УДК 681.3.062

Марченко О. О., д.ф.- м.н., доц.

Метод формального концептуального аналізу на основі векторів семантико-синтаксичної валентності слів

Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова 4д, e-mail: rozenkrans@yandex.ua

O.O. Marchenko, Doctor of Sciences (Physics & Mathematics), Associate Professor

A method of formal concept analysis based on semantic-syntactic valence vectors of words

Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4d,
e-mail: rozenkrans@yandex.ua

В роботі запропоновано застосування векторів семантико-синтаксичної валентності слів як контекстних векторів в методах формального концептуального аналізу для автоматичної побудови таксономій високої якості. Дослідження та експерименти підтвердили значне зростання якості побудови таксономій із збільшенням вимірності тензорної моделі при генерації векторів семантико-синтаксичної валентності слів. Підвищення арності тензорів дає моделі можливість більш точного опису багатомірних семантико-синтаксичних зв'язків та дозволяє виділяти більше комутативних семантико-синтаксичних властивостей слів, що використовуються формальним концептуальним аналізом для побудови таксономій більш високої якості.

Ключові слова: штучний інтелект, комп'ютерна лінгвістика, онтології, бази знань.

The article is devoted to research and development of methods for automatic construction of taxonomies that is very actual and much demanded area in computational linguistics and artificial intelligence at all, given that the taxonomy is a fundamental hierarchical basis for the construction of ontological networks in knowledge bases.

The paper provides the use of semantic-syntactic valence vectors of words as context vectors in formal concept analysis methods for the automatic construction of taxonomies of high quality. Research and experiments have confirmed a significant increase in the quality of taxonomies construction using semantic-syntactic valence vectors of words in the method of formal concept analysis. And along with the increase of dimensions number of the tensor model for generating semantic-syntactic valence vectors of words comes a considerable increase in the quality of the taxonomies construction. Increasing arity of tensor model provides a more precise description of the multidimensional semantic and syntactic relations and allows to extract more commutative semantic and syntactic properties of words used by the formal concept analysis for building taxonomies of higher quality.

Keywords: artificial intelligence, computational linguistics, ontologies, knowledge bases.

Статтю представив д.ф.-м.н., проф. Анісімов А.В.

Розробка методів автоматизації побудови таксономій є актуальним та надзвичайно затребуваним напрямом у комп'ютерній лінгвістиці та взагалі у штучному інтелекті з огляду на те, що таксономії є фундаментальною ієрархічною основою для конструювання онтологічних мереж баз знань.

Формально задача полягає у побудові ієрархічного графу з вхідної множини іменників N із застосуванням відношення гіпонімії-гіперонімії (клас-підклас).

Серед розроблених методів автоматичної побудови таксономій обробкою текстових корпусів можна виділити два основних класи: методи, основані на кластеризації слів з мірами семантичної близькості і теоретико-множинні методи впорядкування слів-понять. Обидва класи методів працюють з моделлю векторного простору, в якому слова або терми представлені у

вигляді відповідних їм векторів ознак, отриманих при обробці та аналізі деякого текстового корпусу.

Для першого класу методів характерне використання деякої міри семантичної близькості для визначення відстані між векторами слів, щоб визначити, наскільки вони семантично подібні, і чи мають вони бути зараховані в один кластер. Наприклад, може бути використана міра косинусів кута між векторами слів. Представники цього класу методів у свою чергу поділяються на агломеративні (кластеризація знизу-вверх) і розділяючі методи (кластеризація зверху-вниз). Кращі представники методів даного напряму описуються в роботах [1, 2, 3].

Теоретико-множинні методи здійснюють побудову таксономії встановленням часткового порядку на множині слів-понять по відношенню включення між їх множинами ознак. Одним з кращих представників цього напряму є *формальний концептуальний аналіз* (Formal Concept Analysis, FCA) [4].

