ISSN 2224-087X. Електроніка та інформаційні технології. 2012. Випуск 2. С. 164–172
Electronics and information technologies. 2012. Issue 2. P. 164–172

УДК 519.765:519.767:004.89

ЙМОВІРНІСНА КЛАСИФІКАЦІЯ ТЕКСТОВИХ ДОКУМЕНТІВ У ПРОСТОРІ СЕМАНТИЧНИХ ПОЛІВ

Б. Павлишенко

Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, 79005 Львів, Україна
e-mail:pavlsh@yahoo.com

Запропоновано модель класифікації текстових документів у векторному семантичному просторі на основі байесівського класифікатора. Розглянуто формування некорельованих ознак текстових документів за допомогою методу головних компонент та сингулярного розкладу матриці частот семантичних полів.

Ключові слова: інтелектуальний аналіз текстів, байесівський класифікатор, семантичні поля.

Розвиток методів класифікації текстових документів є одним із перспективних напрямів інтелектуального аналізу даних [1, 2]. У працях [3–5] наведено результати аналізу текстових масивів на підставі концепції семантичних полів. Семантичні поля розглядають як групи лексем, об'єднаних спільним поняттям. Такі групи лексем утворюють нові характеристики текстових даних, використання яких є ефективним у задачах кластеризації та класифікації текстових документів. Одна з поширених моделей в інтелектуальному аналізі текстових даних – векторна модель, у якій текстові документи зображають у вигляді векторів у деякому фазовому просторі [1]. Базис цього простору утворюють частотні характеристики лексем. У працях [3–5] текстові документи розглядають як вектори, складовими яких є частоти семантичних полів у цих документах. У праці [3] описано теоретико-множинну концепцію семантичних полів у масивах текстових даних. З'ясовано, що семантичні класи утворюються як відношення еквівалентності. Семантичне поле визначене як сегмент, утворюваний семантичними класами, об'єднаними бінарним кластером у структурному відношенні семантичного розбиття лексемного словника текстових масивів. Розглянуто відношення, яке описує розбиття словника на семантичні класи зі структурою, яка визначає семантичні поля лексемного словника. Проаналізовано утворення семантичних полів на підставі лексемних відношень, зокрема, таких як сполучення в тексті лексем семантичного поля та лексем полеутворювальної множини. Доведено, що використання концепції семантичних полів є ефективним у векторній моделі текстових документів унаслідок зменшення розмірності фазового простору відображення документів. У праці [4] запропоновано модель кластеризації текстових документів у семантичному просторі, яка дає змогу отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності, ніж простір, утворений лексемним складом текстової вибірки. Такий структурний поділ відображає класифікацію документів за новими ознаками, зокрема, за

авторством текстів. Текстові вибірки деяких авторів можуть мати свої чіткі області в семантичному просторі. Це дає змогу вивчати авторство текстових документів через аналіз належності семантичних векторів цих документів до заданих областей простору семантичних полів. У праці [5] наголошено, що сингулярний розклад матриці семантичних ознак типу "частоти_семантичних_полів–документи" дає змогу аналізувати текстові документи в новому просторі семантичних концептів. Ієрархічна кластеризація документів у такому просторі відображає класифікаційну структуру документів за різними ознаками. Розмірність простору семантичних концептів визначена рангом апроксимації матриці семантичних ознак у разі сингулярного розкладу і може бути суттєво меншою, ніж розмірність простору семантичних полів. У випадку дослідження авторства текстів вибір розмірності простору семантичних концептів зумовлений рівнем відображення класифікаційного поділу документів за авторами в кластерній структурі, що визначена наявністю переважних кластерів для документів окремих авторів. Поряд із розглянутими методами кластеризації текстових документів у просторі семантичних полів перспективним є розвиток методів класифікації текстових документів із використанням концепції семантичних полів.

