УДК 681.3

Марченко О. О., д.ф.- м.н., доц.

Методи невід'ємної тензорної та матричної факторизації в задачах комп'ютерної лінгвістики

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

Non-negative tensor and matrix factorization methods for natural language processing

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

В роботі розглянуто застосування методів невід'ємної матричної та тензорної факторизації в задачах комп'ютерної лінгвістики, зокрема для латентного семантичного аналізу, кластеризації великих масивів текстів природною мовою, для автоматизації генерації таких лінгвістичних структур як селективні преференції та опису структур фреймової семантики дієслів. Також представлено модель опису композиційної семантики речень природною мовою, де за допомогою тензорної факторизації семантика кожного окремо взятого слова визначається та уточнюється у контексті інших слів речення та у контексті всього речення у цілому.

Ключові слова: комп'ютерна лінгвістика, семантичний аналіз текстів природною мовою, невід'ємна факторизація тензорів, невід'ємна матрична факторизація.

In the paper the use of non-negative matrix and tensor factorization methods in classical tasks of computational linguistics is considered. These tasks include latent semantic analysis and large natural language texts corpora clustering, automation methods for generation of such linguistic structures as selectional preferences and descriptions of the verbs semantics frame structures etc. The N-dimensional tensor is implemented as a multiway array of data obtained from the frequency analysis of large text corpora. Factorization of N-dimensional tensor with decomposition rank k generates N matrixes consisting of k columns that represent a mapping of each individual dimension of the tensor on k factor-dimensions of latent semantic space. The paper describes an algorithm for unsupervised generation of verb subcategorization frames and selectional preference information. The method improves previous non-negative tensor factorization approaches by predicting whether a syntactic argument is likely to be used with a verb.

The paper also describes the model of compositional semantics of natural language sentences where non-negative tensor factorization is employed. Semantics of every single word is defined and clarified in the context of other words of the sentence and in the context of the entire sentence.

Keywords: natural language processing, semantic analysis, non-negative tensor and matrix factorization.

Статтю представив д.ф.-м.н., проф., чл.-кор. НАН України Анісімов А.В.

Методи невід'ємної тензорної та матричної факторизації є ефективними методами аналізу, класифікації та кластеризації багатовимірних даних, що зберігаються у великих масивах.

Дані методи широко застосовуються у багатьох напрямах штучного інтелекту, а також у інформаційних технологіях загалом, коли є необхідність аналізу залежностей та взаємозв'язків у даних, що мають різну природу та модальність.

Серед основних напрямів застосувань методів невід'ємної тензорної і матричної факторизації великих багатовимірних масивів даних слід відзначити обробку та аналіз графічних зображень, обробку відеозображень, кластеризацію великих текстових корпусів,

© О. О. Марченко, 2015

виділення лінгвістичних даних з великих корпусів текстів природною мовою [1, 2].

Методи невід'ємної тензорної і матричної факторизації дозволяють автоматизувати заповнення контентом лінгвістичних баз знань онтологічного типу [3].

Застосування невід'ємної матричної факторизації в задачах семантичного аналізу природномовних текстів

Теми, концепції та семантика є ключовими ознаками тексту природною мовою, які дають уявлення про його зміст та смисл. Для отримання таких ознак з тексту були розроблені спеціальні методи, відомі під загальною назвою "виділення ознак" (feature extraction). Метою таких методів є виділення основних концепцій (або тем) текстів та представлення документів у вигляді їх комбінації. Ці методи успішно застосовуються у задачах обробки природномовних текстів, таких як кластеризація документів, знаходження семантичної відстані та багато інших.

Матричні розклади, такі як сингулярний розклад матриці (відомий під назвою «латентний семантичний аналіз») [4] або невід'ємний матричний розклад (невід'ємна матрична факторизація, NMF) [5], використовуються в якості основи даних методів.

Загалом, NMF є розкладом (або факторизацією) матриці V в добуток двох матриць W і H. Причому, на елементи матриць V, W та H накладаються додаткові обмеження невід'ємності. Потрібно зазначити, що такий розклад не є обов'язково унікальним і матриці W та H не повинні бути ортогональними.

Ця факторизація матриць стала популярною після публікації статті Лі та Сунга [5]. В цій роботі було запропоновано алгоритм побудови невід'ємного розкладу матриці.

В цій статті використовується оригінальний метод Лі та Суна, ще відомий під назвою «мультиплікативні правила оновлення» В якості цільової функції використовується норма Фробеніуса.

