УДК 004.82

Шабінський А. С.

ІНТЕГРАЦІЯ ТЕМАТИЧНО-СУТНІСНОЇ ОНТОЛОГІЇ В ІНФРАСТРУКТУРУ ІНФОРМАЦІЙНО-ПОШУКОВОЇ СИСТЕМИ

У статті розглянуто тематично-сутнісна онтологія покращеної тематичної векторної моделі у контексті загальної інфраструктури інформаційно-пошукової системи. Запропоновано варіанти адаптації процесів у системі під онтолого-керований сценарій роботи. Наведено моделі та способи реалізації основних компонентів з урахуванням практичних аспектів та особливостей використаної онтології.

Ключові слова: індексація, інформаційний пошук, онтологія, ранжування, тематична векторна модель.

Вступ

Застосування онтологій у інформаційно-пошукових системах вимагає адаптації певних компонентів та процесів системи під нові сценарії використання. Масштаби та особливості такої адаптації залежать від місця онтології у системі та ролі у загальному процесі пошуку. Якщо ми говоримо про онтолого-керовані інформаційно-пошукові системи (ОКІПС), то у таких системах онтології посідають центральне місце і скеровують роботу системи. Онтології в ОКІПС впливають на всі ключові процеси: індексацію, побудову моделі документа, інтерпретацію інформаційної потреби, фільтрацію, пошук, ранжування. Проте конкретні варіанти та практичні аспекти адаптації залежать від особливостей обраної онтологічної структури. Ми розглянемо питання, пов'язані з інтеграцією тематично-сутнісної онтології моделі eTVSM [6] в загальносистемну пошукову інфраструктуру та покажемо, як наведена теоретична модель співвідноситься із практичними аспектами організації пошуку.

Тематично-сутнісна онтологія моделі eTVSM

Наведемо формалізм для змішаної онтології. Онтологія є структурою виду

$\Omega = \langle E, \Theta, A_H, A_C, A_\Theta \rangle$,

$E \subset T$ – іменовані сутності,

T – всі терміни,

$\Theta = \{\theta_1, \theta_2, \ldots \theta_t\}, |\Theta| = t$ – теми,

A_H, A_C, A_Θ – відношення у онтології.

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

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

при цьому $A_H(\theta_i) \subseteq (\Theta \setminus \theta_i)$ є множиною тем, для яких θ_i є *безпосереднім* нащадком (тобто тільки теми на рівень вище від заданої). Далі в окремому пункті ми розглянемо можливу інтерпретацію 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 таким чином:

$A_C : E \times \Theta \to R$,

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, а у темі «Благодійність» – найменша.

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

$A_\Theta : \Theta \times \Theta' \to R$,

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

$\forall \theta_i \in \Theta, \Theta_i' = \{\theta' : \theta' \notin A_H^*(\theta_i) \wedge \theta_i \notin A_H^*(\theta')\}$,

де $A_H^*(\theta)$ – усі батьківські теми усіх ієрархічно вищих рівнів для θ. Іншими словами, для теми θ_i множина Θ_i' є множиною усіх тем, які

не входять до ієрархії θ_i, тобто не 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 пов'язані ієрархічно. Відношення A_Θ надає додаткову свободу розробникам кінцевих систем, оскільки дозволяє 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 з інтерпретацій, а не безпосередньо з тем. У нашому випадку інтерпретації повинні базуватися на семантиці, яка задана онтологією запропонованої структури, тобто враховувати згадані вище 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 в 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 по черзі для кожної теми із онтології у три кроки:

$$|\Phi| = |\Theta|, \forall \theta_i \in \Theta, \exists \phi_i \in \Phi : \Omega'(\phi_i) =$$

$$= \begin{cases} \theta_j, \theta_j \in \Theta : \theta_i \in A_H{}^*(\theta_j) \vee \theta_j \in A_H{}^*(\theta_i) \vee i = j & (1) \\ e, e \in E : A_C(e, \theta_i) > 0 \vee A_C(e, \theta_j) > 0 & (2) \\ \theta_k, \theta_k \in \Theta : A_\Theta(\theta_k, \theta_i) > 0 \vee A_\Theta(\theta_k, \theta_j) > 0 & (3) \end{cases}$$

