УДК 004.82

ЗМІШАНА ТЕМАТИЧНО-СУТНІСНА ОНТОЛОГІЯ У ПОКРАЩЕНІЙ ТЕМАТИЧНІЙ ВЕКТОРНІЙ МОДЕЛІ

А.С. Шабінський

Національний університет «Києво-Могилянська академія»,
Україна, 04655, Київ, вул. Сковороди 2.
E-mail: anton.shabinskiy@gmail.com

Розглядається моделювання онтологій у покращеній тематично векторній моделі інформаційного пошуку. Запропонований підхід орієнтований на автоматизацію видобування онтологій. Проаналізовано методи моделювання тематичної структури колекцій документів ймовірнісними тематичними моделями та основні підходи у видобуванні та розв'язанні іменованих сутностей.

Ключові слова: інформаційний пошук, онтологія, тематична векторна модель, тематична карта**.**

The paper considers to modelling ontologies in enhanced topic-based vector-space model of information retrieval. Proposed approach is oriented on ontology extraction automation. Methods of modelling topical structure of collections of documents with probabilistic topical models and named entity recognition, as well as possible interpretation are reviewed.

Key words: information retrieval, ontology, topic vector-space model, topic map

Вступ

Онтології є однією із форм подання знань у інформаційних системах (ІС). Саме поняття «онтологія» є дещо абстрактним і не визначено строго. Натомість, воно лише окреслює загальну концепцію формалізації предметних областей для застосування у ІС, для використання як людьми, так і комп'ютерними системами у складі ІС.

За роллю у кінцевій ІС онтології можна поділити на два великі класи: онтології, що є самостійними базами знань (наприклад, у якості центральних компонентів експертних систем, або як цілком автономні бази знань (БЗ) для обслуговування зовнішніх запитів); та онтології у онтолого-керованих системах (ОКІС, див. напр. [1–3]), де кінцева функціональність системи не пов'язана безпосередньо із онтологією і представленими у ній знаннями, а лише скеровується онтологією. Прикладом ІС другого класу може бути онтолого-керовано інформацій-но-пошукова система (ОКІПС), у якій кінцевого користувача не цікавить внутрішній механізм поведінки системи, а лише вирішення задачі пошуку релевантної інформації. Проте система здійснює пошук саме на основі БЗ, що представлена внутрішньою онтологією.

Покращена тематична векторна модель

Покращена тематична векторна модель (enhanced topic-based vector space model – eTVSM) [4] еволюціонувала із тематичної векторної моделі (TVSM), а та у свою чергу – із класичної векторної моделі (VSM). У моделі TVSM зроблено спробу подолати обмеження VSM, при цьому зберігається формальний підхід, модель подано у загальному вигляді. Модель TVSM було запропоновано у [5] як послідовницю VSM, покращену за рахунок чутливості до відношень між словами. Найперше поліпшення моделі було досягнуто шляхом усунення припущення про ортогональність термінів. Натомість було введено поняття фундаментальних тем, які є векторами у ортогональному базисі векторного простору.

Модель TVSM еволюціонувала у покращену TVSM, де детально пророблена концепція визначення відношень між поняттями завдяки усуненню незалежності між темами і використання онтологій як джерела знань про семантичні зв'язки між поняттями предметних областей. У eTVSM спосіб визначення схожості документів побудований не на принципі схожості термінів, а на основі концепції інтерпретацій термінів. Модель оперує поняттями слова, основи слова, терміна, інтерпретації та теми. У моделі eTVSM інтерпретації використовуються як проміжні ланки між темами та термінами, несуть семантичне навантаження. Моделі TVSM та eTVSM ми розглядали у [6], де зокрема наведено формальний метод обрахунків схожості документів.

Модель eTVSM є гнучкою в реалізації завдяки ряду місць, де розробники ОКІПС можуть застосувати власні рішення і досягнути відповідних результатів, цілком незалежно від основних концепцій моделі. Наведемо тут короткий опис таких можливостей.

Моделювання онтологій. Підхід до моделювання онтологій ніяк не регламентується у eTVSM. Розробники отримують повну свободу і можуть на власний розсуд застосувати такий спосіб представлення онтології, який більш відповідає потребам, задачам, вимогам тощо.

