УДК 519.765:519.767:004.89

МОДЕЛЬ ВТОРИННИХ НЕКОРЕЛЬОВАНИХ СЕМАНТИЧНИХ ПОЛІВ ДЛЯ АНАЛІЗУ ТЕКСТОВИХ ДАНИХ

Б.М. ПАВЛИШЕНКО

Розглянуто модель некорельованих вторинних семантичних полів утворених на основі методу головних компонент та сингулярного розкладу матриці частот семантичних полів. Ця модель характеризує новий семантичний простір відображення текстових документів із ортонормованим базисом. Розмірність простору вторинних семантичних полів є суттєво меншою за розмірність простору первинних семантичних полів внаслідок заміни взаємопов'язаних складових некорельованими семантичними характеристиками. Аналіз тестової вибірки текстових документів показав можливість брати до розгляду лише ті складові вторинних семантичних полів, які описуються першими сингулярними числами. Використання низькорозмірного ортонормованого базису вторинних семантичних полів може бути ефективним в задачах класифікації та кластеризації текстових даних.

ВСТУП

Однією з поширених моделей в інтелектуальному аналізі текстових даних є векторна модель, в якій текстові документи представлені у вигляді векторів у деякому фазовому просторі [1]. Базис цього простору утворюють частотні характеристики лексем. У роботах [2–5] наведено результати аналізу текстових масивів на основі концепції семантичних полів. Семантичні поля розглянуто як групи лексем, об'єднаних спільним поняттям. Такі групи лексем утворюють нові характеристики текстових даних, використання яких є ефективним у задачах кластеризації та класифікації тектстових документів. Формування додаткових семантичних ознак на основі концепції семантичних полів утворює новий семантичний простір, що збільшує можливості аналізу векторного простору текстових документів. Основним методом формування семантичних полів є експертний метод лексикографічного аналізу. В такому методі неможливо сформувати структуру семантичних полів так, щоб вони були не зв'язані між собою і не корелювали у статистичних розподілах алгоритмів аналізу текстових даних. Однак, припустимо, що внаслідок лінійної комбінації частотних характеристик семантичних полів можна утворити нові семантичні поля, частотні характеристики яких будуть не корельовані. Такі поля назвемо некорельованими вторинними семантичними полями. Утворення нових некорельованих вторинних семантичних полів оптимізує задачі аналізу текстових даних та зменшує розмірність семантичного простору текстових документів. Задача зводиться до представлення текстових документів у новому семантичному ортонормованому базисі. Очевидно, що в такому базисі коефіцієнти коваріації між різними частотними складовими текстових документів будуть дорівнювати нулю.

ПОСТАНОВКА ЗАДАЧІ

Мета роботи — розглянути модель некорельованих вторинних семантичних полів утворених на основі методу головних компонент та сингулярного розгладу матриці частот семантичних полів.

Проаналізуємо коваріаційну матрицю для частотних характеристик семантичних полів. Використовуючи перетворення Карунена-Лоева в методі головних компонент визначимо матрицю перетворення семантичних векторів у вторинні некорельовані семантичні вектори.

Проаналізуємо зв'язок між методом головних компонент та сингулярним розкладом у задачі формування ортонормованого семантичного простору. Проаналізуємо сингулярні числа для тестового масиву текстових документів.

МОДЕЛЬ ТЕКСТОВИХ ДОКУМЕНТІВ У СЕМАНТИЧНОМУ ПРОСТОРІ

Розглянемо модель на основі теорії множин, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Нехай існує певний словник лексем, які зустрічаються у текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \{w_i \mid i = 1,2,\ldots,N_w\}. \tag{1}$$

Сукупність текстових документів опишемо множиною

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

Під документом з $j = 0$, будемо вважати документ з нейтральним текстом, який відповідає лінгвостатистичні нормі. Введемо множину семантичних полів

$$S = \{s_k \mid k = 1,2,\ldots,N_s\}. \tag{3}$$

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 тощо.