Розглянемо зображення текстових документів у вигляді векторів у фазовому просторі семантичних полів. Проаналізуємо головні принципи семантичної класифікації текстових документів. Опишемо принципи формування некорельованих характеристик для байєсівської класифікації текстових документів. Схарактеризуємо класифікаційний потенціал семантичних полів.

Векторна модель текстових документів у семантичному просторі. Сукупність текстових документів опишемо такою множиною:

$$D = \left\{ d_j \mid j = 0,1,2..., N_d \right\}. \tag{1}$$

Уведемо множину семантичних полів

$$S = \left\{ s_k \mid k = 1,2..., N_s \right\}. \tag{2}$$

Під семантичним полем розуміють таку множину лексем, які об'єднані певним спільним поняттям [6, 7]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та ін. Нехай існує певний словник лексем, які трапляються у текстових масивах $W = \left\{ w_i \mid i = 1,2..., N_w \right\}$. Уведемо відображення лексемного складу словника W на множину семантичних полів S за допомогою деякого оператора U_{ws}:

$$U_{ws} : w_i \to s_k, i = 1,2..., N_w; k = 1,2..., N_s. \tag{3}$$

Оператор U_{ws} задамо таблицею, яку визначимо експертним лексикографічним аналізом. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\to} s_k, i = 1,2..., N_w \right\}. \tag{4}$$

Уведемо оператор відображення семантичного складу текстового документа d_j на множину квантитативних ознак :

$$U_{sd} : s_k \to p_{kj}^{sd}, k = 1,2,...N_s, j = 1,2...N_d \tag{5}$$

Величина p_{kj}^{sd} визначає структурну частоту лексем семантичного поля s_k у текстовому документі d_j. Визначимо p_{kj}^{sd} за такою формулою:

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \quad f_s(w_i, s_k) = \begin{cases} 1, w_i \in W_k^s \\ 0, w_i \notin W_k^s \end{cases}, \tag{6}$$

де p_{ij}^{wd} – текстова частота лексеми w_i в документі d_j, яка визначена відношенням наявної кількості лексеми w_i до загальної кількості лексем у документі d_j. Сукупність значень p_{kj}^{sd} утворює матрицю ознака–документ, у якій ознаками є частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d}. \tag{7}$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}\right) \tag{8}$$

відображає документ d_j в N_s-вимірному просторі текстових документів.

Байесівський класифікатор текстових документів у семантичному просторі. Нехай існують деякі категорії текстових документів. Ці категорії можуть мати різну природу, наприклад, можуть визначати авторський ідеолект, дискурс, характеризувати різні об'єкти, явища, події тощо. Множину цих категорій позначимо

$$Categories = \{Ctg_m \mid m = 1,2,..., N_{ctg} \ \}, \tag{9}$$

де $N_{ctg} = |\ Categories\ |$ визначає розмір множини категорій. За цими категоріями розподілені текстові документи множини D (1). Завдання полягає в пошуку цільової функції, яку описує відображення

$$F_{d\to ctg} : Categories \times D \to \{0,1\} \tag{10}$$

З використанням імовірнісних методів класифікації можна отримати апроксимацію функції $F_{d\to ctg}$ наближеною функцією

$$\widetilde{F}_{d\to ctg} : Categories \times D \to [0,1]. \tag{11}$$

Значення наближеної функції $\widetilde{F}_{d\to ctg}$, на відміну від $F_{d\to ctg}$, є не дискретними, а неперервними і відображають відповідні апостеріорні ймовірності. У методах текстової класифікації на підставі наївного байесівського класифікатора використовують зображення документів за допомогою частот відповідних ключових слів [8]. Підхід, який ґрунтується на відображенні документів частотними характеристиками семантичних полів, є перспективним з огляду на меншу розмірність семантичного фазового простору. В деяких випадках класифікації аналіз текстів у просторі семантичних полів більш диференційований. Враховуючи різні підходи у формуванні семантичних полів, можна

підібрати такі зважені комбінації частот лексем в семантичних полях, які матимуть найбільший категорійно-розділювальний потенціал.

