УДК 519.7:378.147

МЕТОД ФОРМУВАННЯ ОНТОЛОГІЧНОГО НАПОВНЕННЯ НА ОСНОВІ АНАЛІЗУ ЗАШУМЛЕНОЇ СЛАБКОСТРУКТУРОВАНОЇ ІНФОРМАЦІЇ СПЕЦІАЛІЗОВАНИХ ВЕБ-САЙТІВ

Н.Р. Пасічник, М.П. Дивак

Тернопільський національний економічний університет,

natalia.pasichnyk@gmail.com, mdy@tneu.edu.ua

У статті запропоновано метод автоматизованого формування та підтримки соціально значимих онтологій, який уможливлює зменшення завантаженості експертів. Створено алгоритм формування онтологічного наповнення на основі структурного аналізу наповнення тематичних Веб-сторінок із відсівом неключових термінів за частотним принципом.
Ключові слова: Веб-сайт, онтологійне наповнення, структурний аналіз, частотний відсів.

The article offers a method of automated generation and support of socially meaningful ontologies, which makes the decrease of experts' overload possible. The algorithm of generation of ontological content is created basing on the structural analysis of the thematic Webpage's content, with filtration of non-key terms basing on the frequency principle.
Keywords: Web-site,ontological content, structural analysis, frequency-based filtration.

В статье предложен метод автоматизированного формирования и поддержки социально значимых онтологий, который обеспечивает уменшение загруженности экспертов. Создан алгоритм формирования онтологического наполнения на базе структурного анализа наполнения тематических Веб-страниц с отсевом неключевых терминов за чатотным принципом
Ключевые слова: Веб-сайт, онтологическое наполнение, структурный анализ, частотный отсев.

Вступ

Одним із можливих шляхів зниження витрат на розробку та підтримку тематичних Веб-сайтів є створення методів, алгоритмів та засобів для автоматизованого генерування його структури та інформаційного наповнення. Питання автоматизації процесів генерування структури сайту розглянуто в роботі [1], а задачі автоматизації наповнення сайту потребують подальшої розробки. Виготовлення Веб-сайту розпочинається із проектування його структури, яку доцільно розбивати на типову семантичну частину та частину, що подає особливості об'єкта, який репрезентується даним сайтом. Формування першої із згаданих частин можна формалізувати [1]. Для підвищення інтересу Веб-спільноти до окремого сайту, його інформаційне наповнення повинно відповідати критеріям актуальності та унікальності. Саме вимога унікальності ускладнює формалізацію методів побудови наповнення Веб-сайту. Очевидно, що вона може бути застосована лише до окремих аспектів такої побудови.

Серед типів сторінок Веб-сайту особливе місце займають он-лайн сервіси, які за рахунок динамічності свого наповнення забезпечують значне підвищення числа можливих користувачів. Такий підхід є одним із базових технології Веб 2.0 [2-3]. В окрему групу виділимо сервіси, які формують певні тематичні онтології, тим самим розбудовуючи Семантичний Веб. В залежності від того, яку інформацію вони обробляють (структурну із формалізованих меню або виділену іншими методами інформацію Веб-сторінок), ми поділяємо їх на сервіси синтезу структур та понять. Формалізація методу синтезу структур подана в роботі [1]. Розробці методів структурування понять, а також поповнення таких онтологій понять присвячена дана робота.

Понятійні онтології ефективно використовують для аналізу формалізованих характеристик понять, що зазнають динамічних структурних змін. Структуру загальних понять визначають через систему пов'язаних часткових понять із урахуванням багатомовності та синонімічності їх можливих мовних реалізацій. Після розробки онтологічних структур можна здійснювати їх наповнення та використання. Однак створення онтологій на основі лише експертних суджень є достатньо трудомістким процесом, із важко контрольованою прийнятністю для окремих користувацьких спільнот та суб'єктивізмом у проведенні реструктуризації цих онтологій. Поряд із цим, зовсім виключити втручання експертів можна лише у випадку, коли варіанти онтологічних структур уже реалізовані і потребують лише систематизації та узагальнення. Така ситуація зустрічається при аналізі структур сайтів, але для формалізації змісту понять на основі Веб-контенту вона не актуальна. Тому необхідно розробити метод формування онтологічного наповнення із мінімізацією зусиль залучених експертів.

