УДК 519.765:519.767:004.89

Б.М. Павлишенко

Львівський національний університет імені Івана Франка, Львів, Україна

МОДЕЛЮВАННЯ НЕЧІТКИХ СЕМАНТИЧНИХ ПОЛІВ У МАСИВАХ ТЕКСТОВИХ ДОКУМЕНТІВ

На основі теорії нечітких множин створено моделі нечіткого семантичного поля та статистичної лінгвістичної змінної для словникового складу текстових масивів. Формування синтаксичних та семантичних правил для лінгвістичних змінних дає можливість визначити ієрархічну структуру нечітких семантичних полів. Запропоновані в роботі характеристики семантичних полів та лексем дають можливість відобразити нечіткість семантичної структури словника в алгоритмах інтелектуального аналізу текстів.

Ключові слова: *нечіткі множини, семантичні поля, інтелектуальний аналіз текстів.*

Вступ

В інтелектуальному аналізі текстів широко використовують векторну модель представлення текстових документів. Одним із методів аналізу є кластеризація текстових об'єктів у векторному просторі частот лексемного словника [1, 2]. При цьому виникає проблема, пов'язана із великою розмірністю векторного простору, яка зумовлена розміром лексемного словника аналізованого масиву текстів. Одним із можливих шляхів вирішення цієї проблеми є використання концепції семантичних полів, які розглядають як множини лексем, об'єднаних спільним поняттям [3, 4]. Враховуючи денотативні закономірності лексикографічного складу словника, можна зауважити, що одні і ті ж лексеми можуть знаходитись у різних семантичних полях. Отже, не існує чіткого розмежування семантичних полів. Таку особливість доцільно враховувати в алгоритмах аналізу текстів. Перспективним є створення моделі семантичних полів на основі теорії нечітких множин [5, 6]. Отже, моделювання та аналіз нечіткої структури семантичних полів є актуальною задачею при створенні алгоритмів аналізу даних із використанням векторного представлення текстових об'єктів.

Аналіз останніх досліджень. В роботі [7] проаналізовано формування семантичного векторного простору, утвореного частотними характеристиками семантичних полів словникового складу текстових масивів для кластерного аналізу. Розглянемо модель, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Нехай існує деякий словник лексем, які зустрічаються у текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \left\{ w_i \mid i = 1, 2..., N_w \right\}. \quad (1)$$

Сукупність текстових документів опишемо такою множиною

$$D = \left\{ d_j \mid j = 1, 2..., N_d \right\}. \quad (2)$$

Введемо множину семантичних полів

$$S = \left\{ s_k \mid k = 1, 2..., N_s \right\}. \quad (3)$$

Під семантичним полем розуміють таку множину лексем, які об'єднані деяким спільним поняттям [3, 4]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та інші. Документ d_j з множини текстових документів D можна представити як впорядковану множину слів, порядок елементів якої відповідає порядку слів у цьому документі

$$T_j^d = \left\{ t_{lj} \mid l = 1, 2...N_j^t \right\}. \quad (4)$$

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

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

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

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

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

$$S_j^d = \left\{ n_{kj}^{sd}(s_k) \mid k = 1, 2..., N_s \right\}, \quad (10)$$

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

$$n_{kj}^{sd} = \sum_{l=1}^{N_j^t} f_s(t_{lj}, s_k), \quad (11)$$

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

Введемо матрицю семантичних ознак типу "частоти_семантичних_полів – документи"

$$M_{sd} = \left(p_{kj}^{sd} \right)_{k=1, j=1}^{N_s, N_d}, \quad (12)$$

де p_{kj}^{sd} – частота семантичного поля s_k в лексемному складі документа d_j, яку обчислимо за формулою

$$p_{kj}^{sd} = \frac{n_{kj}^{sd}}{N_j^t}. \quad (13)$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right) \quad (14)$$

відображає документ d_j в N_s-мірному семантичному просторі текстових документів. Запропонована модель дає можливість визначити матрицю частотних семантичних ознак типу "частоти_семантичних_полів–документи" і ввести новий базис для текстових характеристик. В семантичному базисі можуть спостерігатись якісно нові групування текстових документів. В роботі [7] запропонований алгоритм ієрархічної кластеризації текстових документів у векторному просторі, утвореному частотними характеристиками семантичних полів. На прикладі текстових масивів англомовної прози показана ефективність використання низькорозмірних матриць семантичних характеристик документів в алгоритмах кластерного аналізу текстових масивів.