Ваги інтерпретацій. У роботі [6] наведено формалізм розрахунку векторів документів як зваженої суми векторів інтерпретацій. Проте, не вводиться жодних обмежень на механізм визначення самих ваг, окрім вимоги їх належності інтервалу [0; 1]. Це ще одна можливість для розробників самостійно визначити поведінку системи.

Зв'язування інтерпретацій із поняттями. Кожному поняттю у документів має відповідати строго одна інтерпретація, відповідно до значення, у якому вжито поняття. Оскільки при моделюванні онтології поняттям може бути зіставлено декілька інтерпретацій, необхідно визначити принцип, за яким із кількох обиратиметься найбільш доречна інтерпретація. Одним із варіантів, запропонованим у [4], є супутні поняття, які зазвичай зустрічаються у мові поряд із певним поняттям. Такі супутні поняття можуть використовуватись як індикатори контексту, у якому вжито певне поняття. Знову, розробники вільні у імплементації бажаного механізму вибору інтерпретацій.

Визначення понять та попередня обробка. Від розробників системи ІП залежатиме, наскільки вдало групи слів будуть розпізнані як складені поняття. Зокрема, важливу роль відіграє черговість попередньої обробки тексту та визначення понять. Проведення попередньої обробки тексту перед визначенням понять може призвести до хибної трактовки багатьох складених термінів, адже будуть втрачені стоп-слова, форми слів тощо.

Онтологія у eTVSM

Модель eTVSM виглядає привабливою для побудови високоефективних пошукових систем, здебільшого за рахунок онтологій і тих можливостей, що вони дають, а саме чутливість до семантичних зв'язків між поняттями у документах. Очевидно, це дає суттєву перевагу у порівнянні із класичними пошуковими системами, де весь процес пошуку ґрунтується, так чи інакше, на ключових словах та їх словоформах. Але головною проблемою ОКІПС, і зокрема заснованої на eTVSM, є розбудова онтологій. Від якості онтологій залежить ефективність пошукових систем. Зокрема у [4] запропоновано підхід до автоматичної побудови онтології eTVSM на основі WordNet. Для оцінки цього та інших підходів у [7] проведено ряд порівнянь, у результаті чого зроблено висновок, що eTVSM із онтологією на основі WordNet має гіршу ефективність аніж eTVSM із онтологією синонімів і навіть VSM. Причиною цього є те, що WordNet – онтологія загального призначення і не може відобразити більшості усталених складних понять і особливості їх значення та контексту. Також, подібний підхід до реалізації семантичних можливостей ІП запропоновано у [8]. Щоправда, у праці замість інтерпретацій eTVSM використовуються анотації цілих документів, які зіставляють поняття у документі із певними предметними областями.

У роботі [9] аналізували роль та місце онтологій у ОКІПС, розглядали різні ступені деталізації онтології у контексті інженерії та автоматизації. Пропонований нами підхід передбачав використання тематичних карт у загальному вигляді як збалансованого варіанту між відтворенням семантики документів та придатності до автоматизованої розбудови. Тепер, конкретизуючи використання онтології у моделі eTVSM, ми пропонуємо використати особливу за структурою онтологію, яку будуватимемо із двох частин. Перша – іменовані сутності, представлені у документах, отримані автоматичними методами розпізнавання (*англ.* named entity recognition). Друга частина онтології – тематична анотація колекції у формі тематичної карти, отриманої за допомогою ймовірнісної тематичної моделі. Таким чином маємо змішану тематично-сутнісну онтологію.

Наведемо формалізм для змішаної онтології та супутніх понять. Онтологія є структурою виду:

$$\Omega = \langle E, \Theta, A_H, A_C, A_\Theta \rangle, \tag{1}$$

де $E \subset T$ – іменовані сутності, T – всі терміни, $\Theta = \{\theta_1, \theta_2, ..., \theta_t\}$ – теми, A_H, A_C, A_Θ – зв'язки у онтології.

