратегії розпізнавання в вигляді дерева рішень в оптимізаційній процедурі побудови ефективної стратегії розпізнавання образів не виключає узагальненого підходу, так як часто вживані алгоритми статистичного розпізнавання в n-вимірному просторі ознак можна подати у вигляді дерева рішень, в якому є один кореневий вузол, а всі інші вузли - термінальні. Кількість гілок такого дерева (кроків класифікації) дорівнює кількості термінальних вузлів (образів, що розпізнаються).

Оскільки складність C_p системи розпізнавання є адитивною сумою складностей C_i кожного з ієрархічних рівнів розпізнавання, а інформативність I_P є неспадною функцією ймовірності правильного розпізнавання, то оптимальна стратегія є композицією алгоритмів розпізнавання, що максимізують відношення $\frac{I_i}{C_i}$ на кожному з рівнів. Послідовність композиції алгоритмів в оптимальній стратегії повинна відповідати послідовності розміщення рівнів дерева класифікації, а ознаки на кожному рівні повинні вибиратися з умови їх мінімальної складності:

$$\widetilde{S}_{Gopt} = A_1(\widetilde{I}(S^1)) \otimes A_2(\widetilde{I}(S^2)) \otimes \cdots \otimes A_W(\widetilde{I}(W\ )), \quad (6)$$

$$\widetilde{I}(S^i) = x^i_{opt}, x^i_{opt} = \arg u \min C_i(x^i_l),$$

де $\widetilde{I}(W)$ - ознаковий опис еталонів образів; $C_i(x^i_l)$ - обчислювальна складність l-ї ознаки, що використовується для опису елементів i-го рівня; $A_h(\widetilde{I}(S^h))$ - алгоритм попередньої класифікації образів тексту на групи, а $\widetilde{I}(S^h))$ - ознаковий опис образів S^h, що розпізнаються на даному рівні ієрархії. Як було досліджено авторами в роботі [7], для сумісної оптимізації дерева рішень відносно помилок, швидкості і вартості необхідно на верхніх рівнях дерева використовувати більш інформативні ознаки.

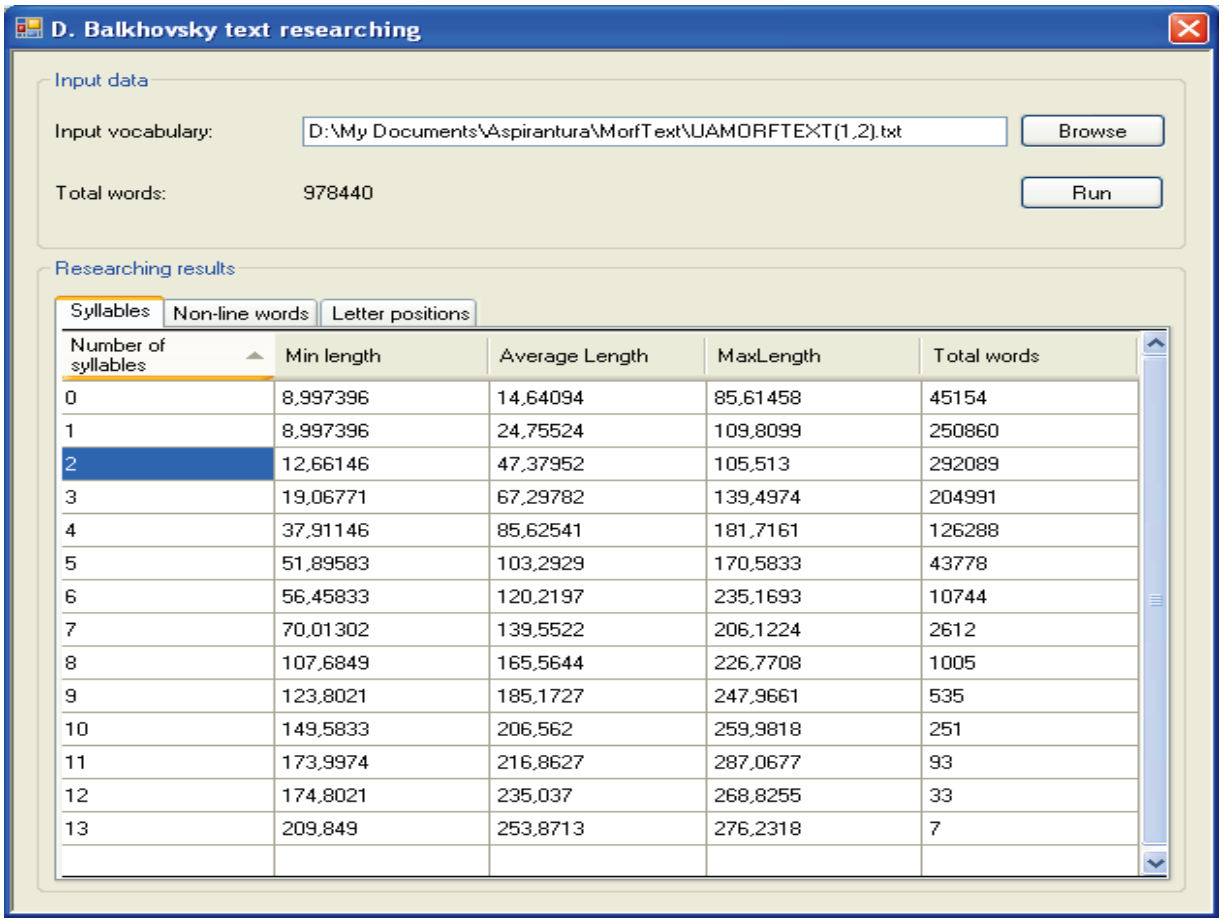
Оптимальний розподіл процедури розпізнавання передбачає передачу на мікропроцесорний пристрій виконання тих функцій обробки графічного зображення, що лежать не нижче рівня h_p:

$$h_p = arg(C_{max}), \quad (7)$$

де h_p - номер рівня дерева прийняття рішень, на якому складність обчислень досягає граничних обчислювальних можливостей C_{max} мікропроцесорного пристрою введення зображення тексту.