Постановка задачі. Використовуючи теоретико-множинний підхід, створимо модель нечітких семантичних полів. Введемо клас семантичних полів як клас нечітких множин α-рівня. Використаємо поняття лінгвістичної змінної для відображення структурних нечітких зв'язків між елементами семантичних полів.

Семантичні поля як нечіткі множини

Розглянемо словник W як універсальну множину із теорії нечітких множин [5, 6]. При моделюванні семантичних полів на основі звичайних множин розглянемо характеристичну функцію множини семантичного поля S_k:

$$\mu_k^c(w_i) = \begin{cases} 0, w_i \notin S_k, \\ 1, w_i \in S_k. \end{cases} \quad (15)$$

Функція $\mu_k^c(w_i)$ набуває лише двох значень – одиниці у випадку приналежності лексеми w_i семантичному полю S_k і нуля у випадку, якщо w_i не належить цій множині. Якщо допустити, що функція $\mu_k^c(w_i)$ може набувати проміжних значень в інтервалі [0,1], тоді множина, яка буде описуватись такою функцією, буде називатись нечіткою множиною [5, 6]. Нечітким семантичним полем $\tilde{S}_k$ назвемо пару $(W, \mu_k^s(w_i))$, де W – універсальна множина словника лексем, $\mu_k^s(w_i)$ – функція, визначена на множині W, яка приймає значення на відрізку [0,1]. Таку функцію $\mu_k^s(w_i)$ назвемо функцією приналежності лексеми w_i нечіткому семантичному полю $\tilde{S}_k$. Загальну форму запису нечіткого семантичного поля будемо розглядати у вигляді

$$\tilde{S}_k = \sum_{i=1}^{N_w} \frac{\mu_k^s(w_i)}{w_i}, \ (w_i \in W). \quad (16)$$

Точками переходу нечіткого семантичного поля назвемо лексеми w_i, для яких $\mu_k^s(w_i) = 0.5$. Нечітке семантичне поле можна зобразити за допомогою діаграми Заде [5, 6], яка є графіком функції $\mu_k^s(w_i)$. Також можна зобразити у вигляді сингельтона – пари $(w_i, \mu_k^s(w_i))$, де на першому місці знаходиться назва лексеми, а на другому – величина її приналежності семантичному полю $\tilde{S}_k$. Сингельтон називають чітким, якщо $\mu_k^s(w_i) = 1$.

Очевидно, що розглядаючи деяке семантичне поле, можна знайти ненульові значення приналежності цьому полю для більшості лексем словника. Тому доцільно визначити деякі додаткові критерії формування нечіткого семантичного поля. Такі критерії визначимо використовуючи поняття нечітких множин α-рівня [5, 6]. Нечітким семантичним полем α-рівня $\tilde{S}_k^\alpha$ назвемо множину лексем $(w_i \in W)$, для яких виконується умова $\mu_k^s(w_i) > \alpha$. Отже,

$$\tilde{S}_k^\alpha = \left\{ w_i \mid \mu_k^s(w_i) > \alpha \right\}. \quad (17)$$

Можна показати, що нечітке семантичне поле $\tilde{S}_k$ можна розкласти по полях всіх α-рівнів у вигляді

$$\tilde{S}_k = \sum_\alpha \alpha \cdot \tilde{S}_k^\alpha. \quad (18)$$

Складним питаннями у теорії нечітких множин є побудова функції приналежності $\mu_k^s(w_i)$. Основними методами побудови таких функцій є методи експертних оцінок [5, 6]. Можна виділити прямі та опосередковані методи. В прямому методі експерти напряму задають значення функції приналежності для кожної лексеми семантичного поля, наприклад, на основі лексикографічного аналізу. В опосередкованому методі здійснюють попарні порівняння елементів нечіткої множини. Такі попарні порівняння можуть здійснюватись на основі порівняння текстових частот лексемних відношень, які характерні для заданих лексемних шаблонів аналізованого семантичного поля. В результаті отримаємо квадратну матрицю попарних порівнянь лексем семантичного поля із одиничними діагональними елементами $A = \left\{ a_{ij} \right\}$. Можна припустити, що