Формальний концептуальний аналіз

Метод формального концептуального аналізу працює із структурою даних, що має назву контекст K. Він представляє собою дані про лінгвістичний контекст використання слів певної предметної області поряд з деяким базовим набором лексем у текстових корпусах, що були оброблені та проаналізовані для створення таблиці K. Наприклад, це може бути матриця, що містить дані про сполучуваність вхідного набору термінів-іменників поряд з базовим набором якісних прикметників. В загальному вигляді можна сказати, що метод працює з таблицею *Об'єкт* × *Атрибут*. В роботі [5] наводиться контекст методу формального концептуального аналізу, представлений таблицею, яка зберігає дані про сполучуваність термінів-іменників поряд із певним набором дієслів у позиції додатку (відношення *дієслово-присудок – іменник-додаток*). Терміни-іменники представлені у контексті K векторами, ненульові елементи яких позначають наявність сполучуваності даного іменника у якості додатку дієсловам, що відповідають позиціям ненульових елементів. У якості значень зазвичай виступає частота сумісної появи у текстовому корпусі відповідної пари *дієслово-присудок – іменник-додаток*. Існують також варіанти із застосуванням бінарних матриць, що передбачають лише значення 1 та 0.

Слід відзначити, що при формуванні контексту K у якості базового набору контекстних слів, зокрема набору дієслів, як описано у прикладі вище, обирають ті слова, з якими існує стабільний зв'язок хоча б у одного із термінів-іменників з N. Тобто у базовий контекстний набір потраплятимуть лише такі дієслова, для яких існує хоча б одна така пара *дієслово-присудок – іменник-додаток,* для якої відносна частота використання у текстовому корпусі має перевищувати певний пороговий рівень, тобто $\forall v \in V \ \exists n \in N: \upsilon(v,n) \geq Th$, де V – множина дієслів, N – множина термінів-іменників, $\upsilon(v,n)$ – частота використання словосполучення v n у корпусі текстів, Th – пороговий рівень.

Формальний концептуальний аналіз використовує теорію порядку для аналізу зв'язків між об'єктами G та їх ознаками M. Формальний концептуальний аналіз ідентифікує з формального контексту K множину ознак $B \subseteq M$, що має бієктивне відношення з множиною об'єктів $A \subseteq G$. Така бієктивно зв'язана пара має назву формальний концепт (A, B). Формальні концепти є частково впорядкованими $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow (A_1 \subseteq A_2)$, або, що є еквівалентним, $B_2 \subseteq B_1$. Об'єкти $o_1, o_2 \in G$ – концептуально кластеризовані, якщо $\{o_1, o_2\} \subseteq A$, де (A, B) – формальний концепт в K.

З бінарного представлення об'єктів формальний концептуальний аналіз будує решітку формальних концептів. Для того, щоб інтерпретувати решітку як таксономію, застосовуються наступні два правила:

1. Вводиться один концепт таксономії C_B з міткою B для кожного формального концепту (A, B), якщо $|A| \geq 2$. Концепти впорядковуються згідно порядку у решітці.

2. Вводиться один таксономічний концепт C_O для кожного об'єкту $o \in G$ і помічається міткою *o*. Концепти впорядковуються таким чином, що $C_O \leq C_B$, де (A, B) – формальний концепт та $o \in A$, і не існує формального концепту (A', B'), щоб $(A', B') \leq (A, B)$ та $o \in A'$.

Серед переваг методу формального концептуального аналізу слід зазначити якість побудованих таксономій, що є на порядок вищою, ніж в ієрархій, побудованих методами кластеризації [5]. Також відзначимо прозорість та добру інтерпретованість таксономій, згенерованих формальним концептуальним аналізом.

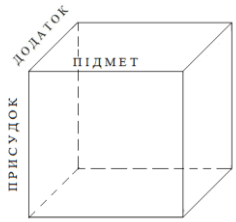
Серед недоліків методу формального концептуального аналізу слід зазначити його експоненційну часову складність $O(2^n)$, де n – кількість термінів-іменників, які треба впорядкувати у таксономію. Але на практиці метод формального концептуального аналізу демонструє час роботи, порівняний з лінійним, що відбувається через надмірну розрідженість векторного представлення термінів-іменників.

