УДК 519.68; 681.513.7; 612.8.001.57; 007.51/.52

АНАЛІЗ МЕТОДІВ ВИЗНАЧЕННЯ ВАГ ОЗНАК ТЕКСТОВИХ ДОКУМЕНТІВ

Олійник Ю. О., Катющенко Д. О.

НТУУ «КПІ ім. Ігоря Сікорського», Україна, Київ

Робота присвячена розвитку методів визначення ваг ознак документа при розв'язанні задачі автоматичної класифікації текстової інформації. Аналізується вплив зменшення розмірності ознак документа на роботу векторного класифікатора. В якості пропонованих методів розглядаються TF-IDF, TF-SLF, покрапрова взаємна інформація, умовні випадкові поля.

Мета даної роботи полягає в підвищенні якості класифікації текстової інформації за рахунок використання доцільного методу визначення ваг ознак документу та його поєднання з методом побудови класифікатора.

В статті виконаний зрівняльний аналіз методів за такими характеристиками, як повнота, точність та F-міра.

Розглянуті методи використовуються для вирішення задач визначення тематичної приналежності текстів, визначення автора документу, емоційного забарвлення, фільтрації спаму тощо.

Ключові слова: інтелектуальний аналіз даних, класифікація текстової інформації, аналіз контенту, машинне навчання, алгоритми класифікації.

Олейник Ю. А., Катющенко Д.А. Анализ методов определения веса признаков текстовых документов/ НТУУ «КПИ им. Игоря Сикорского», Украина, Киев.

Работа посвящена развитию методов определения весов признаков документа при решении задачи автоматической классификации текстовой информации. Рассматривается влияние уменьшения размерности признаков документа на работу векторного классификатора. В качестве предлагаемых методов рассматриваются TF-IDF, TF-SLF, поточечная взаимная информация, условные случайные поля.

Цель данной работы заключается в повышении качества классификации текстовой информации за счет использования целесообразного метода определения весов признаков документа и его сочетание с методом построения и обучения классификатора.

В статье выполнен сравнительный анализ методов по таким характеристикам, как полнота, точность и F-мера.

Рассмотренные методы применяются при решении задач определения тематической принадлежности текстов, определение автора документа, определение эмоциональной окраски документа, фильтрации спама и т.д.

Ключевые слова: интеллектуальный анализ данных, классификация текстовой информации, анализ контента, машинное обучение, алгоритмы классификации

Oliynik Yuri, Katiushchenko Daria Analysis of methods of determining the term weight at textual documents/ National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, Kyiv.

The work is devoted to the development of methods for determining the term weight of the document during automatic classification of text information. The influence of diminishing the dimension of a document terms on the work of vector classifier is considered. In the quality of the proposed methods are considered such methods as TF-IDF, TF-SLF, pointwise mutual information, conditional random fields.

The purpose of this work is to improve the quality of the classification of textual information due to the fact that the appropriate method for determining the weight of the document is documented, and their combination with the method will induce the beginning of the classifier.

The comparative analysis of methods on characteristics such as precision, recall and F-measure were performed.

The considered methods are part of solution of determining the thematic belonging of texts, determining the author of the document, determining the emotional color of the document, spam filtering, etc.

Key words: data mining, classification of textual information, content analysis, machine learning, classification algorithms.

Вступ. Потреба в автоматичній обробці текстових документів зараз є надзвичайно високою, і постійно зростає. Це обумовлено щоденним збільшенням текстової інформації на просторах Всесвітньої мережі. За даними на березень 2016 року в Інтернеті знаходиться близько 4,66 млрд сторінок, при чому ця цифра включає лише сторінки, які індексовані в найбільш розповсюджених пошукових системах. Тож, без комп'ютерної обробки виконати аналіз такого об'єму інформації за прийнятний час не можливо.

Одною із задач інтелектуального аналізу текстів є їх класифікація на задані категорії, яка знаходить використання в різних сферах людської діяльності. Так, для забезпечення інформаційної та суспільної безпеки, важливе значення має аналіз даних соціальних мереж, блогів тощо, з метою виявлення даних пов'язаних з тероризмом, наркоторговлею і т.д. Також в комерційній та суспільній діяльності часто постає потреба обробки відгуків та коментарів, з метою виявлення їх емоційного забарвлення, визначення тематичної приналежності текстів тощо.

Частіше всього постає задача класифікації текстової інформації між категоріями в умовах обмеженості за часом та ресурсами обчислювальних пристроїв. Обчислювальна складність різноманітних методів класифікації, напряму залежить від розмірності множини ознак. Тому метою даної роботи є аналіз впливу алгоритмів зменшення розмірності ознак на ефективність класифікації.