$$a_{ij} = \frac{\mu_k^s(w_i)}{\mu_k^s(w_j)}. \quad (19)$$

На основі утвореної матриці А розглянемо рівняння

$$A\vec{\mu}_k^s = \lambda \vec{\mu}_k^s. \quad (20)$$

Із розв'язку рівняння (20) виберемо вектор $\vec{\mu}_k^s$, який відповідає найбільшому власному значенню матриці А. Елементи вектора $\vec{\mu}_k^s$ можна розглядати

як наближення значень приналежності лексем нечіткого семантичного поля $\tilde{S}_k$.

Звичайне семантичне поле можна розглядати як частковий випадок нечіткого семантичного поля. Зв'язок між звичайним та нечітким семантичним полем можна визначити через наближену апроксимацію функції приналежності характеристичною функцією

$$\mu_k^c(w_i) = \begin{cases} 1, & \mu_k^s(w_i) \geq 0.5, \\ 0, & \mu_k^s(w_i) < 0.5. \end{cases} \quad (21)$$

Звичайним семантичним полем, найближчим до нечіткого семантичного поля, назвемо поле із характеристичною функцією $\mu_k^c(w_i)$, яка визначається виразом (21). Вимір нечіткості семантичного поля можна визначити, як відстань від його множини до множини найближчого до нього звичайного семантичного поля у заданій метриці. В евклідовій метриці індекс нечіткості семантичного поля можна обрахувати так

$$\mathrm{Ind}_f^e(\tilde{S}_k) = \frac{2}{\sqrt{N_w}} \sqrt{\sum_{i=1}^{N_w} (\mu_k^s(w_i) - \mu_k^c(w_i))^2}\ . \quad (22)$$

В семантичному полі можна виділити структурні семантичні групи, зокрема, синонімічні ряди, які можна розглядати як підмножини семантичного поля. Підмножиною нечіткої множини семантичного поля $\left\{ W, \mu_k^s(w_i) \right\}$ будемо називати нечітку множину $\left\{ W, \mu_a(w_i) \right\}$, для якої виконується нерівність

$$\mu_a(w_i) < \mu_k^s(w_i). \quad (23)$$

Аналогічно до визначення нечіткого семантичного поля можна визначити поняття семантично нечіткої лексеми $\tilde{w}_i$. Такою лексемою будемо називати пару

$$\tilde{w}_i = (S, \mu_k^w(s_k)), \quad (24)$$

де S – універсальна множина семантичних полів; $\mu_k^w(s_k)$ – функція приналежності, визначена на множині S, яка приймає значення на відрізку [0, 1]. За допомогою такої функції приналежності можна характеризувати спектр семантичних полів лексеми. Цю функцію можна визначати як за допомогою експертного аналізу, так і на основі статистичних характеристик текстового розподілу даної лексеми у заданих шаблонах, які відповідають певним семантичним полям. Розглянемо опис нечітких лексем $\tilde{w}_i$ за допомогою лінгвістичної змінної [5, 6] із врахуванням особливостей семантичних полів. Лінгвістичною змінною лексеми $\tilde{w}_i$ назвемо набір

$$L_s(\tilde{w}_i) = \left\{ w_i, T_w(w_i), S, G_{synt}, M_{sem} \right\}, \quad (25)$$

де w_i – назва змінної; $T_w(w_i)$ – терм-множина імен значень змінної w_i; S – універсальна множина семантичних полів; G_{synt} – синтаксичне правило утворення імен значень змінної w_i; M_{sem} – семантичне правило, яке ставить у відповідність нечітку підмножину універсальної множини семантичних полів S кожному елементу терм-множини $T_w(w_i)$.

