computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

МОДЕЛЬ СЕМАНТИЧНОГО КОНТЕКСТУ ЛЕКСЕМ В АЛГОРИТМАХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТІВ

Богдан Павлишенко

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

Резюме: *Запропонована модель семантичного контексту лексем, яка відображає структурну семантичну організацію лексемного складу текстових масивів. Показано, що в семантичному контексті лексемного словника формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – лексемами.*

Ключові слова: *аналіз формальних понять, семантичний контекст, семантичні поля, інтелектуальний аналіз текстів.*

MODEL OF SEMANTIC CONTEXT OF LEXEMES IN THE TEXT MINING ALGORITHMS

Bohdan M. Pavlyshenko

Ivan Franko Lviv National University,
Drahomanov Str. 50, Lviv, 79005 Ukraine, e-mail:pavlsh@yahoo.com

Abstract: *The model of semantic context of lexemes which represent the structure semantic configuration of lexems corpus of text arrays has been proposed. It is shown that partially ordered set of semantic concepts are formed in the lexem semantic context. Concepts' intents are defined by semantic fields, concepts extents – by lexems.*

Key words: *Formal Concepts Analysis, Semantic Context, Semantic Fields, Text Mining.*

ВСТУП

Теорія формального аналізу понять (Formal Concept Analysis) є однією із складових сучасних методів інтелектуального аналізу даних [1, 2, 3, 4]. В цій теорії аналізують формальні поняття та їх ієрархії використовуючи математичний апарат теорії алгебраїчних решіток. В алгоритмах аналізу текстів часто використовують матриці ознак. У матриці ознак колонки визначають документи, а рядки – частоти лексем у цих документах. У поширеній векторній моделі документи відображають як вектори у багатомірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми із словників текстових масивів [5, 6]. Кожна колонка матриці ознак є вектором частот лексем для певного документа. Мірою відстані між двома документами може бути кут між векторами цих документів в утвореному векторному просторі. Такий підхід має також ряд проблем, зокрема розмірність аналізованого простору є великою, оскільки зумовлена розміром словника. Одним із можливих шляхів вирішення цієї проблеми є використання кількісних характеристик різних лексемних об'єднань, зокрема, семантичних полів. Лексеми можуть належати тим чи іншим семантичним полям за деякою спільною семантичною ознакою [7, 8]. Враховуючи багатозначність лексем, одні і ті ж лексеми можуть одночасно належати різним семантичним поля, наслідком чого є утворення деякої семантичної структури. Формальний аналіз понять дає можливість описати

семантичну структуру лексемних об'єднань. Створення моделі семантичного контексту лексем, яка б відображала семантичну структуру лексемного словника дасть можливість ввести нові параметри в алгоритми аналізу текстових документів.

1. ПОСТАНОВКА ЗАДАЧІ

Розглянемо теоретико-множинну модель сукупності текстових документів, словника, та семантичних полів. Побудуємо модель семантичної структури словника текстових масивів використовуючи теорію аналізу формальних понять. Визначимо поняття семантичного контексту та семантичного концепту. Проаналізуємо використання моделі семантичного контексту у векторній моделі текстових документів.

2. ТЕОРЕТИКО-МНОЖИННА МОДЕЛЬ СЕМАНТИЧНИХ ПОЛІВ

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

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

Сукупність текстових документів опишемо такою множиною

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

Введемо множину семантичних полів

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

Під семантичним полем розуміють таку множину лексем, які об'єднані деяким спільним поняттям [7, 8]. 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

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

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

$$U_{ws} : w_i \to s_k, \ i = 1,2..., N_w; k = 1,2..., N_s. \quad (5)$$

Оператор U_{ws} задамо таблицею, яка визначається експертним лексикографічним аналізом [7, 8]. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\to} s_k, i = 1,2..., N_w \right\}. \quad (6)$$

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

$$S_j^d = \left\{ n_{kj}^{sd}(s_k) \mid k = 1,2..., N_s \right\}. \quad (7)$$

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

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

$$f_s(t_{lj}, s_k) = \begin{cases} 1, t_{lj} \in W_k^s \\ 0, t_{lj} \notin W_k^s \end{cases}. \quad (8)$$

Введемо матрицю семантичних ознак типу "частоти_семантичних_полів–документи"

$$M_{sd} = \left(p_{kj}^{sd} \right)_{k=1,j=1}^{N_s, N_d} \quad (9)$$

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

$$p_{kj}^{sd} = \frac{n_{kj}^{sd}}{N_j^t}. \quad (10)$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right) \quad (11)$$

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".

3. МОДЕЛЬ СЕМАНТИЧНОГО КОНТЕКСТУ ЛЕКСЕМНОГО СЛОВНИКА

Розглянемо модель семантичної структури словника текстових масивів використовуючи теорію аналізу формальни понять [1-4]. Визначимо семантичний контекст як трійку

$$\mathrm{K}_s = (W, S, I), \tag{12}$$

де W – лексемний склад словника, S – множина семантичних полів, I – відношення належності лексеми до семантичного поля

$$I \subseteq W \times S, \quad I = \{\langle w_i, s_k \rangle\}. \tag{13}$$

Пара $\langle w_i, s_k \rangle$ означає, що лексема w_i належить до семантичного поля s_k. Семантичний контекст можна представити у вигляді бінарної матриці, рядки якої означають лексеми, а стовпці – семантичні поля.