У області обробки природномовних текстів NMF зазвичай застосовується до TD-матриці (матриці терм×документ). Кожен рядок матриці TD відповідає певному документу з корпусу текстів, а кожен стовпець матриці відповідає певному слову. Якщо задано m документів і загальна кількість слів у них рівна n, то розмір TD матриці V рівний m на n. NMF використовується для розкладу матриці V у добуток двох матриць W і H з розмірами m на k і k на n відповідно, де зазвичай $k << \min(m,n)$. Наприклад, k може бути встановлено в значення очікуваної кількості кластерів. Часто параметр k називають «кількість ознак».

Інтуїтивно, параметр k може бути пояснений наступним чином: TD матриця V («документи» ×«слова») розкладається у добуток двох матриць – V=WH. Матриця W пов'язує «документи» і «ознаки», а матриця H пов'язує «ознаки» і «слова». Тому, як випливає з властивостей добутку матриць, кожен документ представляється у вигляді лінійної комбінації виділених ознак. Після цього, традиційні методи кластеризації можуть бути застосовані до рядків матриці W. Рядки матриці W можна розглядати як вектори семантичних або тематичних ознак m документів. Для визначення тематичної або семантичної близькості двох документів можна використовувати косинусну міру, що обчислюється через скалярний добуток векторів ознак документів.

В цілому повний процес кластеризації виглядає наступним чином:

1. Побудувати TD-матрицю V
2. Нормалізувати стовпці матриці V
3. Застосувати NMF: V = W H
4. Побудувати кластеризацію на основі матриць W і H

NMF, як засіб обробки природної мови, має кілька переваг над іншими методами виділення ознак. По-перше, матриці W і H мають тільки невід'ємні елементи, що робить процес їх інтерпретації у термінах обробки природної мови простішим і більш інтуїтивним. По-друге, стовпці W не повинні бути ортогональними, а отже виділені теми можуть мати спільні риси, що є досить звичайним для реальних документів.

Використання невід'ємної факторизації тензорів для моделювання та виділення лінгвістичних семантико-синтаксичних структур з великих текстових корпусів

У роботі [1] було запропоновано модель невід'ємної факторизації тензорів, у яких зберігаються дані про частоту використання різних словосполучень у великих текстових корпусах із врахуванням синтаксичної позиції слів. Факторизовані тензорні лінгвістичні моделі дозволяють автоматизовано виділяти з корпусів текстів опис таких лінгвістичних структур, як селективні преференції (selectional preferences)[1]

та субкатегоріальні фрейми дієслів [2], що містять дані про семантичні та синтаксичні властивості зв'язків між дієсловами та їх аргументами-іменниками у реченнях.

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

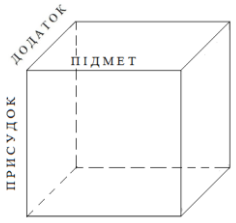

Рисунок 1. Тривимірний тензор для зберігання частотних даних сполучностей слів – підметів, присудків та прямих додатків.

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.

Основна ідея методу полягає в мінімізації суми квадратів різниць значень 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 тензорних добутків трійок векторів (по числу вимірів тензору).

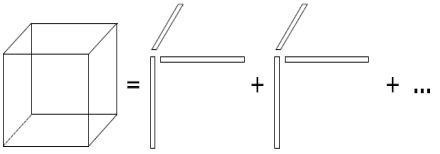

Рисунок 2. Графічне представлення невід'ємної тензорної факторизації як суми зовнішніх добутків.

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

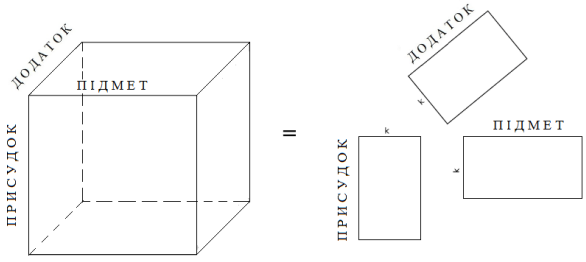

Рисунок 3. Графічне представлення невід'ємної тензорної факторизації для лінгвістичного тензору.

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 *кухар готує борщ* потрібно знайти вектор-підмет для слова *кухар*, потім вектор-присудок для значення *готує*, і, нарешті, вектор-додаток для слова *борщ*. 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 вони були названі векторами семантико-синтаксичної валентності слів.