Знайдемо апостеріорну ймовірність того, що за деяким набором частот семантичних полів документ d_j відноситься до категорії ctg_m. За теоремою Байеса визначимо

$$P(ctg_m \mid p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) = \frac{P(ctg_m)P(p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd} | ctg_m)}{P(p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd})} \tag{12}$$

Чисельник можна розглядати як спільну ймовірність розподілу тексту за категорією та розподілу частот семантичних полів:

$$\begin{aligned} &P(ctg_m)P(p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd} | ctg_m) = P(ctg_m, p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) = \\ &= P(ctg_m) \cdot P(p_{1j}^{sd} | ctg_m) \cdot P(p_{2j}^{sd} | ctg_m, p_{1j}^{sd}) \cdots P(p_{N_s j}^{sd} | ctg_m, p_{1j}^{sd} p_{2j}^{sd},..., p_{N_s - 1 j}^{sd}) \end{aligned} \tag{13}$$

Формулу (12) складно застосувати на практиці внаслідок складності розрахунку розподілів умовних ймовірностей (13). У реалізації наївного байесівського класифікатора роблять суттєве припущення про умовну незалежність ознак об'єктів [8]. У такому випадку спільний розподіл категорій та семантичних полів можна виразити так:

$$P(ctg_m, p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) = P(ctg_m) \cdot \prod_{i=1}^{i=N_s} P(p_{ij}^{sd} | ctg_m) \tag{14}$$

Неперервні розподіли $P(p_{ij}^{sd} | ctg_m)$ можна апроксимувати нормальним гаусовим розподілом. Параметрами цього розподілу можна розглядати математичне сподівання та дисперсію семантичних полів. Доповненням до розрахунку наївного байесівського класифікатора є правило прийняття рішень про віднесення аналізованого документа до тієї чи іншої категорії [8]. У найпростішому випадку таке правило може приймати рішення про належність документа до заданої категорії, якщо розрахована апостеріорна ймовірність для такої категорії за заданих частот семантичних полів є найбільшою, тобто

$$\begin{aligned} &Category(d_j) = ctg_m : P(ctg_m \mid p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) = \\ &= \max\left\{ P(ctg_k \mid p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) \mid k = 1,2,...., N_{ctg} \right\} \end{aligned} \tag{15}$$

Ймовірності $P(p_{ij}^{sd} | ctg_m)$, які використовують у формулі (14), формуються на деякому навчальному категоризованому масиві текстових документів. Якщо цей масив недостатньо об'ємний, то для деяких семантичних полів і категорій можна отримати нульові значення ймовірностей. У такому випадку ці семантичні поля можна вилучити з розгляду як поля з недостатнім категорійно-розділювальним потенціалом на навчальній вибірці або вносити деяку ненульову поправку в імовірність, щоб уникнути множення на нуль у разі розрахунку повної ймовірності. Оскільки, згідно із формулою Байеса (12), усі апостеріорні ймовірності розраховують з однаковим знаменником, то його можна вилучити з розгляду і вважати

$$\begin{aligned} &Category(d_j) = ctg_m : P(ctg_m \mid p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) = \\ &= \max\left\{ P(ctg_m) \cdot \prod_{i=1}^{i=N_s} P(p_{ij}^{sd} | ctg_m) \mid k = 1,2,...., N_{ctg} \right\}. \end{aligned} \tag{16}$$

З погляду зручності розрахунків можна замість добутку розглядати суму логарифмів імовірностей, враховуючи те, що логарифм є монотонно зростаючою функцією. Тоді отримаємо

$$Category(d_j) = ctg_k : ctg_k = \underset{ctg_m \in Categories}{\operatorname{argmax}} \left[\lg\left(P(ctg_m)\right) + \sum_{i=1}^{N_s} \lg\left(P(p_{ij}^{sd} | ctg_m)\right) \right] \quad (17)$$