Відношення A_H – це ієрархічні асоціації між темами, коли одні теми є підтемами інших. Відношення задано так:

$$A_H : \Theta \to 2^{(\Theta \setminus \theta_i)}, \tag{2}$$

при цьому $A_H(\theta_i) \subseteq \Theta \setminus \theta_i$ є множиною усіх батьківських тем деякої теми θ_i. Далі у окремому пункті ми розглянемо можливу інтерпретації ієрархічних зв'язків у векторне подання тем і числові характеристики спорідненості тем.

Відношення A_C – це зважені асоціації між іменованими сутностями та темами. Відношення задано наступним чином:

$$A_C : E \times \Theta \to R, \tag{3}$$

де $A_C(e_i, \theta_j) \in [0;1]$ – ваги. Тобто відношення A_C кожній парі сутність-тема співставляє дійсну вагу із проміжку [0;1]. Це є відображенням того факту, що кожна іменована сутність може належати різним темам одночасно, але мати у цих темах різне значення. Наприклад, сутність «Білл Гейтс» належить одночасно темам «Майкрософт», «Програмне забезпечення» та «Благодійність», але у темі «Майкрософт» вага сутності найбільша, у темі «Програмне забезпечення» – менша, і у темі «Благодійність» – найменша.

Нарешті, відношення A_Θ моделює прості семантичні зв'язки між темами. Відношення визначається так:

$$A_\Theta : \Theta \times \Theta' \to R, \tag{4}$$

де $A_\Theta(\theta_i, \theta_j) \in [0;1]$ – ваги, при цьому Θ' – множина тем, така, що

$$\forall \theta_i \in \Theta, \Theta'_i = \left\{ \theta' : \theta' \notin A_H^*(\theta_i) \wedge \theta_i \notin A_H^*(\theta') \right\}, \tag{5}$$

де $A_H^*(\theta)$ – усі батьківські теми для θ. Іншими словами, для теми θ_i множина Θ'_i є множиною усіх тем, які не входять до ієрархії θ_i, тобто ні є ані батьками, ані нащадками теми θ_i. Таким чином, відношення A_Θ дозволяє задавати довільні семантичні зв'язки між тими темами, які ніяк не зв'язані ієрархічно. Відношення A_Θ надає додаткову свободу розробникам кінцевих систем, оскільки дає змогу впливати на вагу окремих тем у документах незалежно від тематичних ієрархій, за потреби підсилюючи чи послаблюючи вплив тих чи інших тем на інтерпретацію документа.

Нагадаємо, основна ідея моделі eTVSM полягає у використанні інтерпретацій як проміжних об'єктів між темами та документами. Модель документа у eTVSM будується із інтерпретацій, а не з безпосередньо тем. У нашому випадку інтерпретації повинні базуватися на семантиці, яка задана онтологією запропонованої структури, тобто враховувати вищезгадані зв'язки і ваги. Важливим аспектом є те, що на цьому етапі ми залишаємо достатньо гнучкості і свободи у моделі. При розробці системи можна моделювати лінгвістичні та семантичні особливості, обираючи власні вагові схеми та по-різному інтерпретуючи зв'язки у онтології. Ми визначаємо загальний підхід у поданні моделі документа, але не обмежуємо способи її обрахунку. У термінології eTVSM та згідно із нашим підходом інтерпретації задані $\phi_f \in \Phi$, та задано множину $\Omega'(\phi_f) \in 2^{\Theta \cup E}$, яку побудуємо у три кроки наступним чином:

1. $\theta_i, \theta_j \in \Theta : \theta_i \in A_H^*(\theta_j) \vee \theta_j \in A_H^*(\theta_i)$
2. $e \in E : A_C(e, \theta_i) > 0 \vee A_C(e, \theta_j) > 0$
3. $\theta_k \in \Theta : A_\Theta(\theta_k, \theta_i) > 0 \vee A_\Theta(\theta_k, \theta_j) > 0$

Тобто $\Omega'(\phi_f)$ – множина довільних об'єктів онтології, які

- або є темами, що пов'язані ієрархічно (крок 1);
- або є сутностями, безпосередньо пов'язаними із цими темами (крок 2);
- або є темами, що безпосередньо пов'язані семантичними зв'язками із темами з кроку 1.