Використання векторів семантико-синтаксичної валентності у методах формального концептуального аналізу

Пропонується у якості контекстних векторів термінів-іменників використовувати вектори їх семантико-синтаксичної валентності [6].

Вектори семантико-синтаксичної валентності слів генеруються невід'ємною факторизацією тензорів частотних оцінок сумісної сполучності слів в різних синтаксичних позиціях у великих текстових корпусах.

Тривимірний тензор для зберігання частотних даних сполучностей слів – підметів, присудків та прямих додатків, отриманих в результаті аналізу великих текстових корпусів, зображений на малюнку 1.



Малюнок 1. Тривимірний тензор для зберігання частотних даних сполучностей слів – підметів, присудків та прямих додатків.

В результаті частотного аналізу текстових корпусів формується розріджений тензор великої розмірності. З метою отримання більш стислого та зручного представлення до тензору застосовується невід'ємна тензорна факторизація [6].

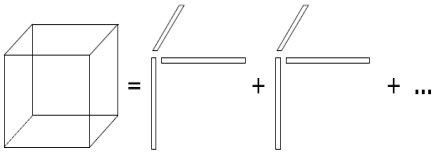
Основна ідея методу полягає в мінімізації суми квадратів різниць між оригіналом тензору та його факторизованою моделлю. Для тривимірного випадку тензору $T \in R^{D1 \times D2 \times D3}$ це відповідає рівнянню

$$\min_{x_i \in R^{D1}, y_i \in R^{D2}, z_i \in R^{D3}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i \|_F^2$$, де к – число вимірів у факторизованій моделі.

При невід'ємній тензорній факторизації додається обмеження невід'ємності, перетворюючи модель у

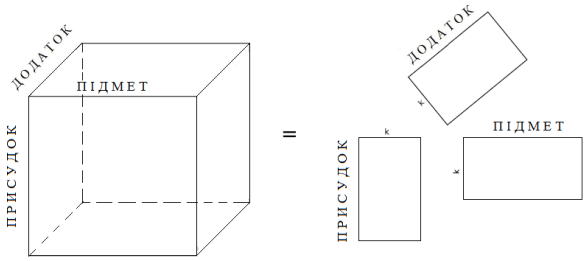
$$\min_{x_i \in R_{\geq 0}^{D1}, y_i \in R_{\geq 0}^{D1}, z_i \in R_{\geq 0}^{D1}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i \|_F^2$$

Модель можна представити графічно наступним чином, враховуючи, що дана декомпозиція складається з суми зовнішніх добутків трійок векторів (по числу вимірів тензору).



Малюнок 2. Графічне представлення невід'ємної тензорної факторизації як суми зовнішніх добутків.

Декомпозиція невід'ємної тензорної факторизації для тензору частотних оцінок сполучностей слів «підмет» × «присудок» × «додаток» у вигляді трьох матриць представлена на малюнку 3.



Малюнок 3. Графічне представлення невід'ємної тензорної факторизації для лінгвістичного тензору.

Використовуючи модель невід'ємної тензорної факторизації для аналізу тривимірних сполучностей (s,v,o) («підмет»×«присудок»×«додаток»), з текстових корпусів виділяють узагальнені моделі селективних преференцій, а також деякі елементи моделей фреймової семантики (семантичних відмінків).

В результаті факторизації кожен підмет-іменник, присудок-дієслово та додаток-іменник отримує власний вектор розмірності k з відповідних матриць.

Оригінальне значення з тензору T для трійки (s,v,o) x_{svo} може бути відновлене у факторизованій моделі обчисленням суми

$$x_{svo}=\sum_{i=1}^{k} s_{si} v_{vi} o_{oi} .$$

Для того, щоб обчислити оцінку частоти для сполучення (*мисливець застрелив вовка*) потрібно знайти вектор-підмет для слова *мисливець*, потім вектор-присудок для значення *застрелив*, і, нарешті, вектор-додаток для слова *вовк*. Потім згідно попередньої формули обчислюється оцінка частоти для даного сполучення слів. Якщо оцінка більша за деякий пороговий рівень, то можна зробити висновок про можливість існування даного речення у мові.