Тобто $\Omega'(\phi_f)$ – множина довільних об'єктів онтології, які обрані у такому порядку:

крок 1: теми, що пов'язані ієрархічно із θ_i;

крок 2: сутності, безпосередньо пов'язані із темами з кроку 1;

крок 3: теми, 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:

$$\vec{e_i} = \left(A_C(e_i, \theta_1), A_C(e_i, \theta_2), \ldots, A_C(e_i, \theta_t)\right)$$

Теми аналогічно представлено векторами:

$$\vec{\theta_i} = \left(g_\Theta(\theta_i, \theta_1), g_\Theta(\theta_i, \theta_2), \ldots, g_\Theta(\theta_i, \theta_t)\right),$$

тут $g_\Theta(\theta_i, \theta_k)$ – деяка узагальнена функція зважування, яка інкапсулює у собі і $A_H(\theta_i)$, і $A_\Theta(\theta_i, \theta_i')$. 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:

$$\vec{\phi_i} = \frac{g_\Phi(\phi_i)}{\left|\sum_{\omega_k \in \Omega'(\phi_i)} \vec{\omega_k}\right|} \sum_{\omega_k \in \Omega'(\phi_i)} \vec{\omega_k}.$$

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 [1].

Практичні аспекти моделі документа

У загальному вигляді модель документа у eTVSM вимагає співставлення інтерпретацій усім термінам із документа [1]. При цьому ми визначаємо інтерпретації $\phi_i \in \Phi$ як деякі абстрактні об'єкти, що є проміжними ланками між елементами онтології та реальними термінами зі словника колекції. Зазначимо, що з практичної точки зору, із формальним апаратом eTVSM, спосіб задання інтерпретацій залежить від реалізації системи. Ми можемо подати цей спосіб у загальному вигляді як деяку функцію

$$\psi : \phi_i \to 2^{\Theta \cup E},$$

яка співставляє кожній інтерпретації підмножину елементів онтології $\Omega'(\phi_i)$. Це дає змогу будувати модель документа (інтерпретувати документи) відповідно до схеми eTVSM. На практиці цей процес пов'язаний із двома аспектами:

1) подання зважувальної функції $g_\Phi(\phi)$, яка впливає на вагу теми/сутності у векторі інтерпретації;

2) вибір інтерпретації для кожного терміна у документі.

Функція $g_\Phi(\phi)$ у найпростішій імплементації може надавати кожному компоненту вектора вагу 1, але це суттєво обмежує семантичну виразність системи. Натомість, надання ваг із інтервалу [0;1], як запропоновано у [6], дає змогу урізноманітнити векторний простір інтерпретацій і досягти повнішої і точнішої передачі тематично-сутнісного змісту кожної інтерпретації.

Розглянемо приклад застосування зважувальної функції для інтерпретацій. Визначимо функцію $g_\Ö(\phi)$ так:

$$g_{\Phi}(\phi) \overset{def}{=} g:(\phi,\omega) \to R,$$
$$\phi \in \Phi, \omega \in \Theta \cup E, g(\phi,\omega) \in [0;1],$$

$$g(\phi_i,\omega_j) = \begin{cases} \dfrac{1}{|l_i - l_\omega + 1|^{|l_i - l_\omega + 1|}}, \\ \text{якщо } \omega_j \in \wedge\left(\theta_i \in A_H^*(\omega_j) \vee \omega_j \in A_H^*(\theta_i) \vee i = j\right), \\ A_C(\omega_j,\theta_i), \text{ якщо } \omega_j \in E \\ A_\Theta(\omega_j,\theta_i), \text{ якщо } \omega_j \in \Theta \wedge A_\Theta(\omega_j,\theta_i) > 0 \end{cases}$$

де l – рівень відповідного вузла у дереві тем, а $A_H^*(\theta)$ – усі батьківські теми для θ. Зважувальну функцію ми визначили як величину, що залежить від типу об'єкта. Для тем з ієрархії – це величина, що обернено залежить від віддаленості тем одна від одної. Тобто для тем на рівень вище (чи нижче) заданої, $g(\phi,\omega) = 0{,}25$, для наступного рівня – $0{,}037$, для заданої теми – 1. Так теми, що ближчі ієрархічно, мають більшу вагу у кінцевому векторі інтерпретації. Для сутностей та семантично пов'язаних тем ми використали існуючі ваги, задані відношеннями в онтології. Певна річ, наведена зважувальна функція є лише одним із можливих підходів до побудови векторів інтерпретацій.