Моделювання композиційної семантики речень природної мови шляхом застосування невід'ємної факторизації лінгвістичних тензорів

У роботі [6] було запропоновано застосувати декомпозицію Такера для факторизації

лінгвістичних тензорів з метою моделювання семантики слів в контексті семантики всього речення. Таким чином обчислення семантичних значень окремих слів йде від визначення семантики цілого речення.

Побудова латентних факторів іменників. Першим етапом методу є побудова латентної факторної моделі для іменників, базуючись на словах з їх контексту. Для цієї цілі використовується невід'ємна матрична факторизація (алгоритм Лі та Сунга). Невід'ємна матрична факторизація мінімізує цільову функцію (у даному випадку збіжності Кулбека-Ліблера моделі добутку двох матриць $\mathbf{W}_{I\times K}$ $\mathbf{H}_{K\times J}$ до початкової матриці $\mathbf{V}_{I\times J}$ при умові невід'ємності елементів усіх матриць. При цьому $K<<I, J$, таким чином має місце значна редукція над даними початкової матриці. Модель факторизації представлено на рисунку 4.

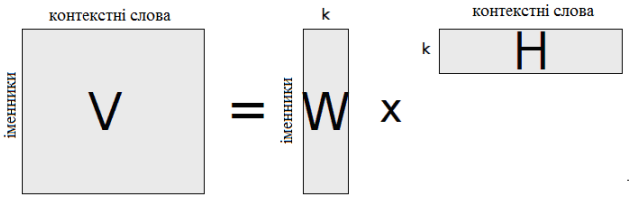

Рисунок 4. Модель невід'ємної факторизації матриці контексту

Для такої цільової функції Кулбека-Ліблера та для двох початкових матриць W_0 і H_0 NMF алгоритм складається з ітераційного виконання двох кроків:

1. $(H_k)_{i,j} = (H_{k-1})_{i,j} \times \dfrac{(W_{k-1}^T V)_{i,j}}{(W_{k-1}^T W_{k-1} H_{k-1})_{i,j}}$,

2. $(W_k)_{i,j} = (W_{k-1})_{i,j} \times \dfrac{(VH_{k-1}^T)_{i,j}}{(W_{k-1} H_{k-1} H_{k-1}^T)_{i,j}}$.

Метод гарантовано збігається до локального мінімуму функції Кулбека-Ліблера.

Моделювання багатовимірної взаємодії. На другому етапі будується багатовимірна модель взаємодії для трійок підмет-присудок-додаток, базуючись на латентних факторах, отриманих на першому етапі. Дана латентна модель взаємодії розроблена на основі декомпозиції Такера [7], хоча конкретна реалізація суттєво відрізняється в деталях.

Декомпозиція Такера – мультилінійне узагальнення добре відомої сингулярної декомпозиції матриць, що використовується у латентному семантичному аналізі. У декомпозиції Такера тензор розкладається на ядерний тензор, який множиться на матриці по кожному з вимірів. Для тривимірного тензору $\boldsymbol{\mathcal{X}} \in R^{I\times J\times L}$ модель визначається як

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\, \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r$$

При цьому $P, Q, R << I, J, L$. Ядерний тензор $\boldsymbol{\mathcal{G}}$ представляє стиснутий латентний варіант оригінального тензору $\boldsymbol{\mathcal{X}}$; матриці $\mathbf{A} \in R^{I\times P}$, $\mathbf{B} \in R^{J\times Q}$ та $\mathbf{C} \in R^{L\times R}$ представляють латентні фактори для кожного виміру, у той час як $\boldsymbol{\mathcal{G}} \in R^{P\times Q\times R}$ показує рівень взаємодії між різними латентними факторами. На рисунку 5 зображено графічне представлення декомпозиції Такера.