Документ d_i з множини текстових документів D можна представити як упорядковану множину слів, порядок елементів якої відповідає порядку слів у цьому документі

$$T_j^d = \{t_{lj} \mid l = 1,2,\ldots,N_j^t\}. \tag{4}$$

Упорядкований за алфавітом словник текстового документа d_i розглянемо як мультимножину W_j^d над множиною словника W:

$$W_j^d = \{n_{ij}^{wd}(w_i) \mid w_i \in d_j,\ i = 1,2,\ldots,N_w\}, \tag{5}$$

де n_{ij}^{wd} — кількість входжень лексеми w_i із словника W у множину лексем текстового документа d_j, яку можна визначити як

$$n_{ij}^{wd} = \sum_{l=1}^{N_j^t} f_{wd}(t_{lj}, w_i), \tag{6}$$

де
$$f_{wd}(t_{lj}, w_i) = \begin{cases} 1, & t_{lj} = w_i, \\ 0, & w_{lj}^d \neq w_i. \end{cases} \tag{7}$$

Введемо відображення лексемного складу словника W на множину семантичних полів S за допомогою деякого оператора U_{ws}

$$U_{ws} : w_i \to s_k, \; i = 1,2,\ldots,N_w; \; k = 1,2,\ldots,N_s. \tag{8}$$

Оператор U_{ws} задамо таблицею, яка визначається експертним лексикографічним аналізом [6, 7]. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\to} s_k, i = 1,2,\ldots,N_w \right\}. \tag{9}$$

Множину образів відображення U_{ws} розглянемо як мультимножину над множиною семантичних полів S:

$$S_f = \{ n_k^s(s_k) \mid k = 1,2,\ldots,N_s \}, \tag{10}$$

де n_k^s — кількість лексем словника W, які відносять до семантичного поля s_k:

$$n_k^s = \sum_{i=1}^{N_w} f_s(w_i, s_k), \tag{11}$$

де
$$f_s(w_i, s_k) = \begin{cases} 1, \; w_i \in W_k^s, \\ 0, \; w_i \notin W_k^s \; . \end{cases} \tag{12}$$

Введемо мультимножину образів відображення U_{ws} семантичних полів для окремого документа d_j:

$$S_j^d = \left\{ n_{kj}^{sd}(s_k) \mid k = 1,2...,N_s \right\}, \tag{13}$$

де n_{kj}^{sd} — кількість лексем семантичного поля s_k в лексемному складі документа d_j

$$n_{kj}^{sd} = \sum_{l=1}^{N_j^t} f_s(t_{lj}, s_k), \tag{14}$$

де
$$f_s(t_{lj}, s_k) = \begin{cases} 1, \; t_{lj} \in W_k^s, \\ 0, \; t_{lj} \notin W_k^s. \end{cases} \tag{15}$$

Введемо деяку множину P квантитативних ознак, за допомогою яких можна порівнювати характеристики текстових документів. Також введемо оператор відображення лексемного словника W на множину квантитативних ознак у масиві документів

$$U_{wd} : w_i \to p_{ij}^{wd},\ i = 1,2,...,N_w,\ j = 1,2,\ldots,N_d. \tag{16}$$

У загальному випадку величина p_{ij}^{wd} може мати довільне походження квантитативної характеристики. У подальшому будемо розглядати цю величину як текстову частоту лексеми w_i у текстовому документі d_j, яка визначається такою функціональною залежністю

$$p_{ij}^{wd} = \frac{n_{ij}^{wd}}{N_j^t}. \tag{17}$$

Аналогічно введемо оператор відображення семантичного складу S_j^d текстового документа d_j, на множину квантитативних ознак:

$$U_{sd} : s_k \to p_{kj}^{sd},\ k = 1,2,\ldots,N_s,\ j = 1,2,\ldots,N_d. \tag{18}$$

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

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \tag{19}$$

де

$$f_s(w_i, s_k) = \begin{cases} 1,\ w_i \in W_k^s, \\ 0,\ w_i \notin W_k^s. \end{cases} \tag{20}$$

Сукупність значень p_{kj}^{sd} утворює матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = (p_{kj}^{sd})_{k=1,j=1}^{N_s,N_d}. \tag{21}$$

Вектор

$$V_j = (p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}) \tag{22}$$

відображає документ d_j в N_s-мірному семантичному просторі текстових документів.