Очевидно, що в загальному експертному методі формування семантичних полів частоти деяких полів можуть корелювати між собою, тому припущення, згідно з яким спільна ймовірність розподілу тексту за категорією та розподілу частот семантичних полів визначена формулою (14), може привести до суттєвих похибок наївного байєсівського класифікатора. Розглянемо можливість формування незалежних семантичних ознак документів на підставі лінійних комбінацій семантичних полів, використовуючи метод головних компонент. Такі ознаки мають відповідати ортогональним базисним осям у векторному семантичному просторі і бути максимально незалежні одні від одних.

Формування множини семантичних ознак текстових документів. Задачу формування незалежних семантичних характеристик розглянемо як реалізацію перетворення до нового базису, який описуватиме діагональна коваріаційна матриця. Такий базис може бути утворений за допомогою перетворення Карунена–Лоева, яке є в основі методу головних компонент [9]. Розглянемо це перетворення для просторового базису, утвореного частотними характеристиками семантичних полів. Коваріаційну матрицю розглянемо у вигляді

$$Cov_s = [\mathrm{cov}_{ij}^s], \ \mathrm{cov}_{ij}^s = \mathrm{cov}(p_i^s, p_j^s) = E[(p_{il}^{sd} - E(p_{il}^{sd}))(p_{jl}^{sd} - E(p_{jl}^{sd}))]. \quad (18)$$

Під знаком E маємо на увазі математичне сподівання. Враховуючи вибірку текстових документів, запишемо

$$\mathrm{cov}_{ij}^s = \frac{1}{N_d - 1} \sum_{l=1}^{N_d} (p_{il}^{sd} - \overline{p}_i^{sd})(p_{jl}^{sd} - \overline{p}_j^{sd}), \quad \overline{p}_i^{sd} = E(p_{il}^{sd}), \quad \overline{p}_j^{sd} = E(p_{jl}^{sd}). \quad (19)$$

Для складових частотних векторів $V_j'^s$, які описують незалежні семантичні ознаки, повинна виконуватись умова

$$\mathrm{cov}(p_i'^s, p_j'^s) = 0, i \neq j. \quad (20)$$

Частотні вектори семантичних векторів та незалежних ознак пов’язані такими співвідношеннями:

$$V_j = A_s V_j', \ V_j' = A_s^T V_j. \quad (21)$$

Матриця A_s формується з власних векторів коваріаційної матриці Cov_s. У загальному випадку метод головних компонент можна розглядати як спектральний розклад коваріаційної матриці частотних характеристик семантичних полів. Задачу про спектральний розклад коваріаційної матриці Cov_s можна звести до задачі сингулярного розкладу матриці частоти_семантичних_полів-документи M_{sd} (7). Сингулярний розклад матриці терми-документи є в основі латентно-семантичного аналізу текстів [10, 11]. Нехай існує

матриця типу "частоти_семантичних_полів–документи" M_{sd}, яку описує формула (7). Вектор V_j (8) відображає документ d_j в N_s-вимірному просторі текстових документів. Добуток двох векторів $(V_p)^T V_q$ визначає кількісну міру близькості цих векторів у N_s-вимірному семантичному просторі текстових документів. Відповідно, добуток матриць $(M_{sd})^T M_{sd}$ містить скалярні добутки векторів $(V_p)^T V_q$ усіх документів і відображає їхні кореляції в просторі семантичних векторів. Нехай існує сингулярна декомпозиція матриці

$$M_{sd} = U_{sd}\Sigma_{sd}\mathrm{Y}_{sd}{}^T. \tag{22}$$

Тоді добуток матриць $(M_{sd})^T M_{sd}$ можна розглянути у вигляді

$$(M_{sd})^T M_{sd} = \left(U_{sd}\Sigma_{sd}\mathrm{Y}_{sd}{}^T\right)^T\left(U_{sd}\Sigma_{sd}\mathrm{Y}_{sd}{}^T\right) = \mathrm{Y}_{sd}\Sigma_{sd}{}^T\Sigma_{sd}\mathrm{Y}_{sd}{}^T. \tag{23}$$