Синтаксичне правило розглядають як алгоритмічну процедуру породження елементів множини $T_w(w_i)$, а семантичне правило – як процедуру обрахунку функцій приналежності нечітких підмножин універсальної множини S. Введення поняття лінгвістичної змінної в опис семантично нечітких лексем є ефективним тоді, коли синтаксична процедура полягає в утворенні терм-множини на основі конкатенації назви лінгвістичної змінної з назвами елементів семантичних полів інших частин мови. Наприклад, лінгвістична змінна опису деякого елемента деякого семантичного поля іменників може мати терм-множину, утворену внаслідок конкатенації назви змінної із елементами деякого семантичного поля прикметників, які будуть виступати в ролі модифікаторів змінної, і в такий спосіб розширювати множину значень лінгвістичної змінної. Причому, терм-множина лінгвістичної змінної деякого іменника може бути об'єднанням терм-множин сформованих на основі модифікаторів із різних семантичних полів прикметників. Зв'язок лінгвістичних змінних елементів одних семантичних полів із елементами інших семантичних полів через синтаксичну процедуру формування терм-множин відображає внутрішню ієрархічну структуру нечітких семантичних полів лексемного словника текстових масивів. Окремим питанням є семантична процедура побудови функцій приналежності нечітких підмножин універсальної множини лексемного словника S, імена яких є елементами терм-множини лінгвістичної змінної. Одним із можливих шляхів розв'язку цього завдання є аналіз статистичних характеристик розподілу лексемних сполучень терм-множини в текстових шаблонах, які відповідають аналізованому семантичному полю. На основі знайдених текстових частот лексемних сполучень можна побудувати табличну апроксимацію функції приналежності для кожного лексемного сполучення, яке належить терм-множині аналізованої лінгвістичної змінної. Аналізуючи статистичні розподіли лексемних сполучень, можна об'єднати синтаксичне та семантичне правила формування лінгвістичної змінної. Нехай, наприклад, існує деякий елемент w_j множини семантичного поля іменників S_1 і деяка множина прикметників семантичного поля S_2. Розглянемо бінарне відношення

$$I_b(w_j) \subseteq S_2 \times \left\{ w_j \right\}. \quad (26)$$

Будемо розглядати пари $(w_i, w_j) \in S_2 \times \left\{ w_j \right\}$, які утворюють лексемні сполучення $w_i w_j$ в тексті. На основі статистичного аналізу текстових масивів можна виявити текстові частоти $p(w_i w_j)$ для цих сполучень. Виберемо деяке мінімальне значення текстових частот, за яким будемо формувати бінарне відношення $I_b(w_j)$. Будемо вважати, що

$$I_b(w_j) = \left\{ \begin{array}{l} (w_i, w_j) \mid \\ w_i \in S_2, \\ (w_i, w_j) \in S_2 \times \{ w_j \}, \\ p(w_i w_j) > p_{min} \end{array} \right\}. \qquad (27)$$

Правило (27) формування бінарного відношення $I_b(w_j)$ будемо розглядати як статистичне відображення синтаксичного правила формування лінгвістичної змінної, а набір частот $p(w_i w_j)$ із умовою $p(w_i w_j) > p_{min}$ – як деякі коефіцієнти для функції приналежності. Табличну табуляцію функції приналежності з врахуванням цих коефіцієнтів можна здійснювати на основі експертного аналізу, враховуючи як денотативні лексикографічні значення лексем, так і їх конотативні значення в тексті. На основі правила (27) формування бінарного відношення $I_b(w_j)$ можна побудувати терм-множину для лексеми w_j:

$$T_w(w_j) = \{ w_i w_j \mid (w_i, w_j) \in I_b(w_j) \}. \qquad (28)$$

Запис $w_i w_j$ означає словосполучення лексем внаслідок конкатенації w_i та w_j. В загальному випадку словосполучення $w_i w_j$ слід розглядати не просто як з'єднання двох лексем, а як входження цих лексем в один із наперед визначених шаблонів. В результаті ми отримаємо визначення статистичної лінгвістичної змінної для нечіткої лексеми $\tilde{w}_i$ як

$$L(\tilde{w}_i) = \{ w_i, T_w(w_i), St(w_i) \}, \qquad (28)$$

де w_i – назва статистичної лігвістичної змінної; $T_w(w_i)$ – терм-множина; $St(w_i)$ – статистична процедура утворення терм-множини.