МОДЕЛЬ ВТОРИННИХ СЕМАНТИЧНИХ ПОЛІВ У ОРТОНОРМОВАНОМУ БАЗИСІ

Розглянемо представлення текстових документів у новому семантичному ортонормованому базисі, в якому коефіцієнти коваріації між різними семантичними частотними складовими текстових документів будуть дорівнювати

нулю. Тобто задача полягає в реалізації перетворення до нового базису, який буде описуватись діагональною коваріаційною матрицею. Такий базис може бути утворений за допомогою перетворення Карунена-Лоева, яке лежить в основі методу головних компонент [8, 9]. Розглянемо це перетворення для просторового базису утвореного частотними характеристиками семантичних полів. Коваріаційну матрицю розглянемо у вигляді

$$\mathrm{Cov}_s = [\mathrm{cov}_{ij}^s],\ \mathrm{cov}_{ij}^s = \mathrm{cov}(p_i^s, p_j^s) = E[(p_{il}^{sd} - E(p_{il}^{sd})(p_{jl}^{sd} - E(p_{jl}^{sd})]. \quad (23)$$

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

$$\mathrm{cov}_{ij}^s = \frac{1}{N_d - 1}\sum_{l=1}^{N_d}(p_{il}^{sd} - \overline{p}_i^{sd})(p_{jl}^{sd} - \overline{p}_j^{sd}),\ \ \overline{p}_i^{sd} = E(p_{il}^{sd}),\ \ \overline{p}_j^{sd} = E(p_{jl}^{sd}). \quad (24)$$

Знайдемо множину вторинних семантичних полів:

$$S' = \{s'_k \mid k = 1,2,\ldots,|S'|\}, \quad (25)$$

які описують текстові документи d_j за допомогою нових частотних векторів:

$$V_j'^s = (p_{1j}'^{sd}, p_{2j}'^{sd}, \ldots, p_{|S'|j}'^{sd}). \quad (26)$$

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

$$\mathrm{cov}(p_i'^s, p_j'^s) = 0,\ \ i \neq j. \quad (27)$$

Знайти семантичні вектори, для яких виконується умова (27) можна за допомогою методу головних компонент [8, 9]. Розглянемо основні положення цього методу для випадку семантичного простору текстових документів. Нехай відомо матрицю базисних частотних семантичних векторів A_s, яка описує зв'язок між векторами первинних та вторинних семантичних полів. Вважаємо цю матрицю ортогональною, для якої виконується умова

$$A_s^{-1} = A_s^T. \quad (28)$$

Тоді вектори первинних та вторинних семантичних полів зв'язані такими співвідношеннями:

$$V_j = A_s V_j',\ \ V_j' = A_s^T V_j. \quad (29)$$

Складові векторів V_j' називають головними компонентами. Для матриць M_{sd} (21) можна записати аналогічні співвідношення:

$$M_{sd} = A_s M'_{sd},\quad M'_{sd} = A_s^T M_{sd}. \quad (30)$$

Здійснимо центрування семантичних векторів та матриць:

$$\dot{V}_j = V_j - E[V_j], \quad (31)$$

$$\dot{M}_{sd} = M_{sd} - E[M_{sd}].$$

Розглянемо таку коваріаційну матрицю:

$$Cov'_s = \dot{M}'_{sd}(\dot{M}'_{sd})^T. \tag{32}$$

Враховуючи (30) отримаємо:

$$\mathrm{Cov}'_s = A_s^T \dot{M}_{sd}(\dot{M}_{sd})^T A_s = A_s^T \mathrm{Cov}'_s A_s, \tag{33}$$

де $$\mathrm{Cov}'_s = \dot{M}_{sd}(\dot{M}_{sd})^T.$$

Нехай матриця A_s складається із власних векторів матриці Cov'_s, тоді Cov'_s буде діагональною матрицею із власними значеннями матриці Cov'_s.

$$\mathrm{Cov}'_s = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_{|S'|}), \tag{34}$$

де $\lambda_1, \lambda_2, \ldots, \lambda_{|S'|}$ — власні значення матриці Cov'_s в порядку спадання їх величин. Задача знаходження матриці A_s, яка описує зв'язок між векторами первинних та вторинних семантичних полів зводиться до знаходження власних векторів та значень коваріаційної матриці Cov'_s первинних семантичних полів. Визначивши матрицю A_s, частотні семантичні вектори V_j можна розкласти по частотних векторах V'_j вторинних семантичних полів. Характерною властивістю базисних векторів вторинного семантичного простору є їх ортонормованість. Якщо множину вторинних семантичних полів впорядкувати за величиною власних чисел базисних векторів, тоді можна відкинути крайні в цьому ряді вторинні поля як несуттєві для аналізу. В результаті отримаємо

$$\hat{p}_{ij}^{sd} = \sum_{l=1}^{|\widetilde{S}'|} a_{il}\, \hat{p}'^{sd}_{lj}, \tag{35}$$