Оцінка інформативності ознак текстових образів проводилася з метою визначення ефективності їх застосування в вузлах дерева рішень запропонованої ієрархічної процедури розпізнавання. В якості графічних образів (графем) тексту, які можуть бути використані для попереднього розпізнавання в більш ранніх роботах авторами було запропоновано використовувати на лексичному рівні графеми слів і морфем з тих міркувань, що перші можна легко сегментувати в зображенні, а другі представляють собою скінченну множину стійких до спотворень змістовних одиниць інформації. Для дослідження в

якості ознак були вибрані довжини слів і надстрічкові і підстрічкові особливості слів і морфем, що задаються графікою написання окремих слів і морфем. Для дослідження був сформований текст із статистично достатньої вибірки слів розміром більше 900 000 слів. Приклад результатів проведених досліджень з впливу такої ознаки, як довжина слова на скорочення альтернатив під час розпізнавання наведено на рисунку 1.



Number of syllables	Min length	Average Length	MaxLength	Total words
0	8,997396	14,64094	85,61458	45154
1	8,997396	24,75524	109,8099	250860
2	12,66146	47,37952	105,513	292089
3	19,06771	67,29782	139,4974	204991
4	37,91146	85,62541	181,7161	126288
5	51,89583	103,2929	170,5833	43778
6	56,45833	120,2197	235,1693	10744
7	70,01302	139,5522	206,1224	2612
8	107,6849	165,5644	226,7708	1005
9	123,8021	185,1727	247,9661	535
10	149,5833	206,562	259,9818	251
11	173,9974	216,8627	287,0677	93
12	174,8021	235,037	268,8255	33
13	209,849	253,8713	276,2318	7

Рис. 1. Результати досліджень інформативності ознаки “довжина слова”

Результати цих досліджень показали, що в окремих випадках дані ознаки можуть звузити пошук альтернатив для етапу розпізнавання графічних зображень в 3-5 разів.

ВИСНОВКИ

Запропонована нова інформаційна технологія введення і оброблення текстової інформації в автоматизованих інформаційно-пошукових системах, яка відрізняється від існуючих тим, що передбачає використання на етапі введення і розпізнавання не тільки графічного зображення тексту, а й низки мовних складових інформації (лексичної, морфологічної, синтаксичної та інш.), що містяться в цьому зображенні і дозволяють здійснити його часткове розуміння, а також оптимально розподіляє процес обробки документа між пристроєм введення і комп’ютерною системою. Запропоновано критерій ефективності, який дозволяє регулювати побудову оптимального дерева прийняття рішень з оброблення текстового документа і ієрархічну стратегію рзпізнавання тексту, оптимальну з точки зору помилок класифікації, швидкості розпізнавання і вартості. Проведено дослідження низки лексичних і морфологічних ознак, яке виявило їх високу інформативність для розпізнавання мовної і графічної інформації, що міститься в зображенні тексту.

СПИСОК ЛІТЕРАТУРИ

1. Ту Д., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 411 с.

2. Анисимов Б.В., Курганов В.Д., Злобин В.К.- Распознавание и цифровая обработка изображений. – М.- 1983.-С.35-68.
3. Фу К. Структурные методы в распознавании образов. – М.: Мир, 1976. – 284 с.
4. Р.Г. Пиотровский. Текст машина, человек.– Ленинград: Издательство “Наука”,1975.– 326 с.
5. Быков Н.М., Агеев А.С. Модель потенциальной системы для распознавания речи в СЧМ.- В кн.: Исследование и проектирование систем “человек-машина”. – Киев: ИК АНН УССР, 1986. – С. 57-61.
6. Биков М.М.,Грищук Т.В. Розробка методів оцінки ефективності автоматизованих систем розпізнавання мови // Вісник Технологічного університету Поділля – Хмельницький, ТУП, 2003. - №3, том 1 – С. 122-125.
7. Bykov N.M., Kuzmin I.V., Yakovenko A.I. Development of effective strategy of pattern recognition. – Proceedings of SPIE, 2000, Vol.4425, pp. 76-82.

Надійшла до редакції 05.10.2008р.

БАЛХОВСЬКИЙ Д.Є. – аспірант кафедри комп’ютерних систем управління, Вінницький національний технічний університет, Вінниця, Україна.

ГРИЩУК Т.В. – к.т.н., доцент кафедри комп’ютерних систем управління, Вінницький національний технічний університет, Вінниця, Україна.

БИКОВ М.М. – к. т. н., професор кафедри комп’ютерних систем управління, Вінницький національний технічний університет, Вінниця, Україна.