Вектори з матриць факторизованого тензору описують комутативні властивості слів. Вони прописують, які зв'язки утворюють дані слова – з якими лексемами та у яких синтаксичних позиціях. У даних векторів присутня як синтаксична, так і семантична складова. Тому вектори були названі векторами семантико-синтаксичної валентності слів.

Для генерації векторів семантико-синтаксичної валентності іменників з вхідної множини термінів N потрібно було випробувати декілька тензорних моделей.

Пропонується використати двовимірну модель, базовану на матриці частот сумісної сполучуваності слів у парах «присудок-додаток» в текстових корпусах, а також – тривимірну модель для представлення сполучностей типу «підмет-присудок-прямий_додаток» і чотирьохвимірну модель «підмет-присудок-прямий_додаток-непрямий_додаток».

Експерименти

Методика. В якості базового алгоритму формального концептуального аналізу був обраний метод, описаний у [7]. Для застосування у даному методі формального концептуального аналізу контекстних векторів було запропоновано використати вектори семантико-синтаксичної валентності іменників в синтаксичній позиції додатку. Для цього було зібрано та факторизовано наступні тензори:

- двовимірна модель – матриця «присудок-прямий_додаток»;
- тривимірна модель – тензор «підмет-присудок-прямий_додаток»;
- чотиривимірна модель – тензор «підмет-присудок-прямий_додаток-непрямий_додаток».

Для збірки тензорів в якості текстових корпусів були використані тексти статей English Wikipedia. Для кожного слова з вхідної множини іменників N до навчальної вибірки були включені всі статті English Wikipedia, які містили дане слово у назві. Потім до вибірки були включені ті статті English Wikipedia, що пов'язані із вже включеними сторінками зв'язками типу *категорія* та зв'язками-посиланнями.

Далі було виконано фільтрацію навчальної вибірки. Для цього було застосовано кластеризацію зібраних текстів методом латентного семантичного аналізу [8]. Після виконання кластеризації найбільший кластер текстів обирається у якості навчальної вибірки текстів.

Після цього слідує етап збірки матриці та тензорів. Він полягає у виконанні синтаксичного аналізу текстів вибірки із застосуванням Стенфордського парсеру та у заповненні значень елементів масивів. Наприклад, якщо при аналізі синтаксичної структури речення з'ясувалося, що підметом є слово «мисливець», присудком – «вполював», додатком – «кабана», то відбувається збільшення значення елементу T[i,j,k], де i – індекс слова «мисливець», j – індекс слова «вполював», k – індекс слова «кабана»: T[i,j,k]= T[i,j,k]+1. В результаті обробки всієї вибірки елемент T[i,j,k] буде містити частоту використання у ній словосполучення *мисливець вполював кабана*.

У випадку відсутності певного елементу трійки чи четвірки відсутній елемент заповнюється порожнім словом ∅. Наприклад, для чотирьохмірної моделі «підмет-присудок-прямий_додаток-непрямий_додаток» у реченні *мисливець вполював кабана* відсутня позиція непрямого додатку, тому воно замінюється реченням *мисливець_вполював_кабана_*∅. Якщо не вистачає більше одного елементу структури, то такі дані не вносяться у тензор.

Після збірки тензору відбувається обнуління всіх елементів масиву, що не перевищують певного порогового значення, для того, щоб залишити у масиві лише стійкі дані та позбутися помилок.

Після етапу збірки матриці та тензорів виконується їх невід'ємна факторизація. Для невід'ємної факторизації матриці двовимірної моделі був використаний метод Лі-Сунга [9]. Для невід'ємної факторизації тривимірного та чотиривимірного тензорів був використаний метод Parafac [10].

Після того для усіх іменників з вхідної множини N були зібрані вектори з матриць додатків. Ці вектори семантико-сентаксичної валентності іменників були використані методом

формального концептуального аналізу [7] у якості контекстних векторів для побудови таксономій з іменників множини N.