Таким чином, реалізація завдання даної роботи передбачає розробку структури понятійної онтології, методу та алгоритму формування її наповнення, а також проведення чисельних експериментів, що буде розглянуто в ході подальшого викладу.

1. Основні положення методу формування онтологічного наповнення на основі слабко структурованої інформації

Пропонований метод ґрунтується на наступних основних положеннях. Отримати актуальні соціально-значимі характеристики предметних областей, що динамічно розвиваються, можна за допомогою аналізу інформації, представленої у Вебі. Ця інформація отримується зі спеціалізованих сайтів, які здійснюють моніторинг запитів на відповідні види продукції або послуг. Згадані запити містять слабко структуровану та неструктуровану інформацію. Слабко структурована інформація міститься у списках, елементи яких структуруються довільним чином. Характеристики предметних областей, що містяться у слабко структурованих списках можуть бути виявлені на основі частотного аналізу і використані для формування наповнення предметних

онтологій. Базою для формування структури онтологій, а також метаонтологійних понять служать Веб-анотації характеристик предметних областей. Структурування онтології відбувається експертним шляхом, із мінімізацією завантаженості та кваліфікаційних вимог до експерта.

Під онтологіями розуміємо деревоподібну систему соціально-значимих понять (концептів) певної предметної області. Сюди включаються як концепти, що використовуються в робочих документах предметної області, так і поняття, що їх узагальнюють, тобто мета-поняття. Тому онтологію моделюємо наступною деревоподібною структурою

$$O\_Str =< IdCn, \Pr Cn, Meta > \quad (1)$$

яка включає ідентифікатори $IdCn$ концептів, ідентифікатор $\Pr Cn$ батьківського концепту та атрибут $Meta$, який дозволяє відділити робочі поняття предметної області від її мета-понять. При цьому загальне поняття відрізняється від часткових відсутністю батьківського $(\Pr Cn = NULL)$. Кожне із понять допускає різні лінгвістичні представлення у вигляді словосполучень O_Phr. Слова, що визначають онтологічні поняття представляються в конкретних словоформах O_Frm а також своїми основами O_Bs. Словоформи використовуються представлення понять користувачам, а основи – для автоматичної ідентифікації еквівалентності мовних представлень. Атрибути введених понять згруповано в наступні структури:

$$O\_Bs =< IdLg, IdBs, WBase >, \quad (2)$$

$$O\_Frm =< IdLg, IdFm, IdBs, WForm >, \quad (3)$$

$$O\_Phr =< IdCn, IdLg, IdPh, IdBs, IdFm, Id \Pr Bs >, \quad (4)$$

де $IdLg$ ідентифікатор мови реалізації, $IdBs$ ідентифікатор основи слова, $WBase$ основа слова, $IdFm$ ідентифікатор форми слова, $WForm$ форма слова, $IdCn$ ідентифікатор концепта, $IdPh$ ідентифікатор фрази, $Id \Pr Bs$ ідентифікатор батьківської основи поняття.

2. Метод формування онтологічного наповнення

Для формування онтологічного наповнення, значимого для певного сегменту Веб-аудиторії, зручно використати описи, представлені на відповідних Веб-сторінках. Для підвищення значимості такої інформації для аналізу необхідно експертним шляхом відбирати лише певні спеціалізовані Веб-сайти. Наповнення сторінок таких сайтів формується для сприйняття користувачами, а тому не є строго структурованим за певними жорсткими правилами. Окрім того на цих сторінках розташовано багато додаткової інформації, яка з точки зору онтологічного наповнення може розглядатися як шум. Варто вимагати також, щоб інформація на Веб-сторінках була

структурована, а не просто розбита на параграфи чи абзаци. Така вимога дозволяє значно звужувати сферу пошуку, тим самим піднімаючи його ефективність. В даному випадку під структурованістю мається на увазі оформлення інформації у вигляді спискових структур.