Для практичної побудови моделі документа у eTSVM необхідно проінтерпретувати кожний термін документа, тобто подати документ як певну комбінацію інтерпретацій, а формально – як суму їхніх векторів, отримавши відповідний вектор-документ [1]. Зв'язки між інтерпретаціями та термінами можна видобути із даних, отриманих за допомогою ймовірнісних тематичних моделей (ТМ) [3]. Нагадаємо, завдяки ТМ можна отримати розподіли на темах для кожного документа та розподіли на словнику для кожної теми. Наявність цієї інформації відкриває кілька ключових можливостей для інтерпретації термінів у рамках тематично-сутнісної онтології моделі eTVSM. По-перше, можна запозичити спосіб подання термінів із моделі TVSM (з якої еволюціонувала eTVSM). Там одним із ключових елементів є термін, вектор якого визначає зв'язок терміна із темами:

$$t_i \in T, \vec{t_i} = (t_{i,1}, t_{i,2} \ldots t_{i,\tau}), |\Theta| = \tau.$$

Цікаво, що саме компоненти вектора $\vec{t_i}$ нам доступні з результатів ТМ. Оскільки ми знаходимося в рамках нашого векторного простору розмірності τ, можна знайти найближчу до терміна інтерпретацію ϕ_{t_i} знайшовши вектор інтерпретації із найменшим кутом відносно вектора терміна. Втім цей спосіб хоч і є досить інтуїтивним і простим, але не є ефективним. По-перше, обчислення інтерпретацій таким способом є перебірним і досить обтяжливим: для кожного терміна із словника потрібно обрахувати всі скалярні добутки з векторами інтерпретацій, а це означає часову оцінку $O(mn)$, де $m = |\Phi|, n = |T|$. По-друге, вказаний спосіб повністю ігнорує можливості різної інтерпретації термінів залежно від контексту всього документа. Натомість кожний термін із словника отримує єдино можливу інтерпретацію, яка буде спільною для усіх документів, де зустрічається термін. Такий підхід може бути виправданий для вузькоспеціалізованих ОКІПС, де висока ймовірність того, що один термін завжди зустрічається в одному контексті. Для універсальних ОКІПС цей підхід буде вкрай неефективним.

Більш перспективним є підхід із використанням інформації про розподіл на темах для кожного документа. Природа ТМ дозволяє трактувати цей розподіл як пропорції тем у документі, а отже, отримати комбінації векторів тем для документів і точніше передати семантичний зміст документів при їхній інтерпретації.

Інтерпретація інформаційної потреби користувача

Першим кроком до створення задовільної з погляду користувача пошукової системи є забезпечення інтерпретації інформаційної потреби. З погляду користувача саме форма і спосіб вираження інформаційної потреби у системі є бар'єром між системою та користувачем, який зрештою визначає результат роботи. Класичний пошук – умовно назвемо його пошуком за ключовими словами – засновується на моделі запитів і моделі ресурсів. Потреба користувача повністю визначається набором ключових слів, які користувач задає при ініціюванні запиту. Попри можливе розв'язання синонімічних і інших лексичних зв'язків у системі, потреба користувача не інтерпретується у термінах онтологій. З погляду системи початковий запит, навіть збагачений лексичними зв'язками чи іншою інформацією (статистикою запитів, наприклад), залишається в рамках єдиної моделі запитів, що подає потребу користувача явно, без спроби «розуміння». Натомість семантичний пошук, зокрема базований на онтологіях, засновується на чотирьох інших моделях, які у [4] визначаються таким чином:

1) мислена (ментальна) модель $\mathcal{O}_U$: модель формалізує інформаційну потребу користувача на початку процедури інформаційного пошуку;

2) модель запитань користувача $\mathcal{Q}_U$: модель складається з елементів, які в свою чергу конструюються із мовних примітивів $\mathcal{P}_U$ мови користувача $\mathcal{L}_U$. Ця модель виражає елементи $\mathcal{O}_U$ у вигляді елементів $\mathcal{P}_U$ мови користувача $\mathcal{L}_U$;