Висновки

На основі теорії нечітких множин створено модель нечіткого семантичного поля лексемного складу текстових масивів. Визначено характеристики для нечіткого семантичного поля - функцію приналежності, найближче звичайне семантичне поле, міру нечіткості семаничного поля, семантичне поле α-рівня. Поряд із поняттям нечіткого семантичного поля введено поняття семантично нечіткої лексеми, для якої визначено лінгвістичну змінну. Формування синтаксичних та семантичних правил для лінгвістичних змінних нечітких лексем дає можливість визначити ієрархічну структуру нечітких семантичних полів. Визначено статистичну лінгвістичну змінну семантично нечіткої лексеми. Для такої змінної терм-множина формується на основі статистичних характеристик розподілу лексемних сполучень. Запропоновані в роботі характеристики семантичних полів та лексем дають можливість відобразити нечіткість семантичної структури словника в алгоритмах інтелектуального аналізу текстових масивів.

Список літератури

1. *Брасегян А.А. Анализ данных и процессов: учеб. пособие / А.А. Брасегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. – СПб.: БХВ–Петербург, 2009. – 512с.: ил.*
2. *Pantel P. From Frequency to Meaning: Vector Space Models of Semantics / Patrick Pantel, Peter D. Turney // Journal of Artificial Intelligence Research. – 2010. – Vol.37. – P. 141-188.*
3. *Вердиева З.Н. Семантические поля в современном английском языке / З.Н. Вердиева. – М.: Высшая школа, 1986. – 120 с.*
4. *Левицкий В.В. Экспериментальные методы в семасиологии / В.В. Левицкий, И.А. Стернин. – Воронеж: Изд-во ВГУ, 1989. – 192 с.*
5. *Конышева Л.К. Основы теории нечетких множеств: Учебн. пособие / Л.К. Конышева, Д.М. Назаров. – СПб.: Питер, 2011. – 192 с.: ил.*
6. *Заде Л. Понятие лингвистической переменной и его применение к принятию приблизительных решений / Л. Заде. – М.: Мир, 1976. – 165 с.*
7. *Павлишенко Б.М. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів / Б.М. Павлишенко // Електроніка та інформаційні технології. – 2011. – Вип. 1. – С. 212-222.*

Надійшла до редколегії 7.11.2011

Рецензент: д-р фіз.-мат. наук, проф. О.М. Бордун, Львівський національний університет імені Івана Франка, Львів.

МОДЕЛИРОВАНИЕ НЕЧЕТКИХ СЕМАНТИЧЕСКИХ ПОЛЕЙ В МАССИВАХ ТЕКСТОВЫХ ДОКУМЕНТОВ

Б.М. Павлышенко

На основе теории нечетких множеств созданы модели нечеткого семантического поля и статистической лингвистической переменной для словарного состава текстовых массивов. Формирование синтаксических и семантических правил для лингвистических переменных дает возможность определить иерархическую структуру нечетких семантических полей. Предложенные в работе характеристики семантических полей и лексем дают возможность отобразить нечеткость семантической структуры словаря в алгоритмах интеллектуального анализа текстов.

Ключевые слова: *нечеткие множества, семантические поля, интеллектуальный анализ текстов.*

THE MODELLING OF FUZZY SEMANTIC FIELDS IN TEXT DOCUMENTS ARRAYS

B.M. Pavlyshenko

Based on the theory of fuzzy sets the model of fuzzy semantic fields and statistical linguistic variable for the vocabulary of text arrays have been created. The forming of syntactic and semantic rules for the linguistic variables makes it possible to determine the hierarchical structure of fuzzy semantic fields. The suggested characteristics of semantic fields and lexemes make it possible to display the fuzziness of the semantic structure of the dictionary in text mining algorithms.

Keywords: *fuzzy sets, semantic fields, text mining.*