де $$|\widetilde{S}'| < |S'|.$$

Тобто для подальшого аналізу беруть підпростір простору вторинних семантичних полів. Складові семантичних векторів $\widetilde{p}'^{sd}_{lj}$ є проекціями складових p'^{sd}_{lj} на цей підпростір. Якщо базисні ортонормовані вектори розмістити у порядку спадання власних значень коваріаційної матриці, то оцінити похибку такої апроксимації можна за формулою

$$\varepsilon = \sum_{i=|\widetilde{s}'|}^{|S'|} \lambda_i. \tag{36}$$

Тобто, похибка визначається сумою власних значень базисних векторів, які не вносять вклад у апроксимацію. Звідси випливає, що для зменшення похибки при апроксимації необхідно взяти базисні вектори, для яких власні значення є максимальними. Виникає питання, яка розмірність простору вторинних полів є достатня для векторного представлення текстових документів. Одним із простих методів відбору головних компонент є правило

Кайзера, згідно з яким залишають ті компоненти, для яких виконується умова

$$\lambda_i > \frac{1}{|S'|}\operatorname{tr}(\mathrm{Cov}'_s). \tag{37}$$

Умова (37), визначає ті головні компоненти, для яких власне значення коваріаційної матриці є більшим за середнє всіх власних значень.

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}. Сингулярний розклад матриці терми-документи лежить в основі латентно-семантичного аналізу текстів [10, 11]. Нехай існує матриця типу «частоти_семантичних_полів–документи» M_{sd}, яка описується формулою (21). Вектор V_j (22) відображає документ d_j в N_s-мірному просторі текстових документів. Добуток двох векторів $(V_p)^T V_q$ визначає кількісну міру близькості цих векторів у N_s-мірному семантичному просторі текстових документів. Відповідно добуток матриць $(M_{sd})^T M_{sd}$ містить скалярні добутки векторів $(V_p)^T V_q$ всіх документів і відображає їхні кореляції в просторі семантичних векторів. Нехай існує сингулярна декомпозиція матриці M_{sd}:

$$M_{sd} = U_{sd}\Sigma_{sd}{Y_{sd}}^T. \tag{38}$$

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

$$(M_{sd})^T M_{sd} = (U_{sd}\Sigma_{sd}{Y_{sd}}^T)^T(U_{sd}\Sigma_{sd}{Y_{sd}}^T) = Y_{sd}{\Sigma_{sd}}^T\Sigma_{sd}{Y_{sd}}^T. \tag{39}$$

У відповідності 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} й Y_{sd} то отримаємо K-рангову апроксимацію матриці M_{sd}:

$$(M_{sd})_K = (U_{sd})_K(\Sigma_{sd})_K{(Y_{sd})_K}^T. \tag{40}$$

Матриця $(Y_{sd})_K$ відображає зв'язок між векторами документів $\widetilde{V}_j$ у новому комбінованому K-мірному семантичному просторі з ортонормованим семантичним базисом. Зв'язок між вектором V_j документу в первинному семантичному просторі та вектором $\widetilde{V}_j$ у просторі вторинних семантичних полів можна описати так:

$$\begin{aligned} V_j &= (U_{sd})_K(\Sigma_{sd})_K\widetilde{V}_j, \\ \widetilde{V}_j &= (\Sigma_{sd})_K^{-1}(U_{sd})_K^T V_j. \end{aligned} \tag{41}$$

Отже, ранг апроксимації матриці M_{sd}, який визначається числом K, також визначає розмірність простору вторинних семантичних полів. Очевидно, що число K, може бути суттєво меншим за розмірність N_s початкового семантичного простору. Це зменшує розмірність задачі аналізу подібності текстових документів у семантичному векторному просторі. Для чисельної оцінки сингулярних чисел взято текстову вибірку 155 художніх творів англомовної класики чотирьох відомих авторів (Ч. Діккенс, Д. Лондон, В. Скотт, М. Твен). Для утворення семантичного простору сформовано 15 семантичних полів, в які входять близько 5000 неозначених форм дієслова. Для кожного документа було сформрвано частотні словники, на основі яких розраховано частотні спектри семантичних полів документів. Отже, кожний документ розглядається як вектор у 15-мірному початковому семантичному просторі. Далі проведено сингулярний розклад матриці семантичних ознак. На рисунку наведено графічне зображення перших сингулярних чисел семантичних ознак типу «частоти_семантичних_полів–документи» у порядку спадання.