Таким чином ми отримали зв'язок тем та сутностей із інтерпретаціями, як це вимагається у моделі eTVSM. Зазначимо, що ми маємо право об'єднувати теми Θ та сутності E завдяки тому, що у векторному поданні ці об'єкти є сумісними, оскільки всі вектори мають розмірність $t = |\Theta|$. Сутності подаються у вигляді векторів:

$$\vec{e}_i = (A_C(e_i, \theta_1), A_C(e_i, \theta_2), \ldots, A_C(e_i, \theta_t)). \tag{6}$$

Теми аналогічно представлені векторами:

$$\vec{\theta}_i = (g_\Theta(\theta_i, \theta_1), g_\Theta(\theta_i, \theta_2), \ldots, g_\Theta(\theta_i, \theta_t)), \tag{7}$$

тут $g_\Theta(\theta_i, \theta_j)$ – деяка узагальнююча функція зважування, яка інкапсулює у собі і $A_H(\theta_i)$, і $A_\Theta(\theta_i, \theta_j)$. Зрештою, ми можемо перейти до обчислення векторів інтерпретацій. Позначимо вектори $\vec{\theta}_i$ і $\vec{e}_i$ загальним вектором $\vec{\omega}_i \in \Omega'$, тоді вектор інтерпретації має вигляд:

$$\vec{\phi}_i = \frac{g_\Phi(\phi_i)}{\left| \sum_{\omega_k \in \Omega'(\phi_i)} \vec{\omega}_k \right|} \sum_{\omega_k \in \Omega'(\phi_i)} \vec{\omega}_k . \tag{8}$$

Подальша побудова та обрахунок моделі документа є повністю уніфікованою із звичним підходом у моделі eTVSM, де документ є вектором, обрахованим як зважена сума векторів інтерпретацій. Цей формалізм ми уже розглядали у [6].

Розпізнавання іменованих сутностей

Проблема розпізнавання іменованих сутностей вперше була сформульована на 6-й конференції Message Understanding Conference у 1995 р.

Усі підходи до видобування іменованих сутностей можна поділити на три класи: засновані на довідниках, засновані на правилах та статистичні. Є також змішані підходи, які поєднують декілька різних. Зауважимо,

що майже одразу після появи задач із розпізнавання сутностей довідникові підходи були визнані неефективними (і подеколи незастосовними). Зокрема, у статті «Розпізнавання сутностей без довідників» [10] наведено результати для чистого довідникового розпізнавання 90–94 % точності та 75–78 % повноти для географічних місць і 75–85 % та менше 50 % для особистих імен та організацій. Окрім того, довідники неможливо підтримувати для особистих імен та назв організацій. Проте, там же зазначено, що без довідників вкрай складно розпізнавати географічні місця, які зазвичай з'являються у тексті без достатнього контексту, аби бути розпізнаними статистичними методами або правилами. Зокрема, запропонований у [10] підхід поєднує контекстні правила, статистичні методи, та довідники.

Машинне навчання та статистичні методи у обробці природних мов часто залежать від вчителя у сенсі наявності розміченої навчальної вибірки. У розпізнаванні іменованих сутностей проблема наглядового навчання теж має місце, і є методи, які базуються на попередній розмітці навчальної вибірки. Проте, безнаглядові методи актуальніші, оскільки є більш універсальними та легше підтримуваними. У [11] розглянуто можливий перехід від алгоритмів із вчителем до алгоритмів без вчителя і мінімізація обов'язкової розмітки навчальної вибірки. Зокрема, для тестової навчальної вибірки у 90 000 зразків метод досягає 91% точності, а явно заданих правил вимагається всього 7.

Також є кілька запатентованих алгоритмів розпізнавання сутностей з урахуванням їх семантики, зокрема, у компанії Xerox – із пошуком зв'язків між сутностями [12] та із розпізнаванням метонімії [13]; у компанії IBM – із розпізнаванням сутностей з певної предметної області за допомогою N-грамних моделей [14] та із розв'язанням сутностей (entity resolution – розпізнавання різних за формою згадувань однієї сутності) [15].