Відповідно до теорії сингулярного розкладу матриць [10, 11], діагональна матриця Σ_{sd} містить сингулярні числа в порядку їх спадання. Якщо взяти K найбільших сингулярних чисел матриці Σ_{sd} і, відповідно, K сингулярних векторів матриць U_{sd} і Y_{sd}, то отримаємо K-рангову апроксимацію матриці M_{sd}:

$$(M_{sd})_K = (U_{sd})_K(\Sigma_{sd})_K(\mathrm{Y}_{sd})_K{}^T. \tag{24}$$

Матриця $(\mathrm{Y}_{sd})_K$ відображає зв'язок між векторами документів V_j' у новому комбінованому K-вимірному семантичному просторі з ортонормованим семантичним базисом. Зв'язок між вектором V_j документа в первинному семантичному просторі та вектором V_j' у просторі ортонормованих семантичних ознак можна описати так:

$$\begin{aligned} V_j &= (U_{sd})_K(\Sigma_{sd})_K V_j', \\ V_j' &= (\Sigma_{sd})_K^{-1}(U_{sd})_K^T V_j. \end{aligned} \tag{25}$$

Отже, ранг апроксимації матриці M_{sd}, який визначений числом K, також визначає розмірність простору ортонормованих семантичних ознак. Очевидно, що число K може бути суттєво меншим від розмірності N_s початкового семантичного простору. Це зменшує розмірність задач аналізу текстових документів у семантичному векторному просторі.

Класифікаційний потенціал семантичних полів. У задачах категоризації текстових документів одні семантичні поля можуть відігравати суттєву роль у категоріальній диференціації, а в інші можуть бути менш значимими. Для визначення категорійно-диференціювального потенціалу частотних характеристик семантичних полів використаємо поняття ентропії. Невизначеність класифікаційної системи можна схарактеризувати таким виразом:

$$H(Categories) = -\sum_{i=1}^{N_{ctg}} P(ctg_i)\log_2\left(P(ctg_i)\right). \tag{26}$$

Умовну ентропію за заданого значення частоти семантичного поля розглянемо у вигляді

$$H(Categories \mid p_j^s) = -\sum_{i=1}^{N_{ctg}} P(ctg_i \mid p_j^s) \log_2\left(P(ctg_i \mid p_j^s)\right). \quad (27)$$

Умовну ентропію для семантичного поля s_j за всіх значень частот опишемо так:

$$H(Categories \mid s_j) = \int_0^1 f_j^s(p_j^s) H(Categories \mid p_j^s) dp_j^s , \quad (28)$$

де $f_j^s(p_j^s)$ – функція розподілу значень частот p_j^s семантичного поля s_j . Враховуючи (27), отримаємо

$$H(Categories \mid s_j) = -\sum_{i=1}^{N_{ctg}} \int_0^1 f_j^s(p_j^s) P(ctg_i \mid p_j^s) \log_2\left(P(ctg_i \mid p_j^s)\right) dp_j^s . \quad (29)$$

Кількість інформації, яку отримує система семантичної класифікації за наявності семантичного поля s_j,

$$H_{\inf}(s_j) = H(Categories) - H(Categories \mid s_j) \quad (30)$$

Функцію $H_{\inf}(s_j)$ можна розглядати як цільову для вибору множини класифікаційних семантичних полів. Таку множину формують на основі полів з максимальним значенням $H_{\inf}(s_j)$. Очевидно, що значення $H_{\inf}(s_j)$ розраховують для деякої вибірки текстових документів заданого класу задач. Для іншої вибірки необхідно виконати новий розрахунок таких цільових функцій, оскільки для іншого класу задач та вибірки такі функції можуть відрізнятись і оптимальна множина класифікаційних семантичних полів матиме інший склад.

