УДК 681.301

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

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

СИСТЕМА АВТОМАТИЗОВАНОГО РОЗПІЗНАВАННЯ ТА КЛАСИФІКАЦІЇ ДОКУМЕНТІВ

Вступ

Створення автоматизованої системи розпізнавання та класифікації документів можна розділити на 2 етапи: різні види інформації перетворювати в текстові документи, а потім класифікувати текстові документи. Дослідження стосується тільки другої частини питання, тому що одержання зображення зі сканера або фото, перетворення у текстовий документ - суто технічна проблема.

Постановка задачі

Формалізуємо задачу. Нехай задані деяка скінчена множина категорій $C = \{c_1,...,c_{|C|}\}$, скінчена множина документів $D = \{d_1,...,d_{|D|}\}$, але невідома цільова функція Φ, яка для кожної пари <документ, категорія> визначає, чи відповідають вони один одному: $\Phi: D \times C \to \{0,1\}$. Задача полягає в тому, щоб знайти максимально близьку до функції Φ функцію Φ'. Функцію Φ' називають класифікатором. Машинне навчання базується на початковій сукупності документів $\Omega = \{d_1,...,d_{|\Omega|}\} \subset D$. При цьому, значення цільової функції Φ відоме для кожної пари $< d_j, c_i > \in \Omega \times C$. Документ $d \in \Omega$ називається позитивним або негативним прикладом для категорії c, якщо значення функції $\Phi(d,c)$ дорівнює 1 або 0, відповідно (належить - не належить до категорії) [1].

Первина обробка документів

1. Вибір ваги ознак. У статті [2] наведено докладне дослідження різних підходів до вибору ваг ознак, що характеризують категорію (множину ключових слів). Результати експериментів, наведених у цій статті, показують, що однією з кращих формул обчислення ваги є:

$$W_{ij} = TF_{ij} * IDF_i \quad (1)$$

кожен документ - це 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.

Нормалізована вага терма в документі може бути визначена як:

$$w_{ij} = \frac{TF_{ij} * IDF_i}{\sqrt{\sum_{s=1}^{|T|}(TF_{sj} * IDF_s)^2}} \quad (2)$$

2. Зменшення розмірності. Навіть після приведення всіх слів документа до нормалізованої форми з урахуванням ваги, отриманий простір ознак має дуже велику розмірність (десятки тисяч). Цю розмірність можна істотно зменшити без погіршення якості класифікації, якщо виключити слова, що слабо впливають на результати класифікації [3].

Побудова і навчання класифікатора

Існують декілька найбільш поширених методів класифікації:

1. Метод Байєса.
2. Метод опорних векторів SVM (Support Vector Machines).
3. Інші методи: метод k-найближчих сусідів; нейронні мережі; метод Роше (Rocchio);дерева рішень.

Оцінка якості класифікації за різними методами

З метою вибору методу були запропоновані наступні критерії відбору : повнота та точність.

У статтях [4,5,6,7] порівнюються за цими критеріями методи машинного навчання та класифікації, а результати показують, що метод SVM має перевагу над іншими методами.

Було протестовано 1000 документів. Результати тестування для методу SVM наведені в таблиці табл. 1.

Таблиця 1. Результати тестування документів

	Бізнес	Здоров'я	Мистецтво	Комп'ютери	Гуманітарні науки	Середнє
Точність	83. 223	98. 211	80. 398	93. 658	94. 048	89. 9076
Повнота	91. 549	97.03	89. 759	89. 343	93. 941	92. 3244

Програмні реалізації SVM

Відомі програмні реалізації алгоритму SVM. Досить докладний список можна знайти в інтернеті [8]. Для досліджень в цій роботі використний SVM_light.

Функції автоматизованої системи розпізнавання та класифікації документів

1. Попередня обробка текстів: використання OCR (Optical Character Recognition) технології для розпізнавання різних форм документів у текстовому вигляді.
2. Навчання класифікатора.
3. Машинне розпізнавання та класифікація.

Висновки

Вибраний метод та результати експериментів з автоматичного тестування документів підтверджують можливість створення ефективної системи автоматичної класифікації документів за критерії при належності до відповідної області знань, використовуючи сучасні технічі засоби обчислювальної техніки.

Література

1. *Юрий Лифшиц.* Курс "Алгоритмы для Интернета" ПОМИ РАН 2006. http://logic.pdmi.ras.ru/~yura/internet.html
2. *Salton G, Buckley C.* Term-Weighting Approaches in Automatic Text Retrieval. / Information Processing and Management, -1988 - pp. 513-523.
3. *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. — pp. 412-420.
4. *Dumais S., Platt J., Heckerman D., Sahami M.* Inductive learning algorithms and representations for text categorization. // In Proc. Int. Conf. on Inform. and Knowledge Manage., 1998.
5. *Yang Y., Liu X.* A re-examination of text categorization methods. // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99), 1999 — pp. 42-49.
6. *Yang Y.* An Evaluation of Statistical Approaches to Text Categorization. / / Journal of Information Retrieval, 1999 - V.1 - pp. 67--88.
7. *Joachims T.* Text Categorization with Support Vector Machines: Learning with Many Relevant Features. // Proceedings of ECML-98, 10th European Conference on Machine Learning — 1998.
8. http://www.kernel-machines.org/software