3) модель системних ресурсів $\mathcal{O}_S$: ця модель складається із мовних примітивів $\mathcal{P}_S$ мови системи $\mathcal{L}_S$. Модель базується на знаннях, виражених у онтологіях, і на відміну від абстрактної моделі $\mathcal{O}_U$, має чітку і повністю доступну структуру. Елементи моделі формують базу знань пошукової системи;

4) модель запитів системи $\mathcal{Q}_S$: модель представляє фінальну оброблену версію запиту користувача як конструкцію з мовних примітивів $\mathcal{P}'_S$ мовою запитів системи $\mathcal{L}'_S$. Як правило, запити мовою системи виражаються за допомогою елементів онтологій системи.

Модель eTVSM є абстрактною відносно звичних сценаріїв роботи користувача із пошуковою системою. Модель визначає спосіб обрахунку схожості документів між собою, що зокрема передбачає інтерпретацію запиту користувача як документа і подальше зіставлення цього документа із документами в системі. Проте неважко помітити, що запит не вдасться інтерпретувати тим самим способом, що і документи в системі. Це зумовлено тим, що документи у системі зазвичай є цілісними текстами, а у нашому випадку із використанням ймовірнісних тематичних моделей – ми припускаємо, що ці тексти ще й базуються на породжувальних моделях мови. Натомість запити здебільшого представлено окремими фразами, ключовими словами, одиничними іменованими сутностями тощо. Інакше кажучи, природа документів і запитів є різною, а тому інтерпретація інформаційної потреби користувача у ОКІПС вимагає особливих підходів.

У традиційних пошукових системах, що базуються на звичайних словниках термінів та індексах і відповідних методах пошуку, проблема інтерпретації інформаційної потреби здебільшого обмежується переформулюванням та розширенням запитів [8]. По-перше, завдяки статистиці запитів система може пропонувати користувачу більш уживані та успішні запити. По-друге, за рахунок тезаурусів – ручних чи автоматичних – система може збагачувати запит синонімами чи іншими лексично спорідненими словами. Ці дії спрямовані на те, аби при фільтруванні документів за списками словопозицій в індексі отримати для ранжування якомога більшу множину результатів, яка буде впорядкована за відповідністю запиту.

У ОКІПС подібні прийоми не мають сенсу, оскільки модель запитів користувача і модель запитів системи суттєво відрізняються. Натомість користувацький запит має бути перетворений на онтологічно-орієнтований запит. У контексті обраної нами моделі онтології це означає представлення запиту через інтерпретації . Зауважимо, що інтерпретації Φ несуть основне семантичне навантаження у онтологічному просторі системи. Насправді інтерпретації $\phi_i \in \Phi$ є *атомарними інформаційними потребами*. Це означає, що у найпростішому випадку потреба користувача виражена деякою підмножиною елементів онтології (тем, сутностей), поєднаних у різних пропорціях. З практичного погляду як інформаційне наповнення цієї потреби виступає частина певного документа з колекції. Певна річ, у більшості випадків користувач має *складну інформаційну потребу*, що складається з атомарних у різних пропорціях, а інформаційно представлена документами із різним ступенем релевантності (визначеним рангом документа у результатах пошуку).

Ранжування результатів

Для обчислення рангу у ОКІПС може бути застосовано кілька підходів. Окрім стандартного ранжування методом обчислення скалярних добутків векторів запиту та документа, ми звертаємо увагу на два особливі методи. Один базується на кластеризації документів, інший – на попередньо обрахованих *TOP-k* документах за рангом.

Кластеризація у ІП грає особливу роль і базується на *кластерній гіпотезі* (cluster hypothesis): документи в межах кластера поводяться однаково відносно релевантності інформаційній потребі [8]. Якщо певний документ у кластері релевантний запиту, то ймовірно, що й інші документи релевантні цьому ж запиту. Кластеризація може бути плоскою та ієрархічною. Плоска кластеризація групує об'єкти у кластери, які жодним чином не пов'язані між собою. Ієрархічна кластеризація, як виходить із назви, створює ієрархію кластерів. Окрім того, кластеризація може бути жорсткою та м'якою, що впливає на однозначність віднесення об'єкта до того чи іншого кластера. При жорсткій кластеризації кожний об'єкт відноситься строго до одного кластера. При м'якій кластеризації прив'язка об'єкта поширюється між декількома кластерами, отже

об'єкт частково присутній у кожному кластері. Прикладом алгоритму м'якої кластеризації є прихований семантичний аналіз [5].