Для відбору структурованої інформації, що стосується даної предметної області, формуємо множину ключових термінів KWS, що її характеризує. Для підтримки аналізу вмістимого Веб-сторінок створено наступну допоміжну структуру AS

$$AS =< IdPg, IdLst, IdIt, IdBs, IdFm, Id \Pr Bs >, \quad (5)$$

де $IdPg$ - ідентифікатор аналізованої Веб-сторінки, $IdLst$ - ідентифікатор списку сторінки, $IdIt$ - ідентифікатор елемента списку.

Повторюваність мовного представлення є його важливою характеристикою, що дозволяє відділити значимі представлення від несуттєвої інформації даної предметної області. Для її контролю вводиться наступна структура BF частот основ

$$BF =< IdBs, BsFr, IdLPg, Phn > \quad (6)$$

де $IdBs$ - ідентифікатор основи, $BsFr$ - частота появи основи на різних Веб-сторінках, $IdLPg$ - ідентифікатор останньої із Веб-сторінок, де зустрічалася основа, Phn - маркер фоновості поняття, що приймає значення невизначеності $NULL$ за замовчуванням.

При аналізі $HTML$ коду чергової Веб-сторінки спеціалізованого Веб-сайту встановлюємо її ідентифікатор

$$CurPgId := \max\left(\pi_{IdLPg}(BF)\right) + 1. \quad (7)$$

Далі виділяємо елементи $LSTIt$ її спискових структур, що наповнюють теги $< li >$. Елементи списку розбиваються на елементарні поняття It із використанням роздільників, які утворюють спеціальну множину сепараторів:

До слів елементарного поняття застосовуємо процедуру BsC побудови їх основ. Процедура здійснює відкидання від слів типових закінчень. При умові, що

$$BsC(It_{Pg,Lst}) \subset BsC(KWS) \quad (8)$$

всі елементарні поняття аналізованого списку із відповідними атрибутами включаються в структуру AS. Словоформи вибирають безпосередньо із елемента списку $It_{Pg,Lst}$ і розпізнають за допомогою відношення форм O_Frm або поповнюють його.

Нехай $Wrd_k(It_{Pg,Lst})$ - деяке k – те слово, виділене із елемента списку

$It_{Pg,Lst}$. Якщо основа даного слова зареєстрована, то ідентифікатор його основи $IdBsWrd$ визначається із відношення O_Bs:

$$IdBsWrd = \pi_{IdBs}\left(\sigma_{WBase=BsC(Wrd_k(It_{Pg,Lst}))}\left(O\_Bs\right)\right). \quad (9)$$

Якщо основа зареєстрована у відношенні BF і номер поточної сторінки не співпадає із номером останньої врахованої, то індекс частоти $BsFr$ збільшується на 1, а номер поточної сторінки заноситься в поле $IdLPg$

$$\left(Count\left(\pi_{IdBs}\left(\sigma_{IdBs=IdBsWrd\ AND\ IdLPg<>CurPgId}\left(BF\right)\right)\right)\neq 0\right)\Rightarrow \\ \Rightarrow \left(BF.BsFr := BF.BsFr+1\right)\wedge\left(BF.IdLPg := CurPgId\right) \quad (10)$$

При такому підході забезпечується облік частот використання основи на різних Веб-сторінках. Якщо основа у відношенні BF не знайдена, вона заноситься у відношення частот основ із індексом частоти рівним 1 та номером поточної сторінки

$$\left(Count\left(\pi_{IdBs}\left(\sigma_{IdBs=IdBsWrd}\left(BF\right)\right)\right)=0\right)\Rightarrow \\ \Rightarrow\left(BF.IdBs := IdBsWrd\right)\wedge\left(BF.BsFr := 1\right)\wedge\left(BF.IdLPg := CurPgId\right). \quad (10)$$

Якщо основа слова $Wrd_k(It_{Pg,Lst})$ не розпізнана, вона вноситься в список основ, саме слово вноситься в список словоформ, а поповнення відношення частот основ здійснюється згідно співвідношення (10).

