**УДК 681.301**

**Д.т.н., професор Зайцев В.Г., аспірант Лан Чуньлінь**

**Національний технічний університет України**
**«Київський політехнічний інститут»**

# ВПРОВАДЖЕННЯ СИСТЕМИ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ДОКУМЕНТІВ В БАГАТОМОВНИХ ІНФОРМАЦІЙНИХ СИСТЕМАХ

## Abstract

***Zaitsev Vladimir, prof.,Lang Chunlin, PHD student***
***The implementation of automatic classification system for multilingual environment***
*The paper investigates the practical aspects of creating an automated system for automatic language recognition and classification of documents. The architecture of the automatic text classification system for multilingual environment are described.*

## Вступ

У сучасному світі існує декілька систем класифікації великих об'ємів текстової інформації, в основі яких лежать технології комп'ютерної лінгвістики і алгоритмів розпізнавання образів. В даний час існує велика кількість текстових документів на різних мовах. Сьогодні офіційними мовами Організації Об'єднаних Націй (ООН) є: англійська, арабська, іспанська, китайська, російська і французька. Звичайні системи автоматичної класифікації текстів вирішують задачу автоматичної класифікації текстів на одній мові. Таким чином, необхідно вирішити завдання автоматичної класифікації текстів як для однієї мови, так і для завдання автоматичної класифікації текстів на різних мовах.

## Постановка задачі

Систему автоматичної класифікації текстів для скінченої множини мов можна розділити на дві підсистеми. Перша підсистема слугує для автоматичного розпізнавання мови (Language Recognition) і ідентифікації (Language Identification). Друга підсистема - для автоматичної класифікації текстів з першої підсистеми.

**Перша підсистема: автоматичне розпізнавання та кодування мови**

Підсистема для автоматичного розпізнавання мови складається з двох частин. Блок автоматичної ідентифікації мови: вибір тексту з файлу довільного формату, визначення кодування тексту. Блок автоматичного розпізнавання мови поділяє текст на абзаци, речення, слова і визначає мову. Розпізнавання реалізоване з використанням так званої "N-Gram-Based Text Categorization" технології.

У статті [1] запропонований метод для визначення мови і кодування документа по його вмісту на підставі статистик документів, для яких мова і кодування відомі заздалегідь. Метод визначає частоти N-грамм (підрядків або поєднань символів, завдовжки не більше N), передбачає, що приблизно 300 найбільш часто використовуваних N-грамм сильно залежать від мови. Алгоритм включається при знаходженні частот N-грамм для всіх тестових документів, для яких відома мова, а також для кожного документа, мова якого потребує визначення. Після цього серед всіх тестових документів знаходять той, для якого відстань від його N-граммної статистики до статистики тестованого документа мінімальна. Після того за мову тестованого документа вважається мова знайденого тестового документа. Відстань між статистиками визначається таким чином: всі N-грамми сортуються в порядку зменшення частоти їх появи, потім для кожної N-грамми обчислюється різниця її позицій у відсортованому списку N-грамм тестового і тестованого документів. Відстань між статистиками визначається як сума різниць позицій кожної N-грамми:

$$l = \sum_{i=1}^{300} \left| P_i - \tilde{P}_i \right| \quad (1)$$

де, $P_i$ $\tilde{P}_i$ – позиції і-ї N-грамми в тестовому і тестованому документах відповідно. [2]

**Друга підсистема: автоматична класифікація текстів**

Класифікація документів — одне із завдань інформатики, що полягає у віднесенні документа до однієї з декількох категорій на підставі вмісту документа. Використовує методи інформаційного пошуку і машинного навчання. [3]

**Первинна обробка документів**

а) Вибір ваги ознак і зменшення розмірності.