Уведемо решітку семантичних концептів лексем. Для деяких $Extent \subseteq W$, $Intent \subseteq S$ визначимо такі відображення

$$\begin{aligned} Extent' &= \{\, s \in S \mid w \in Extent : wIs \,\} \\ Intent' &= \{\, w \in W \mid s \in Intent : wIs \,\} \end{aligned} \tag{14}$$

Множина $Extent'$ описує семантичні поля, яким належать лексеми множини $Extent$, а множина $Intent'$ описує лексеми, які належать семантичним полям множини $Intent$. Уведемо семантичний концепт як пару

$$Concept = (Extent, Intent), \tag{15}$$

до якої належать лексеми з множини $Extent \subseteq W$ та семантичні поля з множини $Intent \subseteq S$ з такими умовами

$$\begin{aligned} Extent' &= Intent \\ Intent' &= Extent \end{aligned} \tag{16}$$

Множину $Extent$ назвемо об'ємом, а $Intent$ – змістом семантичного концепту лексем $Concept$. В семантичному контексті лексем утворюється частково-впорядкована множина семантичних концептів

$$\begin{aligned} &\Psi(W, S, I) = \{\, Concept_m \mid m = 1,2,...N_{ct} \}, \\ &Concept_m = (Extent_m, Intent_m), \end{aligned} \tag{17}$$

де N_{ct} – кількість виявлених семантичних концептів у формальному семантичному контексті лексемного словника. Семантичний концепт

$$Concept_1 = (Extent_1, Intent_1) \tag{18}$$

є менш загальним чим концепт

$$Concept_2 = (Extent_2, Content_2) \tag{19}$$

тобто виконується умова

$$(Extent_1, Intent_1) \le (Extent_2, Intent_2), \tag{20}$$

якщо

$$Extent_1 \subseteq Extent_2. \tag{21}$$

В цьому випадку концепт $Concept_2$ можна вважати узагальненням концепту $Concept_1$. Семантичний концепт можна розглядати як підматрицю семантичного контексту, яка повністю заповнена одиницями.

Решітку концептів часто відображають за допомогою діаграм Гассе. Розглянемо приклад, в якому лексеми $[L1, L2, L3, L4]$ належать семантичним полям $[S1, S2, S3, S4]$. Контекст, в якому кожна лексема належить лише одному семантичному полю можна розглянути у вигляді таблиці 1:

Таблиця 1

	S1	S2	S3	S4
L1	X			
L2		X		
L3			X	
L4				X

Приклад контексту, в якому деякі лексеми належать різним семантичним полям зообразимо у вигляді таблиці 2.

Таблиця 2

	S1	S2	S3	S4
L1	X		X	X
L2		X	X	
L3	X		X	
L4				X

Діаграма Гассе у випадку однозначної приналежності лексем семантичним полям зображена на рис.1, а у випадку багатозначної

приналежності – на рис.2.

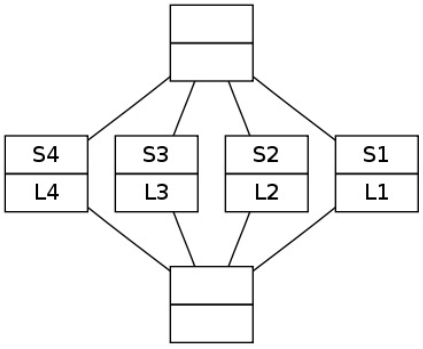

Рис. 1 – Діаграма Гассе у випадку однозначної приналежності лексем семантичним полям

В таких діаграмах кожний елемент представляє семантичний концепт. На верхньому рівні діаграми концепт включає в себе всі лексеми і нульову множину семантичних полів. На другому рівні в елементи діаграми входить одне семантичне поле, на третьому – два семантичних поля і так до найнижчого рівня, який включає в себе всі семантичні поля та нульову множину лексем. В комірку діаграми записують назви семантичних полів, які присутні в цьому концепті і відсутні в більш загальних концептах з верхніх рівнів. Також в комірку записують назви лексем, які присутні в цьому концепті та відсутні в менш загальних концептах з нижніх рівнів. Якщо, наприклад, в менш загальних концептах є ті ж лексеми, що і в більш загальному, тоді відповідна область комірки більш загального концепта є незаповнена. Такі діаграми відображають внутрішню структурну організацію текстових словників на основі теорії формального аналізу понять.