Тематичні карти

Тематичні карти [16, 17] є однією з форм представлення онтологій. Концепти онтології у тематичній карті (ТК) подаються темами та зв'язками (асоціаціями) між ними. Тематичні карти багато в чому подібні до стандарту W3C RDF [18, 19], хоча останні орієнтовані на ресурси, а не на теми як такі. Онтологія в ОКІПС може мати будь-яку форму і походження. Ми обираємо тематичні карти з двох причин. Вони дозволяють інтуїтивно моделювати зміст документів як розкриття їх тематики у різних пропорціях; по-друге – придатні до автоматизованої побудови з мінімальною участю людини-експерта. Тематичні карти як онтології у ОКІПС вже розглядалися нами у [9].

Тема у ТК є машинно-читабельним представленням деякого концепту. Не існує жодних обмежень на природу концептів, що можуть позначатися темами, проте є чотири основні форми ідентифікації тем у межах ТК:

1. Ідентифікатор теми як ресурсу у серіалізованій ТК: таким ідентифікатором виступає URI (Uniform Resource Identifier за стандартом RFC 3986) і є унікальним в межах ТК.
2. Ідентифікатор теми у вигляді людино-читабельного ярлику: тема може мати довільну кількість назв, доступних людині для розуміння.
3. Ідентифікація за посиланням: для ресурсів, що мають власний URI ідентифікатор пов'язної теми є похідним від ідентифікатора ресурсу, що дає змогу означувати кожну тему за відповідним їй ресурсом.
4. Ідентифікатор за описом: деякі теми можуть позначати концепти, що не є ідентифіковними за URI (наприклад, люди), але асоціюються із певними описовими сутностями (реєстраційні картки, фотографії, анкети тощо), а відтак ідентифікуються цими сукупностями описової інформації.

Важливо зауважити, що хоча тема може мати довільну кількість ідентифікаторів, кожний окремий ідентифікатор повинен однозначно вказувати на конкретну тему.

Відношення між темами у ТК є трьох типів:

- тип – екземпляр («is-a»),
- супертип – підтип («kind-of»),
- (рольові) асоціації.

Рольові асоціації є n-арними відношеннями, що можуть включати довільну кількість тем як учасників. Участь кожної теми у асоціації визначається її роллю. За допомогою асоціацій можна моделювати зв'язки з будь-якою семантикою. Наприклад, у предметній області «об'єктно-орієнтоване програмування», визначивши теми «ситуативний поліморфізм» та «перевантаження функцій» між ними можна встановити асоціацію з ролями «concept» та «technique» відповідно. Семантично така асоціація моделює явище, коли перевантаження функцій як механізм у мові програмування є технікою реалізації ситуативного поліморфізму як загальної концепції.

Також, відношення «супертип – підтип» насправді теж є асоціацією, але спеціального типу, визначеного заздалегідь. У цій асоціації вже визначено дві ролі («супертип» та «підтип») і кількість учасників обмежено двома.

На користь тематичних карт слід зауважити, що тематичні карти забезпечені стандартизованим XML синтаксисом [20] та специфікацією прикладного програмного інтерфейсу (API) з різноманітними імплементаціями [21]. Також тематичні карти стандартизовані за ідентифікатором ISO/IEC 13250.

Інтерпретація тематичних карт для eTVSM

У роботі [22] запропоновано один алгоритм перетворення тематичних карт найпростішої структури – дерева лише із відношенням наслідування – у числові характеристики спорідненості тем. Нехай на множині тем визначено ієрархію, задану відношенням $A_H(\theta_i)$. Тоді множина усіх батьківських тем теми θ_i не вище, ніж p рівнів вгору має вигляд:

$$A_H^p(\theta_i) = \bigcup_{\theta_k \in A_H^{p-1}(\theta_i)} A_H(\theta_k), \tag{9}$$

а множина усіх батьківських тем теми θ_i, має вигляд:

$$A_H^*(\theta_i) = \bigcup_{k=1}^{l-1} A_H^k(\theta_i), \tag{10}$$

де l – глибина вузла для теми θ_i у загальному дереві тем.