Формальна постановка задачі. Задача класифікації полягає в наступному: існує множина документів $D = \{d_1, \dots, d|D|\}$та множина можливих категорій (класів)$C = \{c_1, \dots, c|C|\}$. Є невідома функція $F: D \times C \to \{0, 1\}$, яка для кожної пари <документ, категорія> визначає, чи відповідають вони один одному. Задача полягає в тому, що необхідно знайти максимально близьку до функції F функцію F' - класифікатор. [1]

Класифікацію розділяють на однозначну (якщо в задачі кожному документу d∈ D, може відповідаті лише одна категорія $c \in C$) та багатозначну (якщо кожному документу d∈ D, може відповідати довільна кількість класів). [2]

Загальна схема класифікації. Рішення задачі класифікації складається з 4 етапів: попередня обробка та індексація документів, визначення вагомості ознак документів (зменшення розмірності множини ознак), побудова та навчання класифікатора, оцінка якості класифікації (рисунок 1).

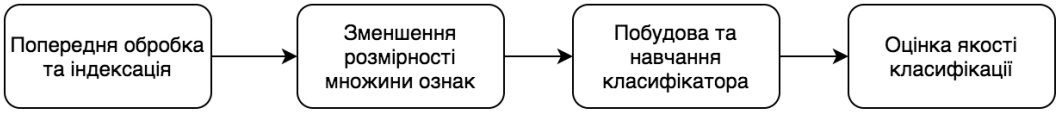

Рис. 1 Загальна схема рішення задачі автоматичної класифікації текстів

На етапі попередньої обробки та індексації документів формуються ознаки документа, в якості яких виступають всі його значимі слова або словосполучення.[3] Цей етап включає в себе токенізацію, тобто розбиття тексту на більш дрібні об'єкти, наприклад, речення, фрази або слова; видалення функціональних слів (семантично нейтральних, таких як сполучники, прийменники, артиклі тощо) та морфологічний аналіз, тобто виявлення частин мови та стематизація або лематизація.[3]

Зменшення розмірності множини ознак - це процес надання ваги словам, в залежності від їх важливості для класифікації тексту, та подальше видалення маловагомих термів з множини ознак.

Зазвичай по завершенню цього етапу документ представляється у вигляді вектору $\underline{d_j} = (w_{1j}, \dots, w_{|T|j})$ в деякому просторі ознак - Т, в якому кожному терму (ознаці) ставиться у відповідність його вага - w_{ij} .

Для видалення термів встановлюється порогова вага, терми нижче якої вважаються не важливими.

За рахунок зменшення розмірності множини термінів можна знизити ефект перенавчання - явище за якого класифікатор орієнтується на випадкові або помилкові характеристики навчаючих даних, а не на насправді важливі.[2]

Після вибору ознак переходять до побудови та навчання класифікатора. Загалом класифікатори поділяються на лінійні та нелінійні. Лінійні класифікатори побудовані на простій лінійній комбінації функцій, які поділяють простір ознак на регіони за допомогою лінійних гіперплощин.

Нелінійні - більш складні класифікатори, але не завжди кращі на лінійні, через зміщення дисперсії. Такі моделі мають багато параметрів, які підходять для обмеженої кількості навчальних даних,

але з більшою ймовірністю припустяться помилки на невеликих та шумних наборах даних.

Методи визначення ваг ознак документа.

До найбільш розповсюджених методів визначення ваг ознак документа відносяться: TF-IDF [4] , TF-SLF[4] покрапкова взаємна інформація (PMI)[4], умовні випадкові поля (CRF)[5].

TF-IDF та PMI розглядають важливість терма в рамках всього корпусу документів. За такої обробки ігнорується важливість терма в рамках окремо взятої категорії. TF-SLF, який заснований на наступних положеннях, дозволяє подолати це обмеження:

1. терм є важливим в рамках категорії, якщо він зустрічається в більшості текстів даної категорії;
2. оцінка терма зменшується якщо він є важливим для декількох категорій.

Детальний опис розрахунку SLF приведений в [4, с. 13].

Також відзначимо, що недоліком PMI є те що термів з однаковою умовною ймовірністю, рідкі терми матимуть більшу оцінку ніж загальні терми. Відповідно, оцінки не можна порівнювати за умови широко визначеної частоти.

Алгоритм умовних випадкових полів (Conditional Random Feilds - CRF) - це неспрямована ймовірнісна графічна модель, яка кодує умовні розподілення ймовірностей з заданим набором особливостей. Схема роботи цього алгоритму на прикладах наведена в [7]. Відзначимо лише, що основними перевагами цього алгоритму є відносно висока якість результатів при автоматичному виділенні сутностей та ключових слів та гнучкість у виборі ознак для навчання, при чому наявність умовної незалежності змінних не є обов’язковою умовою. А основними недоліками CRF, по-перше, є обчислювальна складність аналізу навчальної вибірки, що ускладнює постійне

оновлення моделі при надходжені нових навчальних даних; по-друге, при виділенні ключових слів CRF не працює з невідомими словами (тобто зі словами, які не зустрічались в навчальній вибірці).

