УДК 519.689.3

К.т.н., доцент Олефір О.С., ст. викладач Мальчиков В.В., магістрант Заєздний О.Є.

Національний технічний університету України «Київський політехнічний інститут»

АВТОМАТИЧНА КЛАСИФІКАЦІЯ ТЕКСТІВ

Вступ

Останнім часом все частіше постає проблема автоматичної класифікації текстів. Класифікація текстів – це сортування текстів за заздалегідь відомими категоріями. Задачі класифікації текстів зустрічаються при категоризації сторінок в Інтернеті, виявленні небажаної поштової кореспонденції, автоматичної генерації метаданих, і, взагалі, у будь-яких випадках, коли є потреба автоматичної організації документів.

Проблемам автоматичної класифікації текстів присвячено чимало досліджень. Так, зокрема, в [1] дається загальний підхід до процесу класифікації, виділяються основні етапи, вимоги до результатів кожного з них. Розглянуто лінійний онлайн-класифікатор, метод регресії, метод ДНФ-правил. У [2] дано модифікований наївний Байєсівський алгоритм для класифікації електронних листів за ознакою спам / не спам. У [1] не виконано оцінкювання кожного з методів з точки зору доцільності та оптимальності використання на практиці.

Загалом, у [1] залишилась невивченою практична сторона описаних методів, деякі з них непридатні для якісної класифікації колекцій, кількість документів у яких перевищує кількість документів у навчальній вибірці в десятки разів. Метод класифікації, описаний у [2], не враховує питання оптимального алгоритму, за яким формуються терми та їх вагові коефіцієнти у документі зокрема, та у колекції документів загалом. Також не повністю вирішеною є проблема зменшення розмірності задачі.

Постановка задачі

Мета даної статті – реалізувати автоматичну класифікацію текстів заміток у блогах в Інтернеті. Іншими словами, кожна замітка автоматично може бути віднесена до деякої категорії із заздалегідь визначеного набору можливих категорій. Автоматична класифікація значно поліпшить якість

пошуку за замітками, надасть додаткові можливості для розширеного пошуку. Відзначимо додаткові умови задачі:

1) категорії необов'язково є такими, що не перетинаються;
2) немає жодної додаткової інформації про категорії;
3) немає жодної зовнішньої інформації про документ (невідомими є дата публікації документу, автор, джерело тощо), іншими словами, доступна лише та інформація, яка є в самому тексті документу і, можливо, у його заголовку.

Метод розв'язання задачі

Для розв'язування задачі автоматичної класифікації текстів використано підхід машинного навчання (machine learning). Цей підхід передбачає наступні етапи:

1) вибір навчального набору документів (training-and-validation set) і тестового набору документів (test set);
2) індексація документів з навчального набору (вилучення частовживаних слів, вибір термів, визначення вагових коефіцієнтів термів);
3) зменшення розмірності задачі (вилучення надлишкових та додавання штучних термів);
4) визначення порогів (threshold);
5) застосування класифікатора;
6) оцінка результатів;

Коротко розглянемо основні етапи розв'язання задачі.

Розв'язання задачі

Навчальний набір документів – це множина документів, за допомогою якої проводиться навчання класифікатора. Часто виділяють окремо підмножину – перевірковий набір документів (validation set), – яка використовується для оптимізації параметрів класифікатора.

Перед індексацією необхідно провести первинну обробку документа. До такої обробки відноситься, перш за все, вилучення частовживаних слів (прийменників, сполучників тощо), а також стеммінг (зведення слів до однієї словоформи). Для індексації документа вводиться поняття терму та ваги, з якою він входить до даного документа. Термом (згідно [1]) може бути як окреме слово, так і деяке словосполучення або ціла фраза. Існує два підходи до вибору фраз у якості термів: синтаксичний та статистичний. Доцільність використання кожного з цих підходів окремо один від одного поставлено під сумнів в [1]. Проте, використання їх одночасно може

призвести до поліпшення результатів [3]. У якості ваги терму в більшості випадків використовується або бінарне ранжування (ваговий коефіцієнт може набувати значення 0 та 1), або міра *tfidf* (1).

$$tfidf(t_k, d_j) = c(t_k, d_j) \cdot \log \frac{|T_r|}{df_{T_r}(t_k)} \quad (1),$$

де $c(t_k, d_j)$ - кількість повторень терму t_k у документі d_j,

$df_{T_r}(t_k)$ - частота терму t_k у колекції документів T_r.