Кожній темі θ_i відповідає вектор $\vec{\theta}_i = (\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,t}) \in R^t$. При цьому компоненти вектора для листкової теми визначаються так:

$$\theta_{i,d} = \begin{cases} 1, \theta_d \in A_H^*(\theta_i) \vee i = d, \\ 0. \end{cases} \tag{11}$$

Для внутрішніх вузлів дерева вектори тем обраховуються як нормовані суми векторів всіх прямих нащадків:

$$\vec{\theta}_{\text{int}} = \left\| \sum_{\theta_S \in \Theta : \theta_{\text{int}} \in A_H(\theta_S)} \vec{\theta}_S \right\|. \tag{12}$$

Звідси спорідненість двох тем θ_a та θ_b може бути обрахована як скалярний добуток їх векторів, який, завдяки нормуванню векторів, є косинусом кута між векторами $\omega_{a,b}$:

$$sim(\theta_a, \theta_b) = \left|\vec{\theta}_a\right\| \left|\vec{\theta}_b\right| \cos \omega_{a,b} = \cos \omega_{a,b}. \tag{13}$$

Тематичні карти із зв'язками наслідування найбільш придатні для автоматизованої розбудови. Зокрема, одним із підходів, що уможливлює таку автоматизацію, є ймовірнісні тематичні моделі. Ймовірнісні тематичні моделі (ТМ) – це алгоритми для виявлення тематичного наповнення документів у великих неструктурованих колекціях. [23] Найпростішою тематичною моделлю є приховане розміщення Діріхле (latent Dirichlet allocation).[24] Як одна з перших ймовірнісних моделей для текстів LDA базується на ряді припущень, що практично унеможливлюють її використання у розроблюваних ОКІПС. Модель спирається на поданні текстів як «торби слів» («bag of words»), тобто ігнорує порядок та зв'язок слів у документах. Окрім того, породжувальний процес моделі передбачає лише однорівневе моделювання тем без врахування ієрархічних зв'язків між темами, що є недоречним у практичних застосуваннях. Натомість у контексті побудови ієрархічних тематичних карт для ОКІПС цікавими є корельована тематична модель [25] та модель розміщення патінко (pachinko) [26]. Моделі описують значно складніший породжувальний процес, що покликаний відтворити більш природній стан речей у тематичній ієрархії, тобто врахувати, що деякі теми можуть бути підтемами інших. Як наслідок, зазначені моделі можна використати для навчання на еталонних колекціях документів, а навчені моделі застосувати для виведення (inference – процес, коли навчена модель опрацьовує довільні колекції ресурсів) тематичних структур (ієрархій) на кінцевих репозиторіях ресурсів.

Висновки

У статті розглянуто представлення онтології у покращеній тематичній векторній моделі інформаційного пошуку. Простір можливих реалізацій моделі є досить широким, оскільки модель не задає жодних обмежень на будову онтології та методи її інтерпретації.

Ймовірнісні тематичні моделі та алгоритми розпізнавання іменованих сутностей є перспективними методами автоматизації розбудови та інтерпретації онтологій у eTVSM. Вони дозволяють видобувати вагому частину значущого змісту документів, а отримані онтологічні структури придатні для подальшої інтерпретації.

У статті запропоновано спеціальну змішану онтологію для eTVSM, яка побудована із тем, іменованих сутностей, та зв'язків між ними. Наведено формальну модель онтології та підхід до побудови інтерпретацій, залишивши гнучкість та свободу у конкретних методах обчислення моделі документа.

Подальші дослідження варто зосередити на підвищенні виразності онтології eTVSM, зокрема на методах видобування та інтерпретації рольових асоціацій у тематичних картах. Цікавим бачиться детальніше дослідження механізмів зваженої інтерпретації іменованих сутностей.

