УДК 004.912:81'32

М.М. Биков
Р.С. Калінчук

Методи підвищення ефективності процесу розпізнавання тексту

Вінницький національний технічний університет

Анотація
В роботі розроблено математичну модель процесу розпізнавання тексту і визначено критерії оцінки його ефективності, представлено математичні основи визначення інформаційної структури тексту і його морфологічних характеристик, розроблено метод автоматичного виділення морфем в тексті.
Ключові слова: розпізнавання тексту, критерії оцінки ефективності, морфологічні характеристики тексту.

Abstract
The mathematical model of process of recognition of the text is developed and criteria for evaluation of its efficiency is defined, the mathematical grounds of determination of information structure of the text and its morphological characteristics is presented, the method of automatic allocation of morphemes in the text is developed.
Keywords: text recognition, criteria for efficiency evaluation, morphological characteristics of the text.

Вступ

Аналіз сучасного стану проблеми оцінки ефективності систем розпізнавання образів, а також побудови ефективних стратегій розпізнавання показав, що її опрацюванню приділено недостатню увагу. Разом з тим, за думкою фахівців в області зображень, оптимізація процесу розпізнавання графічних образів можлива тільки за допомогою процедури “від початку до кінця”, тобто одночасно за всіма елементами системи розпізнавання.

Використання традиційних технологій електронізації документів в текстових форматах, які дозволяють представити символи в ASCII кодах, і, таким чином, автоматизувати їх аналіз, передбачає посимвольне розпізнавання графічного зображення тексту за допомогою наявних програмних засобів (наприклад, FineReader). Однак такі технології в своїх історичних витоках орієнтовані на брак апаратних ресурсів (швидкодії і пам’яті), не враховують технічних можливостей сучасних обчислювальних систем і мікропроцесорних засобів, а також не використовують мовних складових в інформаційній ієрархії текстового документа[1,2,3]. Тому тема даної роботи, присвячена розробці ефективних методів оброки текстових документів для електронізації , є актуальною.

Результати досліджень

Традиційні технології обробки текстових документів в інформаційно-пошукових системах передбачають виконання їх посимвольного розпізнавання під час введення. З огляду на невисоку швидкодію сучасних засобів сканування і розпізнавння, на сьогодні процедура введення тексту в найбільшій мірі гальмує процесс електронізації текстів. Підвищення швидкості і точності цієї процедури можливі за рахунок застосування інтелектуальних технологій, які під час обробки інформації наслідують механізми діяльності мозку людини. Ці механізми передбачають в значній мірі використання мовного тезаурусу мозку на основі значних ресурсів його пам’яті і ієрархічної паралельної обробки інформації. Зважаючи на можливості сучасної обчислювальної техніки, реалізація таких інтелектуальних технологій впирається в брак знань про інформативність різних графічних елементів тексту (будемо називати їх графемами) для розуміння текстової інформації.

З лінгвістичної точки зору будь-який текстовий документ можна розглядати як деякий носій мовної інформації, що використовується для її передачі в тій чи іншій комунікативній системі [4]. З цієї точки зору зображення тексту опосередкованим чином відображає різні інформаційні складові, присутні комунікативному акті: прагматичний, семантичний, лексичний, морфологічний, сигматичний і афективний [4]. Виникає задача – в якій послідовності потрібно використовувати інформацію того чи

іншого рівня в автоматизованому процесі введення і розпізнавання текстового документа, щоб отримати максимально можливу швидкість і мінімально можливі помилки і вартість [5]. Для розв'язання цього питання в даній роботі пропонується нова технологія обробки текстових документів, яка передбачає використовувати часткове розуміння тексту під час розпізнаванням графічних образів. Для цього вона використовує низку мовних складових інформації - лексичної, морфологічної, синтаксичної та інш. на етапі введення і розпізнавання поряд з графічним зображенням тексту. Ці виділені в графічному зображенні складові дозволяють здійснити його часткове розуміння, а також оптимально розподілити процес обробки документа між пристроєм введення і комп'ютерною системою.

Для реалізації даної технології в роботі вирішені задачі оцінки інформативності окремих ознак того чи іншого виду інформації, вибору критерію оцінки ефективності процедури електронізації тексту, розробки методів, агоритмів і програмного забезпечення автоматизації створення бази даних морфем української мови, які є елементами розпізнавання в запропонованому підході.

В якості графічних образів (графем) тексту, які можуть бути використані для попереднього розпізнавання 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 надстрокові і підстрокові особливості слів і морфем, що задаються графікою написання окремих слів і морфем. Результати цих досліджень показали, що в окремих випадках дані ознаки можуть звузити пошук альтернатив для етапу розпізнавання графічних зображень в 3-5 разів.

Висновки

Запропоновані в роботі методи підвищення ефективності введення і оброблення текстової інформації в автоматизованих інформаційно-пошукових системах відрізняється від існуючих тим, що передбачують 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 розподілити процес обробки документа з метою його розпізнаваня між пристроєм введення і комп'ютерною системою.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Биков М. М. Використання інтелектуальних методів в розпізнаванні символів / М. М. Биков, Д. Є. Балховський, А. Раїмі // Інформаційні технології та комп'ютерна інженерія. – 2007. – № 2 (9). – С. 121 – 125.
2. Методи розпізнавання тексту.- [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/методи розпізнавання тексту.
3. Репік С. І., Штогріна О. С. Методи розпізнавання тексту / С.І. Репік, О.С. Штогріна // Збірник матеріалів Міжнародної науково-технічної конференції «ПЕРСПЕКТИВИ ТЕЛЕКОМУНІКАЦІЙ», [S.l.], nov. 2016. – [Електронний ресурс]. Режим доступу: <http://conferenc.its.kpi.ua/proc/article/view/71101>.
4. Пиотровский Р.Г. Математическая лингвистика / Р.Г. Пиотровский, К.Б. Бектаев, А.А. Пиотровская. – М.: Высшая школа, 1977. – 384 с.
5. Пиотровский Р.Г. Математическая лингвистика / Р.Г. Пиотровский, К.Б. Бектаев, А.А. Пиотровская. – М.: Высшая школа, 1977. – 384 с.

Науковий керівник: **Микола Максимович Биков** – кандидат технічних наук, доцент, професор кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, e-mail: nkbykov@vntu.edu.ua

Калінчук Роман Сергійович – студент групи 2АКІТ-18м, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця, e-mail: t.v.gayuk@gmail.com

Supervisor: ***M. Bykov*** **–** Ph.D., Professor at the Computer Control Systems Department, Vinnitsia National Technical University

Roman Kalinchuk **–** student of group 2ACIT-18m of Computer Systems and Automation Faculty, Vinnitsia National Technical University