Зазначимо, що при такому підході не береться до уваги позиція терму в документі. Для ліквідації цього недоліку варто надавати більшу вагу термам, що знаходяться у назві документа.

На практиці розмірність задачі може сягати значних обсягів. Тому для спрощення розв'язання задачі доцільно зменшити розмірність задачі. Є два підходи до цього: локальний (для кожної категорії оптимізується набір термів окремо) та глобальний (набір термів оптимізується для всіх категорій одночасно). Під оптимізацією набору термів мається на увазі видалення термів та створення штучних термів. Видалення термів краще всього проводити за алгоритмом TSR [4]. Можна без втрати точності моделі зменшити розмірність моделі у 10 разів, а з невеликою втратою точності – до 100 разів [4]. Штучні терми варто вводити для виключення проблеми омонімів, синонімів, та багатозначності термів. Серед алгоритмів введення штучних термів можна виділити алгоритм кластеризації термів.

Для побудови точного класифікатора необхідно обробити результати роботи ранжувального класифікатора так, щоб можна було визначити, чи належить документ категорії. Для цього вводиться поняття порога (threshold). Для визначення порога в [1] запропоновано використовувати пропорційний метод і метод *k* найближчих категорій. Кращих результатів можна досягти, поєднуючи ці два алгоритми.

Як класифікатор було використано модифікований наївний Байєсівський алгоритм [2], з модифікаціями для *n* категорій.

Оцінка результатів роботи класифікатора може бути отримана за допомогою двох параметрів – повноти (2) та точності (3).

$$p = \frac{TP}{TP + FN} \quad (2)$$

$$c = \frac{TP}{TP + FP} \quad (3),$$

де *TP* – кількість правильно класифікованих документів,

FN – кількість документів, які були не віднесені до певної категорії, хоча насправді вони їй належать,

FP – кількість документів, які неправильно були віднесені до певної категорії.

Приклади застосування

Розроблена методика буде застосовуватися для категоризації записів у блогах в Інтернеті. Крім того, вона може використовуватися для класифікації будь-яких текстів, наприклад, інформаційних повідомлень, публікацій у журналах, текстового наповнення веб-сторінок тощо.

Висновки

Проблема автоматичної класифікації текстів є актуальною. Є велика кількість галузей, у яких може бути використана автоматична класифікація. Розглянута методика дозволяє успішно замінити ручну категоризацію текстів автоматичною. Розроблені модифікації існуючих алгоритмів дозволяють істотно поліпшити результати класифікації.

У подальшому описаний метод може бути вдосконалений завдяки поліпшеному вибору термів на етапі індексації документа. Проблематичним є використання описаної методики для нетекстових даних, наприклад, для зображень. Це пояснюється тим, що для нетекстових даних є дуже велика кількість різноманітних поєднань термів, в той час як для текстових даних така кількість є порівняно невеликою. Але якщо буде вирішена проблема індексації мультимедійних даних, описану методику можна буде застосувати і для нетекстових даних.

Література

1. *Fabrizio Sebastiani.* Machine learning in automated text categorization. // ACM Computing Surveys (CSUR). Volume 34, Issue 1 (March 2002). – P. 1 – 47.

2. *Gary Robinson.* A statistical approach to the spam problem. // Linux Journal, Volume 2003, Issue 107 (March 2003). – P. 3.

3. *Tzeras K., Hartmann S.* Automatic indexing based on Bayesian inference networks. // In Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval (Pittsburgh, PA, 1993). – P. 22–34.

4. *Yang Y., Pedersen J. O.* A comparative study on feature selection in text categorization. // In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN, 1997). – P. 412–420.