Для оцінки якості алгоритмів побудови таксономій зазвичай використовуються методи порівняння згенерованих алгоритмами таксономій з деякою еталонною, в якості якої виступають таксономії, зібрані вручну.

Для гарантування наявності еталонної таксономії, зібраної вручну, вхідна множина іменників N формувалася наступним чином. У якості еталонної таксономії береться певна підмережа з лексико-семантичної бази WordNet, починаючи з деякого вузла N_0 у якості кореня, та всі його нащадки на відстані 4-5 разом із відношеннями гіпонімії-гіперонімії між ними. Для формування вхідної множини N зі всіх вузлів даної підмережі були взяті перші слова їх синсетів.

Для обчислення оцінок структурної близькості таксономій, отриманих алгоритмом, до еталонної була застосована методика, описана в [5].

Результати експериментів. Було проведено 3 серії експериментів – для множини N, що відповідає N_0 = «транспорт», для множини N, що відповідає N_0 = «їжа» та для множини N, що відповідає N_0= «професії».

Метод формального концептуального аналізу використовував у якості контекстних векторів:

1) вектори з простої матриці інцидентності слів у парах «присудок-додаток» (як це описано в роботі [5]) у якості базового оціночного рівня;
2) вектори семантико-синтаксичної валентності іменників-додатків з факторизованої матриці двовимірної моделі;
3) вектори семантико-синтаксичної валентності іменників-додатків з факторизованого тензору тривимірної моделі;
4) вектори семантико-синтаксичної валентності іменників-додатків з факторизованого тензору чотиривимірної моделі.

В результаті кожної серії було побудовано чотири таксономії T_1,T_2,T_3,T_4, які були досліджені на структурну подібність до еталонної таксономії з WordNet. Оцінки представлені у таблиці 1.

Таблиця 1. Оцінки структурної подібності побудованих таксономій до еталонних

	T_1	T_2	T_3	T_4
N_0=«транспорт»	67.39 %	70.21 %	76.47 %	78.73%
N_0= «їжа»	65.98 %	68.29 %	73.41 %	75.12%
N_0= «професії»	68.01 %	70.77 %	77.81 %	79.37%

Отримані оцінки свідчать про беззаперечну перевагу застосування векторів семантико-синтаксичної валентності слів у методах формального концептуального аналізу по відношенню до використання простих контекстних векторів з матриць інцидентності слів. При цьому є помітне зростання точності таксономій, побудованих на основі векторів семантико-синтаксичної валентності слів, із збільшенням вимірності тензорної моделі. Це відбувається через те, що в тензорних багатовимірних моделях враховуються тонкі складні багатовимірні зв'язки предикатно-аргументного типу між дієсловами та їх аргументами-іменниками, що є недоступним для двовимірних моделей. Навіть невід'ємна факторизація двовимірної матриці інцидентності дає помітне покращення якості таксономії у порівнянні з використанням алгоритмом простого контексту, так як невід'ємна факторизація надійно виділяє чіткі та стійкі зв'язки з матриці інцидентності.

Подальше збільшення вимірності тензорної моделі дозволяє накопичувати більший об'єм багатомірних складних семантико-синтаксичних даних та виділяти більше семантико-синтаксичних зв'язків, які використовуються як атрибути у методі формального концептуального аналізу, що підвищує якість побудованих таксономій.

Висновки

В роботі запропоновано застосування векторів семантико-синтаксичної валентності слів як контекстних векторів в методах формального концептуального аналізу для автоматичної побудови таксономій високої якості. Дослідження та експерименти підтвердили значне зростання якості побудови таксономій із збільшенням вимірності тензорної моделі при генерації векторів семантико-синтаксичної валентності слів. Підвищення арності тензорів дає моделі можливість більш точного опису багатомірних семантико-синтаксичних зв'язків та дозволяє виділяти більше комутативних семантико-синтаксичних властивостей слів, що використовуються формальним концептуальним аналізом для побудови таксономій більш високої якості.