1. *Palagin A.V., Petrenko N.G.* Towards designing ontology-driven information system with natural language processing // Mathematical machines and systems. – 2008. – № 2 – P. 14–23.
2. *Palagin A. V., Petrenko N. G.* Architecture-ontological principles of developing intelligent information systems // Mathematical machines and systems. – 2006. – № 4 – P. 15–20.
3. *Qiu R. G.* Towards ontology-driven knowledge synthesis for heterogeneous information systems // Journal of Intelligent Manufacturing. – 2006. – N 1, Vol. 17. – P. 99–109.
4. *Kuropka D.* Modelle zur Repräsentation natürlichsprachlicher Dokumente. – Berlin: Logos Verlag, 2003.
5. *Becker J., Kuropka D.*. Topic-based Vector Space Model // Proceedings of BIS. – Colorado Springs, USA: Business Information Systems, 2003.
6. *Glibovets A.N., Glibovets N.N., Shabinskiy A.S.* Application of Ontologies and Text Mining Methods to the Development of Intelligent Information Retrieval Systems. // Journal of Automation and Information Sciences. – 2011. – N 6. – P. 95–102.
7. *Polyvyanyy A.* Evaluation of a Novel Information Retrieval Model: eTVSM. – Potsdam: HPI, 2007.
8. *Vallet D., Fernández M., Castells P.* An Ontology-Based Information Retrieval Model. – Madrid : Proc. Second European Semantic Web Conf., 2005.
9. *Shabinskiy A.* Ontologies, probabilistic topical models, and topic maps // Scientific Notes of NaUKMA. – 2013. – Vol. 151. – P. 60–65.
10. *Grishman R., Sundheim B.* Message Understanding Conference – 6: A Brief History // Proceedings of the 16th International Conference on Computational Linguistics. – Kopenhagen: [s.n.], 1996. – Vol. I.
11. *Mikheev A., Moens M., Grover C.* Named Entity Recognition without Gazetteers // Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. – 1999.
12. *Collins M., Yoram S.* Unsupervised models for named entity classification // Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora. – 1999.
13. *Brun C., Hagege C.* Semantically-driven extraction of relations between named entities [Patent]: 8,370,128. – USA, 5 February 2013.
14. *Brun C., Ehrmann M., Jacquet G.* Hybrid system for named entity resolution [Patent]: 8,374,844. – USA, February 12, 2013.
15. *Kanungo T., Rhodes J.* System and method for extracting entities of interest from text using n-gram models [Patent]: 7,493,293. – USA, February 17, 2009.
16. *Caceres B. M.* Entity resolution based on relationships to a common entity [Patent]: 13/217,027 (application). – USA, 28 February 2013.
17. *Ahmed K., Moore G.* An Introduction to Topic Maps // The Architecture Journal. – [s.l.]: Microsoft Corporation. – 2005. – N 5. – P. 3–9.
18. ISO/IEC JTC1/SC34/WG3, "Topic Maps — Part 1: Overview and Basic Concepts" [Online]: http://www.itscj.ipsj.or.jp/sc34/open/1045.htm.
19. *Berners-Lee T.* Notation3 (N3): A readable RDF syntax // World Wide Web Consortium. – http://www.w3.org/DesignIssues/Notation3.
20. *Beckett D., McBride B.* RDF/XML Syntax Specification // W3C. – Лютий 10, 2004. – http://www.w3.org/TR/REC-rdf-syntax/.
21. ISO/IEC JTC1/SC34/WG3, "Topic Maps — XML Syntax" [Online]:
22. http://www.isotopicmaps.org/sam/sam-xtm/.
23. "Common Topic Map Application Programming Interface" [Online]. Available: http://www.tmapi.org.
24. *Kuropka D.* A proposal for transformation of topic-maps into similarities of topics. – 2005.
25. *Blei D. M.* Probabilistic Topic Models // Communications of the ACM. – New York: ACM, 2012. – 4, T. 55.
26. *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet Allocation // Journal of Machine Learning Research. – Cambridge, MA : MIT Press, 2003. – Vol. 3. – P. 993–1022.
27. *Blei D., Lafferty J.* A correlated topic model of SCIENCE // The Annals of Applied Statistics. – 2007. – N 1, Vol. 1. – P. 17–35.
28. *Li W., McCallum A.* Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations // Proceedings of the 23rd International Conference on Machine Learning. – Pittsburg: [s.n.], 2006.International Conference on Machine Learning, Pittsburg, 2006.