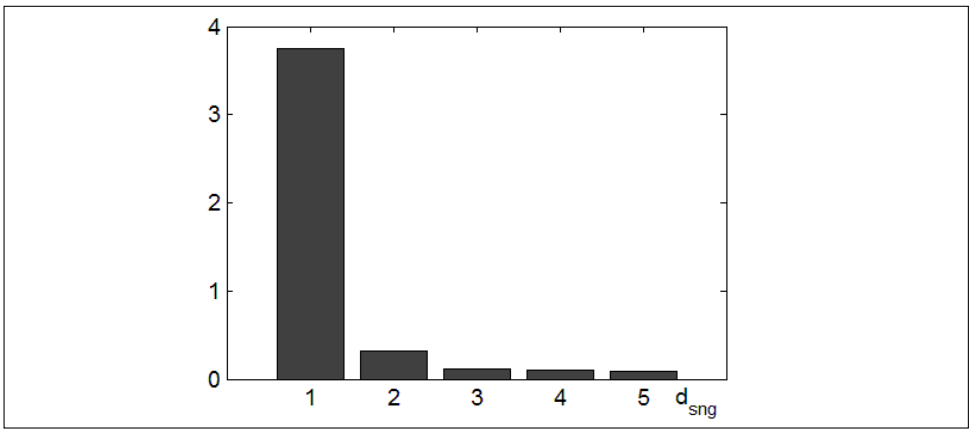

Рисунок. Сингулярні числа матриці семантичних ознак в порядку спадання

Слід відмітити суттєве спадання значень сингулярних чисел, що дає можливість для апроксимації матриці семантичних ознак взяти суттєво менше значення рангу апроксимації K у порівнянні із початковою розмірністю семантичного простору.

ВИСНОВКИ

Розглянуто модель некорельованих вторинних семантичних полів, що формуються на основі методу головних компонент шляхом визначення ортонормованого базису семантичного простору утвореного власними векторами коваріаційної матриці частотних семантичних векторів. Розмірність простору вторинних семантичних полів є суттєво меншою за розмірність простору первинних семантичних полів внаслідок заміни взаємопов'язаних складових некорельованими семантичними характеристиками. Ортонормований базис вторинних семантичних полів може бути також утворений за допомогою сингулярного розкладу матриць «частоти_семантичних_полів–документи».

Аналіз тестової вибірки текстових документів показав різке спадання значень сингулярних чисел. Це дає можливість брати до розгляду лише ті складові вторинних семантичних полів, які описуються першими сингулярними числами. Використання низькорозмірного ортонормованого базису вторинних семантичних полів може бути ефективним у задачах класифікації та кластеризації текстових даних.

ЛІТЕРАТУРА

1. *Pantel P., Peter D. Turney.* From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research. — 2010. — **37**. — P. 141–188.
2. *Павлишенко Б.М.* Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів // Електроніка та інформаційні технології. — 2011. — Випуск 1. — С. 212–222.
3. *Павлишенко Б.М.* Модель семантичного контексту в алгоритмах інтелектуального аналізу текстів // Комп'ютинг. — 2011. — Том 10, випуск 3. — С. 216–222.
4. *Павлишенко Б.М.* Використання концепції семантичного поля у векторній моделі текстових документів // Східно-Європейський журнал передових технологій. — 2011. — № 6/2 (54). — С. 7–11.
5. *Павлишенко Б.М.* Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. — 2012. — № 1. — С. 69–76.
6. *Левицкий В.В., Стернин И.А.* Экспериментальные методы в семасиологии. — Воронеж: Изд-во ВГУ, 1989. — 192с.
7. *Вердиева З.Н.* Семантические поля в соверменном английском языке. — М.: Высшая школа, 1986. — 120 с.
8. *Брасегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И.* Анализ данных и процессов: учеб. пособие. — СПб.: БХВ-Петербург, 2009. — 512 с.
9. *Jolliffe I.T.* Principal Component Analysis. — Series: Springer Series in Statistics, 2nd ed. — Springer, NY, 2002, XXIX — 487 p.
10. *Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman*. Deerwester Scott Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science. — 1990. — **41**, Issue 6. — P. 391–407.
11. *Mirzal Andri.* Clustering and Latent Semantic Indexing Aspects of the Singular Value Decomposition. — http://arxiv.org/abs/1011.4104v2.

Надійшла 01.06.2012