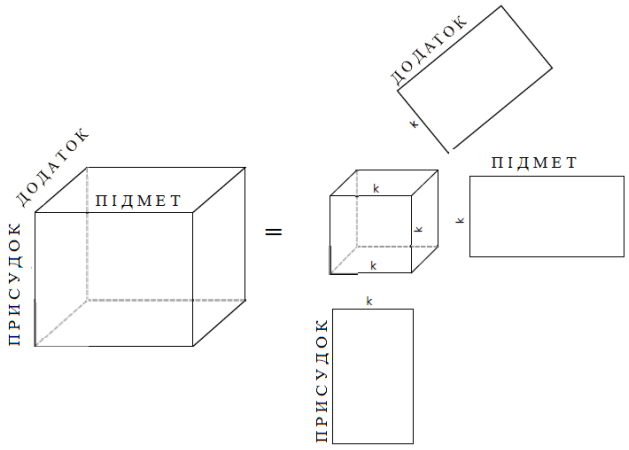

Рисунок 5. Графічне представлення декомпозиції Такера

Побудова моделі Такера з двовимірних факторів. Обчислення моделі декомпозиції Такера для тензору вимагає багато ресурсів у термінах часу та пам'яті. Крім того, декомпозиція не є унікальною. Враховуючи ці недоліки, було розроблено альтернативний метод побудови моделі Такера. Розглядаються фактор-матриці як такі (як вихід першого етапу методу) та обчислюється ядерний тензор $\boldsymbol{\mathcal{G}}$. Крім того, не використовується латентне представлення для першого виміру, що означає, що перший вимір представлений безпосередньо через оригінальні дані. Дана модель напряму використовується до лінгвістичних даних. Ядерний тензор $\boldsymbol{\mathcal{G}}$ моделює латентну взаємодію між дієсловами-присудками, підметами та додатками. Ядерний тензор $\boldsymbol{\mathcal{G}}$

обчислюється застосуванням n-мірного добутку відповідного виміру оригінального тензору:

$$\mathcal{G} = \mathcal{X} \times_2 \mathbf{W}^T \times_3 \mathbf{C}^T,$$

де $\mathcal{X}^{V \times N \times N}$ – оригінальний тензор, що містить частотні оцінки сумісного використання трійок *підмет-присудок-додаток* (отримані з корпусу текстів), а $\mathbf{W}^{N \times K}$ – латентна фактор-матриця для іменників. Латентне представлення для виміру дієслів не використовується. Для ефективного обчислення семантичної близькості дієслів тільки виміри підметів та додатків представлені у моделі факторами, у той час як вимір дієслів представлений безпосередньо оригінальними даними. Отже, ядерний тензор $\mathcal{G}$ матиме розмір $V \times K \times K$. Модель матиме вигляд як на рисунку 6.

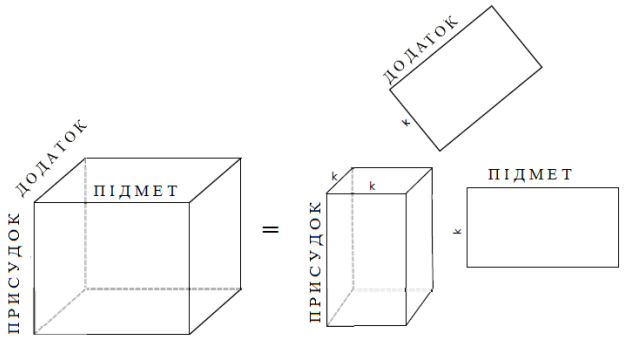

Рисунок 6. Модель без латентного виміру дієслів-присудків

Композиція трійок підмет-присудок-додаток. Для того, щоб обчислити композицію конкретної трійки *підмет-присудок-додаток*, треба виділити відповідний вектор-підмет $\mathbf{w}_s$ та вектор-додаток $\mathbf{w}_o$ з фактор-матриці $\mathbf{W}$ та обчислити тензорний добуток обох векторів, результатом якого буде матриця $\mathbf{Y}$ розміром $K \times K$.

$$\mathbf{Y} = \mathbf{w}_s \circ \mathbf{w}_o$$

Фінальним кроком є порівняння оригінальної матриці дієслів $\mathbf{G}_v$ (матриці латентної взаємодії факторів, яка є шаром тензору $\mathcal{G}$, що відповідає даному дієслову) та матриці $\mathbf{Y}$, що описує латентну взаємодію даного підмета й додатку. Порівняння виконується застосуванням операції добутку Хадамарда над $\mathbf{G}_v$ та $\mathbf{Y}$.

$$\mathbf{Z} = \mathbf{G}_v * \mathbf{Y}.$$

В такий спосіб обчислюється матриця композиційності для словосполучень *підмет-присудок-додаток*.

Приклад. Обчислимо матриці композиційності для дієслова *to damage* у контексті речень:

a) Man damages car.

b) Car damages man.

Зазначимо, що контекст дієслова відрізняється у цих реченнях лише порядком слів. У першому реченні підметом є *man*, додатком – *car*, а у другому все навпаки.