Список використаних джерел

1. Caraballo S.A. Automatic construction of a hypernym-labeled noun hierarchy from text / S.A. Caraballo // Proceedings of the 37th Annual Meeting of the ACL. – 1999. – pp.120-126.
2. Hindle D. Noun classification from predicate-argument structures / D. Hindle // Proceedings of the Annual Meeting of the ACL-90. – 1990. – P. 268-275.
3. Pereira F. Distributional clustering of English words / F. Pereira, N. Tishby, L. Lee // Proceedings of the 31st Annual Meeting of the ACL, Columbus, Ohio, USA. – 1993. – P. 183-190.
4. Ganter B. Formal Concept Analysis: Mathematical Foundations/ B. Ganter, R. Wille. – New York: Springer-Verlag. – Secaucus. Inc. – 1999. – pp. 284.
5. Cimiano P. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text/ P. Cimiano, A. Hotho, S. Staab // In Proceedings of the European Conference on Artificial Intelligence (ECAI). – 2004. – P. 435-439.
6. Анисимов А.В. Определение семантических валентностей концептов онтологий с помощью неотрицательной факторизации тензоров больших текстовых корпусов / А.В. Анисимов, А.А. Марченко, Т.Г Вознюк // Кибернетика и системный анализ .– 2014. – № 3. – С. 3-16.
7. Vychodil V. A new algorithm for computing formal concepts / V. Vychodil // In Proceedings of the 19th European Meeting on Cybernetics and Systems Research, Vienna. – 2008. – P. 15-21.
8. Deerwester S. Indexing by latent semantic analysis / S. Deerwester, S. T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman// Journal of the Association for Information Science. – 1990. – № 41(6). – P. 391-407.
9. Lee D.D. Algorithms for non-negative matrix factorization / D.D. Lee, S.H. Seung // In Proceedings of the NIPS-2000. – 2000. – P. 556-562.
10. Cichocki A. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation / A. Cichocki, R. Zdunek, A.-H. Phan, S.-I. Amari// Chichester: J. Wiley & Sons. – 2009. – p. 477.

References

1. CARBALLO S.A. (1999) *Automatic construction of a hypernym-labeled noun hierarchy from text*, Proceedings of the 37th Annual Meeting of the ACL-99, pp.120-126.
2. HINDLE D. (1990) *Noun classification from predicate-argument structures*, Proceedings of the Annual Meeting of the ACL-1990, pp. 268-275.
3. PEREIRA F., TISHBY N., Lee L. (1993) *Distributional clustering of English words*, Proceedings of the 31-st Annual Meeting of the ACL, Columbus, Ohio, USA, pp. 183-190.
4. GANTER B., WILLE R. (1999) *Formal Concept Analysis: Mathematical Foundations*, New York: Springer-Verlag, Secaucus. Inc., P. 284.
5. CIMIANO P., HOTHO A., STAAB S. (2004) *Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text*, Proceedings of the European Conference on Artificial Intelligence (ECAI-2004), pp. 435-439.
6. ANISIMOV A. V., MARCHENKO A. A., VOZNIUK T. G. (2014), *Determination of semantic valences of ontologies concepts using non-negative tensor factorization of large text corpora*, Cybernetics and System Analysis 3, pp. 3-16.
7. VYCHODIL V. (2008) *A new algorithm for computing formal concepts*, Proceedings of the 19-th European Meeting on Cybernetics and Systems Research, Vienna, pp. 15-21.
8. DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K., HARSHMAN R. (1990) *Indexing by latent semantic analysis*, Journal of the Association for Information Science 41(6), pp. 391-407.
9. LEE D. D., SEUNG S. H. (2000) *Algorithms for non-negative matrix factorization*, Proceedings of the NIPS-2000, pp. 556-562.
10. CICHOCKI A., ZDUNEK R., PHAN A.-H., AMARI S.-I. (2009) *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Chichester: J. Wiley & Sons, P. 477.

Надійшла до редколегії 01.09.15