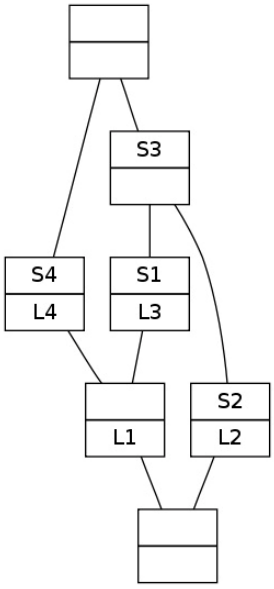

Рис. 2 – Діаграма Гассе у випадку багатозначної приналежності лексем семантичним полям

Наприклад, в контексті, який описується таблицею 1 та рис.1 можна виявити множину таких концептів

$$\left\{\begin{array}{l}([L1,L2,L3,L4],[\varnothing]),\\([L1],[S1]),([L2],[S2]),\\([L3],[S3]),([L4],[S4]),\\([\varnothing],[S1,S2,S3,S4])\end{array}\right\}. \quad (22)$$

В контексті, який описується таблицею 2 та рис.2 можна виявити таку множину концептів

$$\left\{\begin{array}{l}([L1,L2,L3,L4],[\varnothing]),\\([L1,L2,L3],[S3]),([L1,L4],[S4]),\\([L1,L3],[S1,S3]),([L2],[S2]),\\([L1],[S1,S3,S4]),\\([\varnothing],[S1,S2,S3,S4])\end{array}\right\}. \quad (23)$$

Уведемо додаткові квантитативні ознаки текстових об'єктів на основі характеристик об'ємів семантичних концептів виявлених в семантичному контексті лексемного словника текстових масивів. Лексемний склад концепту $Concept_m = (Extent_m, Intent_m)$ визначимо так

$$W_m^{ct} = \{ w_i \mid w_i \in Extent_m,\ i = 1,2...,N_w \},\quad m = 1,2,....N_{ct} \quad (24)$$

Як кількісну характеристику *m-го* концепта в лексемному складі документа d_j розглянемо частоту вживання лексем словника *m-го* концепта в документі d_j

$$p_{mj}^{cd} = \frac{\sum_{l=1}^{N_j^t} f_{ct}(t_{lj}, W_m^{ct})}{N_j^t}, \quad (25)$$

$$f_{ct}(t_{lj}, s_k) = \begin{cases} 1, t_{lj} \in W_m^{ct} \\ 0, t_{lj} \notin W_m^{ct} \end{cases}.$$

Частотні характеристики p_{mj}^{cd} утворюють додатковий семантичний підпростір для квантитативного аналізу текстів в рамках векторної моделі текстових документів. Повний вектор документа у семантичному просторі можна розглядати у вигляді

$$V_j^{sc} = \left(p_{1j}^{sd}, p_{2j}^{sd},..., p_{N_s j}^{sd}, p_{1j}^{cd}, p_{2j}^{cd},..., p_{N_{ct} j}^{cd}\right). \quad (26)$$

3. ВИСНОВКИ

Запропонована модель семантичного контексту відображає структурну семантичну організацію лексемного складу текстових масивів. В семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – лексемами. Введення моделі семантичного контексту дає можливість виявити нові підмножини лексемних семантичних груп для формування векторного простору аналізу текстових масивів. Такі групи утворюють множину формальних об'ємів семантичних концептів лексем. Отримана в роботі модель може бути використана для оптимізації алгоритмів інтелектуального аналізу текстів за допомогою введення додаткових квантитативних характеристик на основі семантичних концептів лексем.

6. СПИСОК ЛІТЕРАТУРИ

[1] B. Ganter, R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer, 1999.

[2] S.O. Kuznetsov, S. A. Obiedkov, Comparing performance of algorithms for generating concept lattices, *Journal of Experimental and Theoretical Artificial Intelligence*. 14 (2002). – pp. 189-216.

[3] P. Cimiano, A. Hotho, S. Staab, Learning concept hierarchies from text corpora, using formal concept analysis, *Journal of Artificial Intelligence Research*. 24 (2005) – pp. 305-339.

[4] Abderrahim El Qadi, Driss Aboutajedine, Yassine Ennouary, Formal concept analysis for information retrieval, *International Journal of Computer Science and Information Security (IJCSIS)*. 7 (2) (2010).

[5] А.А. Брасегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров, *Анализ данных и процессов: учеб. Пособие,* СПб.: БХВ–Петербург, 2009. – 512 с.

[6] Patrick Pantel, Peter D. Turney, From frequency to meaning: vector space models of semantics, *Journal of Artificial Intelligence Research*. 37 (2010). – pp. 141-188.

[7] З.Н. Вердиева, *Семантические поля в соверменном английском языке*, М.: Высшая школа, 1986. – 120 с.

[8] В.В. Левицкий, И.А. Стернин, *Экспериментальные методы в семасиологии*, Воронеж: Изд-во ВГУ, 1989. – 192 с.

Павлишенко Богдан *в 1992 р. закінчив фізичний факультет Львівського національного університету ім. Івана Франка, в 1995 р. захистив кандидатську дисертацію, з 2005 р. доцент факультету електроніки ЛНУ ім. Івана Франка.*

Коло наукових інтересів: фізика напівпровідників, розпізнавання оптичних образів, інтелектуальний аналіз даних, штучний інтелект, квантові алгоритми.