Якщо розглядати жорстку плоску кластеризацію, то формалізувати проблему можна таким чином. Дано множину документів D та кількість кластерів K (кардинальність), а також деяку цільову функцію, яка оцінює якість кластеризації. Необхідно знайти таке відображення $\kappa : D \to \{1..K\}$, яке б максимізувало (мінімізувало) цільову функцію. Як правило також необхідно, аби відображення κ було сюр'єкцією, тобто жоден кластер не був порожнім. Цільова функція може бути визначена через схожість документів, зокрема у нашому випадку – через її онтологічно-орієнтовану трактовку.

Метод кластеризації *K-середніх* (K-means clustering) є одним із основних методів плоскої кластеризації [8]. У методі *K*-середніх кожна точка, що підлягає кластеризації, виражається як вектор у деякому *S*-вимірному просторі, $\left\{x^{(n)}\right\}, де\ n = |D|, D - множина\ точок$. Якщо припустити, що простір над полем дійсних чисел, то можна визначити деяку метрику, що визначає відстань між точками, у загальному вигляді: $d(x,y) = \frac{1}{2}\sum_{i \in S}(x_i - y_i)^2$ [7]. Якщо розглядати документи у колекції, то ціль методу *K*-середніх – мінімізувати вказану відстань між документом $\vec{x}$ та *центроїдом* $\vec{\zeta}$ відповідного кластера λ_k. Формально цільова функція визначається так:

$$f_{obj}(\mathrm{K}) = \sum_{k \in K} f(k);$$

$$f(k) = \sum_{\vec{x} \in \lambda_k} \left|\vec{x} - \vec{\zeta}(k)\right|^2, \text{ де } \vec{\zeta}(k) = \frac{1}{|\lambda_k|}\sum_{\vec{x} \in \lambda_k} \vec{x}.$$

Алгоритм *K*-середніх є ітеративним алгоритмом з двокроковою ітерацією [7]. При ініціалізації центроїди обираються довільним чином, присвоюючи кожному центроїду випадковим чином обраний документ. Ітерація складається з двох кроків – присвоєння та перерахунку. На першому кроці документи присвоюються кластерам із найближчими центроїдами; на другому кроці центроїди обраховуються із урахуванням змін у кластері. Критерієм зупинки може бути відсутність змін центроїдів [8].

Метод із *TOP-k* документів за рангом є значно простішою альтернативою кластеризації (хоча може бути трактований як її варіант). Він полягає у попередньому знаходженні *k* найбільш схожих документів для кожного документа в колекції. Ідея використання в ОКІПС надбудови у вигляді кластеризації чи методу *TOP-k* документів за рангом полягає у посиленні простого обрахунку спорідненості «на вимогу» додатковою інформацією про релевантність документів. Для найвищих позицій у пошуковій видачі кластери повинні так чи інакше містити певний спільний набір документів, що було б ілюстрацією цілісної концепції релевантності у системі. Щоправда, інтеграцію цих механізмів як надбудови над моделлю eTVSM слід досліджувати окремо, із аналізом конкретних методів обчислень та трактовки результатів.

Фільтрація, індекс та попередні обчислення

Ми порушували питання онтологічно-орієнтованої індексації в загальному вигляді у [2]. З оглядом на запропоновану нами онтологію постає ряд додаткових питань, пов'язаних із адаптацією процесів пошуку за індексом під онтологічно-орієнтовані сценарії використання. Традиційний пошук за класичним індексом передбачає фільтрацію документів за ключовими словами запиту та списками словопозицій в індексі. У ОКІПС це неможливо через принципову відсутність списків словопозицій і недоцільність фільтрування документів за ключовими словами. Проте обрахунок скалярних добутків між вектором запиту та усіма векторами документів при кожному запиті користувача є теж неможливим.