У статті [4] наведено докладне дослідження різних підходів до вибору ваг ознак, що характеризують категорію ( множину ключових слів). Результати експериментів, наведених у цій статті, показують, що однією із кращих формул обчислення ваги є:

$$W_{ij} = TF_{ij} * IDF_i \qquad (2)$$

Кожен документ - це просто набір слів (термів). Множину всіх термів позначимо як *T*. Кожен терм $t_i \in T$ має вагу $w_{ij}$ стосовно документа $d_j \in D$ . Таким чином, кожен документ можна представити у вигляді вектора ваг його термів $\overrightarrow{d_j} =< w_{ij},...,w_{|T|j} >$. Вагу документів нормують так, щоб $0 \le w_{ij} \le 1$ для $\forall i, j : 0 \le i \le |T|, 0 \le j \le |D|$. Тут $TF_{ij}$ - відношення числа термів $t_i$ у документі $d_i$ до загального числа термів у цьому документі, а $IDF_i$ - число, обернене кількості документів, у якому зустрічається терм.

Після наведення всіх слів документа до нормалізованої форми, отриманий простір ознак має дуже велику розмірність. Цю розмірність можна істотно зменшити без погіршення якості класифікації, якщо виключити слова, що слабо впливають на результати. Звичайно із списку ознак видаляють так звані "стоп-слова" (stopword) і із списку ознак можна видалити слова, що рідко зустрічаються [5].

б) Метод опорних векторів SVM (Support Vector Machines).

Знаходження оптимальної площини поділу множини методом SVM зводиться до рішення оптимізаційної задачі з лінійними обмеженнями типа рівностей і нерівностей [6]:

$$L_D(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \to \max, 0 \le \alpha_i \le C, \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (3)$$

Тут $K(x_i, x_j)$ - функція ядра SVM, яка в простому випадку дорівнює евклідову скалярному добутку векторів $x_i$ і $x_j$ . Для вирішення завдання (3) запропоновані ефективні методи рішення [7].

**Оцінка якості класифікації**

Повнота (recall) - відношення кількості знайдених документів з категорії до загальної кількості документів категорії. Точність (precision) - визначається як відношення числа релевантних документів, знайдених ІПС (інформаційно-пошукових систем), до загального числа документів.

$$recall = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|} \quad (4) \quad precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|} \quad (5)$$

Де $D_{rel}$ — це множина релевантних документів в базі, а $D_{retr}$ — множина документів, знайдених системою [8].

**Результати роботи**

Система розпізнає наступні мови: російську, китайську, українську, англійську, німецьку, французьку. Було протестовано три групи з 8000 документів. Результати: середня повнота 87.3%, середня точність 89.6%.

**Висновки**

Вибраний метод і результати роботи підтверджують можливість створення ефективної системи автоматичної класифікації документів для скінченої множини мов по критерію належності до певної області знання, використовуючи сучасні засоби обчислювальної техніки.

**Література**


1. *Cavnar, W. B. and J. M. Trenkle*, "N-Gram-Based Text Categorization" In Proceedings of Third Annual Symposiumon Document Analysis and Information Retrieval,Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994
2. *Сотник С.Л.* Идентификация языка UNICODE-текста по N-граммам длиной до 4-х включительно (квадрограммам). Журнал "Математичне моделювання" №1,2(15) 2006, стр. 111-112, Днепродзержинск, издательство ДГТУ.
3. http://ru.wikipedia.org/wiki/Классификация_документов
4. *Salton G, Buckley C.* Term-Weighting Approaches in Automatic Text Retrieval. // Information Processing and Management, —1988 — p. 513-523.
5. *Yang Y., Pedersen J.* A comparative study on feature selection in text categorization. // In: Proc. of ICML-97, 14th International Conf. On machine Learning — Nashville, USA, 1997. — p. 412-420.
6. *Vapnik V.* The Nature of Statistical Learning Theory. — Springer-Verlag — New York, 1995. — p. 123-167.
7. *Joachims T.* Making Large-Scale SVM Learning Practical. Advances in Kernel Methods // Support Vector Learning, Burges C., Smola A. (ed.), — MIT-Press, 1999. —p. 5-12.
8. http://ru.wikipedia.org/wiki/Информационный_поиск