Методи побудови та навчання класифікатора. Методи побудови та навчання класифікатора можна розділити на: ймовірнісні (наївний байєсівський метод (Naive Bayes, MaxEnt), засновані на нейронних мережах, лінійні (SVM, логістична регресія) та метричні (k-NN, Rocchio Classifier).

Кожен з цих методів більш доцільний в порівнянні з іншими в залежності від виду поставленої задачі. (Детальний опис більшості перерахованих методів можна знайти за посиланням [2]). Так k-NN дає хороші результати у випадку з лінійно нероздільними вибірками, але не підходить для вирішення задач великої розмірності за кількістю класів та документів. SVM в свою чергу є одним з найшвидших класифікаторів, зводиться до вирішення задачі квадратичного програмування, яке завжди є єдиним, але дуже чутливий до шумів та стандартизації даних.

Для порівняння методів визначення ваг ознак документу в якості методу побудови класифікатора було обрано класифікатор Роші, який відноситься до векторних моделей. Він базується на гіпотезі суміжності: документи в одному класі з суміжного регіону та регіони різних класів не перетинаються.

Класифікатор Роші розподіляє векторний простір ознак на регіони відносно центроїду (один для кожного класу), обчислений як центр мас всіх документів класу. Межі між класами в загальному вигляді для M-вимірного простору є гіпер площинами. Текст, який класифікується, буде віднесений до того класу, до центроїду якого він найближче знаходиться. Зазвичай для визначення відстані

обчислюють косинус між векторами. Детальний алгоритм побудови та навчання класифікатора наведений в [6].

Класифіктор Роші простий та ефективний, але може бути не точним, якщо класи не є приблизно сферами зі схожими довжинами радіусів.

Ефективність алгоритмів. Для оцінки більшості алгоритмів вилучення інформації зазвичай використовують метрики точності(precision) та повноти(recall), або їх модифікації. Точність – це доля документів, які дійсно належать даному класу відносно всіх документів, яких система віднесла до даного класу. Повнота системи – це доля знайдених класифікатором документів, які належать класу, відносно всіх документів цього класу в тестовій виборці.

Для оцінки ефективності роботи алгоритмів використано набір даних моніторингового проекту згадувань України у закордонних медіа “ОКО” [7].

Один із способів розрахунку точності та повноти - через таблиці контингентності, які будуються для кожного класу окремо [3]. Але на практиці для визначення точності та повноти часто використовують матриці неточностей (confusion matrix). Матриця неточностей - це матриця розмірності N на N, де N - кількість класів. Стовпці відповідають експертним рішення, а рядки - результатам роботи класифікатора. Під час класифікації документу з тестової вибірки, інкрементуємо число, яке стоїть на перетині рядка класу, який був визначений класифікатором та стовпчику класу, до якого насправді відноситься документ.

Нижче наведені матриці неточностей результатів роботи класифікатора Роші в якому в якості визначення ваг ознак документу були: TF-IDF (рис. 2), TF-SLF (рис. 3) та PMI (рис. 4).

Category		1	2	3	4
		1.0	0.879	0.75	0.968
1	1.0	11.0	0.0	0.0	0.0
2	0.784	0.0	29.0	7.0	1.0
3	0.948	0.0	1.0	36.0	1.0
4	0.883	0.0	3.0	5.0	60.0

Рис. 2 Матриця неточностей Роші класифікації з використанням TF-IDF

Category		1	2	3	4
		1.0	0.94	0.834	0.952
1	1.0	11.0	0.0	0.0	0.0
2	0.816	0.0	31.0	5.0	2.0
3	0.931	0.0	2.0	40.0	1.0
4	0.952	0.0	0.0	3.0	59.0

Рис. 3 Матриця неточностей Роші класифікації з використанням TF-SLF

Category		1	2	3	4
		1.0	0.304	0.709	1.0
1	0.734	11.0	1.0	3.0	0.0
2	1.0	0.0	10.0	0.0	0.0
3	0.895	0.0	4.0	34.0	0.0
4	0.682	0.0	18.0	11.0	62.0

Рис. 4 Матриця неточностей Роші класифікації з використанням PMI

З такої матриці точність та повнота для кожного класу розраховується досить легко: точність дорівнює співвідношенню відповідного діагонального елементу матриці до суми усього рядка, а повнота - співвідношенню діагонального елементу до суми всього стовпчику:

$$Precision_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}},$$

$$Recall_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}}.$$

Результуюча точність, як і повнота, класифікатора розраховується як середнє арифметичне його точності по класам.