Варіантом адаптації індексу під онтолого-керований сценарій пошуку є побудова *інвертованої тематичної карти* за ймовірнісними тематичними моделями. Маючи інформацію про розподіл на темах для кожного документа можна побудувати структуру, аналогічну звичайному інвертованому індексу, де замість словника – список тем, а замість списків словопозицій – списки документів із пропорціями даної теми в них. Тоді за інтерпретацією запиту користувача можна здійснювати фільтрацію документів за темами, а отриману вибірку ефективніше опрацьовувати при ранжуванні результатів та формуванні пошукової видачі.

У розглянутій нами моделі ОКІПС, окрім власне індексації, є ще потреба у двох попередньо обчислюваних групах об'єктів. По-перше, це моделі документів. Інтерпретація документів є процесом, що може бути вільно виокремлений у відповідну фазу. По-друге, це згадана вище кластеризація та/або попереднє обчислення *TOP-k* документів за рангом. Слід дбати про те, аби ці процеси не перешкоджали перебігу поточних процесів.

Висновки

У статті розглянуто інтеграцію тематично-сутнісної онтології покращеної тематичної векторної моделі в інфраструктуру інформаційно-пошукової системи. Акцент у роботі зроблено на практичні аспекти адаптації стандартної пошукової інфраструктури під онтолого-керований сценарій.

Ключовим аспектом інтеграції онтології у загальносистемний контекст є процес побудови моделі документа. Запропоновано способи видобування зв'язків інтерпретацій із темами та відображення змісту документів у термінах онтології моделі.

Розглянуто також проблему інтерпретації інформаційної потреби користувача та представлено шляхи її вирішення.

Окрім цього, запропоновано варіанти адаптації механізмів ранжування та трактовки результатів через використання кластеризації документів. Також надано способи підвищення ефективності системи за рахунок переглянутих процесів індексації та додаткових попередніх обчислень.

Подальші дослідження варто зосередити на вивченні методів та засобів впровадження запропонованих рішень, зокрема на конкретних підходах до формування пошукової видачі. Цікавим бачиться детальніше дослідження інтерпретації інформаційної потреби користувачів.

Список літератури

1. Глибовец Н. Н. Применение онтологий и методов текстового анализа при создании интеллектуальных поисковых систем / Н. Н. Глибовец, А. Н. Глибовец, А. С. Шабинский // Проблемы управления и информатики. – 2011. – № 6. – С. 96–102.
2. Шабінський А. С. Один підхід до індексації у онтологічно-керованій інформаційно-пошуковій системі / Д. Б. Буй та ін. // Матеріали десятої міжнародної науково-практичної конференції TAAPSD'2013 (Ялта, 25 травня – 2 червня 2013 р.). – Кіровоград : ПП Центр оперативної поліграфії «Авангард», 2013. – 196 с.
3. Шабінський А. С. Онтології, ймовірнісні тематичні моделі та тематичні карти / А. С. Шабінський // Наукові записки НаУКМА. – 2013. – Т. 151: Комп'ютерні науки. – С. 60–65.
4. Ontology-based Interpretation of Keywords for Semantic Search / T. Tran, P. Cimiano, S. Rudolph, R. Studer // Proceedings of the 6th International Semantic Web Conference. – Busan, Korea, 2007.
5. Indexing by Latent Semantic Analysis / S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman // Journal of the American Society for Information Science. – 1990. – № 41.
6. Kuropka D. Modelle zur Repräsentation natürlichsprachlicher Dokumente / D. Kuropka. – Berlin : Logos Verlag, 2003. – 253 p.
7. MacKay D. J. Information Theory, Inference and Learning Algorithms / D. J. MacKay. – Cambridge University Press, 2003. – 411 p.
8. Manning C. D. Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schütze. – New York : Cambridge University Press, 2008. – 348 p.

A. Shabinskiy

INTEGRATION OF TOPIC-ENTITY ONTOLOGY INTO THE INFRASTRUCTURE OF INFORMATION RETRIEVAL SYSTEM

In this paper we consider topic-entity ontology of the enhanced topic-based vector-space model in the context of the common information retrieval infrastructure. Adaptation of the information retrieval processes to the ontology-driven scenario is proposed. Models and approaches for the implementation of key components are proposed considering practical aspects and peculiarities of the suggested ontology.

Keywords: indexing, information retrieval, ontology, ranking, topic-based vector-space model.

Матеріал надійшов 15.05.2014