Першим кроком потрібно виділити латентні вектори для підмету та додатку з матриці $\mathbf{W}$($\mathbf{w}_{man}$ та $\mathbf{w}_{car}$ для першого речення та $\mathbf{w}_{car}$ та $\mathbf{w}_{man}$ – для другого). Далі треба обчислити тензорні добутки векторів підметів та векторів додатків – $\mathbf{w}_{man} \circ \mathbf{w}_{car}$ та $\mathbf{w}_{car} \circ \mathbf{w}_{man}$, результатом чого будуть матриці $\mathbf{Y}_{(man,car)}$ та $\mathbf{Y}_{(car,man)}$. Далі береться матриця дієслова $\mathbf{G}_{damage}$, що є шаром тензору $\mathcal{G}$, якій відповідає даному дієслову, та обчислюються добутки Хадамарда:

$$\mathbf{Z}_1 = \mathbf{G}_{damage} * \mathbf{Y}_{(man,car)};$$

$$\mathbf{Z}_2 = \mathbf{G}_{damage} * \mathbf{Y}_{(car,man)}.$$

У матрицях композиційності виділяються найбільші значення, що відповідають найтиповішим парам *підмет-додаток* у рамках контекстів наведених речень. Далі у тензорі $\mathcal{G}$ можна пошукати такі матриці дієслів, які мають подібні набори значень у тих самих елементів, що і в $\mathbf{Z}_1$ та $\mathbf{Z}_2$. Матриця $\mathbf{Z}_1$ виявила значну кореляцію з матрицями дієслів *crash*, *drive* та *ride*, що свідчить про семантичну зв'язність дієслова *to damage* у контексті першого речення до даного набору дієслів. Матриця $\mathbf{Z}_2$ виявила кореляцію з матрицями дієслів *scare*, *kill*, *hurt*, що демонструє суттєву відмінність семантики дієслова *to damage* у контексті другого речення. Таким чином, модель при обчисленні контекстної семантики речення враховує не лише набір контекстних слів, але і їх синтаксичну позицію.

Висновки

В роботі розглянуто застосування методів невід'ємної матричної та тензорної факторизації в задачах комп'ютерної лінгвістики, зокрема для латентного семантичного аналізу та кластеризації великих масивів текстів природною мовою, для автоматизації генерації таких лінгвістичних структур як селективні преференції та опису структур фреймової семантики дієслів. Також представлено модель опису композиційної семантики речень природною мовою, де за допомогою тензорної факторизації семантика кожного окремо взятого слова визначається та уточнюється у контексті інших слів речення та у контексті всього речення в цілому.

Список використаних джерел

1. Van de Cruys T. A Non-negative Tensor Factorization Model for Selectional Preference Induction / T. Van de Cruys // Journal of Natural Language Engineering. – 2010. – №16 (4). – P. 417-437.
2. Van de Cruys T. Multi-way Tensor Factorization for Unsupervised Lexical Acquisition / T. Van de Cruys, L. Rimell, T. Poibeau, A. Korhonen // In Proceedings of the International Conference on Computational Linguistics (COLING-12), Mumbai, India. – 2012. – P. 2703-2720.
3. 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.
4. 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.
5. 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.
6. Van de Cruys T. A Tensor-Based Factorization Model of Semantic Compositionality / T. Van de Cruys , T. Poibeau, A.Korhonen // Proceedings of NAACL-2013. – 2013. – P. 1142–1151.
7. 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. VAN DE CRUYS T. (2010) *A Non-negative Tensor Factorization Model for Selectional Preference Induction*, Journal of Natural Language Engineering, 16 (4), pp. 417-437.
2. VAN DE CRUYS T., RIMELL L., POIBEAU T., KORHONEN A. (2012) *Multi-way Tensor Factorization for Unsupervised Lexical Acquisition*, Proceedings of the International Conference on Computational Linguistics (COLING-12), pp. 2703-2720.
3. 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.
4. 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.
5. LEE D. D., SEUNG S. H. (2000) *Algorithms for non-negative matrix factorization*, Proceedings of the NIPS-2000, pp. 556-562.
6. VAN DE CRUYS T., T. POIBEAU T., KORHONEN A. (2013) *Tensor-Based Factorization Model of Semantic Compositionality*, Proceedings of NAACL-2013, pp. 1142–1151.
7. 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 01.12.15