Отже, розглянуто теоретико-множинну модель класифікації текстових документів у векторному семантичному просторі на основі байесівського класифікатора. Класифікаційні характеристики текстових документів у цій моделі утворюються на підставі частотного розподілу семантичних полів. Для реалізації наївного байесівського класифікатора запропоновано формувати кількісні некорельовані характеристики текстових документів за допомогою методу головних компонент та сингулярного розкладу матриці частот семантичних полів. Проаналізовано кількісні інформаційні характеристики класифікаційного потенціалу семантичних полів у текстових документах, які можна використовувати для відбору семантичних полів у задачах класифікації текстових документів.

1. *Pantel P., Turney P. D*. From Frequency to Meaning: Vector Space Models of Semantics // J. of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141–188.
2. *Брасегян А. А., Куприянов М. С., Холод И. И.* и др. Анализ данных и процессов: учеб. Пособие. – СПб.: БХВ–Петербург, 2009. – 512 с.:ил.
3. *Павлишенко Б. М*. Використання концепції семантичного поля у векторній моделі текстових документів // Східно-Європ. журн. передових технологій. – 2011. – № 6/2(54). – С. 7–11.
4. *Павлишенко Б. М*. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів // Електроніка та інформ. технології. – 2011. – Вип. 1 .– С. 212–222.
5. *Павлишенко Б. М*. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Матем. машини і системи. – 2012. – № 1. – С. 69–76.
6. *Левицкий В. В., Стернин В. В*. Экспериментальные методы в семасиологии. – Воронеж: Изд-во ВГУ, 1989. – 192 с.
7. *Вердиева З. Н*. Семантические поля в современном английском языке. – М.: Высшая школа, 1986. – 120 с.
8. *Sebastiani F*. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34. – N 1. – P. 1–47.
9. *Jolliffe I. T*. Principal Component Analysis. – Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002. – Vol. 29. – 487 p. 28 il.
10. *Deerwester S., Dumais S., Furnas G*. et al. Indexing by Latent Semantic Analysis // J. of the American Society for Information Science. – 1990. – Vol. 41. – Is. 6. – P. 391–407.
11. *Mirzal A*. Clustering and Latent Semantic Indexing Aspects of the Singular Value Decomposition // [Електронний ресурс] arXiv:1011.4104v2, 2011, http://arxiv.org/abs/1011.4104v2

PROBABILISTIC CLASSIFICATION OF TEXT DOCUMENTS IN THE SPACE OF SEMANTIC FIELDS

B. Pavlyshenko

Ivan Franko Lviv National University,
50 Dragomanov St., Lviv, UA–79005 Ukraine
e-mail:pavlsh@yahoo.com

A classification of text documents in the semantic vector space has been proposed using a Bayesian classifier. The formation of uncorrelated attributes of text documents has been considered, using the method of principal components and singular decomposition of the frequency matrix of semantic fields.

Key words: text mining, Bayesian classification, semantic fields.

ISSN 2224-087X. Електроніка та інформаційні технології. 2012. Випуск 2

ВЕРОЯТНОСТНАЯ КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ В ПРОСТРАНСТВЕ СЕМАНТИЧЕСКИХ ПОЛЕЙ

Б. Павлышенко

Львовский национальный университет имени Ивана Франко
ул. Драгоманова, 50, 79005 Львов, Украина
e-mail:pavlsh@yahoo.com

Предложена модель классификации текстовых документов в векторном семантическом пространстве на основе байесовского классификатора. Рассмотрено формирование некоррелированных признаков текстовых документов с помощью метода главных компонент и сингулярного разложения матрицы частот семантических полей.

Ключевые слова: интеллектуальный анализ текстов, байесовский классификатор, семантические поля.

Стаття надійшла до редколегії 03.04.2012
Прийнята до друку 19.06.2012