Для об’єднання цих двох метрик в одну та визначення загальної ефективності алгоритму часто використовують метрику F-міри, яка є гармонічним середнім між точністю та повнотою[8].

Тож для алгоритмів які розглядаються, отримаємо:

Таблиця 1

Порівняльна таблиця результуючої ефективності використання методів визначення ваг ознак документа з класифікатором Роші

	Rocchio and TF-IDF	Rocchio and TF-SLF	Rocchio and PMI
Precision	0.903	0.924	0.827
Recall	0.899	0.931	0.753
F-міра	**0.901**	**0.927**	**0.788**

Як видно з таблиці 1 найкращі результати класифікації були отримані при застосуванні TF-SLF в якості методу визначення ваг ознак документа, в порівнянні із PMI та TF-IDF.

Висновок.

В даній роботи розглянуто методи зменшення розмірності множини ознак при вирішенні задачі автоматичної класифікації текстової інформації. В якості методів визначення ваг ознак документу розглядалися TF-IDF, TF-SLF, PMI та CRF. Відзначимо, що для кожної окремої задачі необхідно окремо підбирати та комбінувати методи кожного з етапів розв’язку задачі класифікації, оскільки ефективність одних методів в порівнянні з іншими залежить від розмірності задачі, роздрібності класів та типу класифікації.

Велика множина ознак в деяких алгоритмах значно збільшує час навчання, а час роботи не змінюється (нейроні мережі), в інших - час навчання залишаються незмінним, але значно зростає час роботи (наприклад SVM). Для уникнення ефекту перенавчання кількість прикладів з навчальної вибірки приблизно не повинна значно

відрізнятись від кількості термів. Але все ж таки задача з визначенням кількості термів є все ще не вирішеною.

Література:

1. Joachims T. "Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms" / Thorsten Joachims. – USA, MA: Kluwer Academic Publishers Norwel, 2002. – 224 p.

2. Батура Т. В. Методы автоматической классификации текстов / Т. В. Батура // Международный научно-практический журнал программные продукты и системы. – 2017. – №1 – С. 85 – 99.

3. Гавриленко О. В. Огляд та аналіз алгоритмів TEXT MINING / О.В. Гавриленко, Ю. О. Олійник, Г. В. Ханько. // Управління проектами, системний аналіз ілогістика. – К.: НТУ, 2017. – Вип.

4. Попков М.И. Автоматическая система классификации текстов для базы знаний предприятия: дис. магист. / М.И. Попков – Москва, 2014. – 57 с.

5. Zhang C. Automatic Keyword Extraction from Documents Using Conditional Random Fields / C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, B. Wang // Journal of Computational Information Systems 4:3. – 2008. – pp. 1169-1180.

6. Manning D. Christopher Introduction to information retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze – USA, NY : Cambridge University Press. – 2008. – 482 p.

7. Моніторинговий проект "ОКО" [Електронний ресурс] / Режим доступу: http://www.ukroko.org

8. Оценка классификатора (точность, полнота, F-мера) [Електронний ресурс] / Суровая реальность – 2012. – Режим доступу: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html

References:

1. Joachims T. "Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms" / Thorsten Joachims. – USA, MA: Kluwer Academic Publishers Norwel, 2002. – 224 p.

2. Batura T. V. Metody avtomaticheskoy klassifikatsii tekstov / T. V. Batura // Mezhdunarodnyy nauchno-prakticheskiy zhurnal programmnye produkty i sistemy. – 2017. – №1 – С. 85 – 99.

3. Havrylenko O. V. Ohliad ta analiz alhorytmiv TEXT MINING / O. V. Havrylenko, Yu. O. Oliinyk, H. V. Khanko. // Upravlinnia proektamy, systemnyi analiz i lohistyka. – K.: NTU, 2017. – Vyp.

4. Popkov M. I. Avtomaticheskaya sistema klassifikatsii tekstov dlya bazy znaniy predpriyatiya: dis. magisterskaya / M. I. Popkov – Moskva, 2014. – 57 c.

5. Zhang C. Automatic Keyword Extraction from Documents Using Conditional Random Fields / C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, B. Wang // Journal of Computational Information Systems 4:3. – 2008. – pp. 1169-1180.

6. Manning D. Christopher Introduction to information retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze – USA, NY : Cambridge University Press. – 2008. – 482 p.

7. Monitorynhovyi proekt "OKO" [Elektronnyi resurs] / Rezhym dostupu: http://www.ukroko.org

8. Otsenka klassifikatora (tochnost, polnota, F-mera) [Elektronnyi resurs] / Surovaya realnost – 2012. – Rezhym dostupu: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html