Для включення в онтологію експерту пропонуються лише основи, частота яких буде перевищує деяке мінімальне значення $BF0 \geq 2$, яке вибирається користувачем. Експертові пропонується список основ для включення в онтологію, коли вищезгаданій умові задовольняють не менше $BC0$ основ

$Count\left(\pi_{IdBs}\left(\sigma_{BsFr>=BF0}\left(BF\right)\right)\right) > BC0$.

Для прийняття адекватного рішення основи пропонуються в тому контексті, в якому вони зустрічаються на Веб-сторінках. Це дає змогу виділяти поняття, які складаються із кількох слів а також не пропонувати повторно основи, які не вибрані експертом для включення в онтологію при аналізі попередніх контекстів. Основи, які ввійшли в онтологію отримують фоновий індекс $Phn := 2$, щоб повторно не подаватися для аналізу. Основи, які не були вибрані жодного разу не можуть складати основи контекстів, вони помічаються як фонові $Phn := 1$.

Після вибору елементів онтологічного наповнення, вони будуть включені в онтологічну ієрархію. Таке включення робиться експертом, який закладає свої знання в онтологію. Щоб зробити його вибір соціально-значимим та мінімізувати суб'єктивізм оцінювання, використовується механізм автоматизованого анотування понять, який підтримується інформаційною структурою *ConcAnot*

$$ConcAnot = \langle IdCn, AnURL, AnTitle, Anot, Desc \rangle. \quad (11)$$

де $IdCn$ ідентифікатор концепту, $AnURL$ адреса Веб-сторінки анотації, $AnTitle$ тег title Веб-сторінки анотації, $Desc$ опис концепту.

Пошук анотації здійснюється на основі запиту до пошукового сервера, який включає перелік основ даного концепта. Із списку анотацій, виданого пошуковим сервером, вибирається сторінка, тег title якої містить слова із переліку словоформ. В анотацію включаються не більше 3-х абзаців із відібраної сторінки, які містять елементи запиту. Експерт здійснює перегляд множини відібраних концептів із їх анотаціями. Він може редагувати анотації а також отримувати нову анотацію, або переглядати всю сторінку, що містить анотацію.

На основі анотацій експерт формує опис поняття, відбирає підмножини вкладених понять за принципом ”частина-ціле”, впорядкувати однорівневі поняття за принципом ”від загального до конкретного” та ”від простого до складного”. На основі термінів анотації експерт групує поняття та вводить метапоняття, що їх об’єднують. При великій кількості метапонять формуються метапоняття вищих порядків.

Представлені вище алгоритм та метод демонструють достатньо громіздкі підходи до переробки інформації, яка проте не виключає і активної роботи експерта. Представляє інтерес оціннка вливу згаданих підходів на ефективність роботи експерта. В якості критерію ефективності таких оцінок виберемо відношення кількості відібраних термінів NOI до кількості переглянутих стрічок NSE:

$$EE = \frac{NOI}{NSE} \quad (12)$$

3. Алгоритм формування онтологічного наповнення

На основі наведених теоретичних положень сформуємо алгоритм автоматизованого формування онтологічного наповнення:

1. Встановлюємо перелік релевантних спеціалізованих сайтів StL а також множину KWS ключових слів мінімальної потужності, які характеризують найважливішу особливість предметної області.
2. Будуємо запит, що включає слова із множини KWS до кожного сайту із множини StL та формуємо множини $HTML$ кодів веб-сторінок .
3. Якщо сторінка містить список, хоча б один елемент якого містить слова із множини KWS, або, що описують один із термінів побудованої онтології, то елементи списку заносяться у відношення AS. При цьому вони розбиваються на елементарні терміни за допомогою роздільників

із множини SS, а елементарні терміни розбиваються на слова. Основи відібраних слів заносяться у відношення BF, а якщо вони вже там зареєстровані із сторінки, що не співпадає із поточною і також не належать до фону, їх кратність збільшується на 1 та оновлюється ідентифікатор сторінки реєстрації.

4. Якщо при зміні кратність основи перевищить деяке наперед задане значення $BF0$, кількість кандидатів на включення в онтологію OMC збільшується на 1. Якщо $OMC \geq BC0$, то контекст кандидатів на включення в онтологію подається експертові.
5. Для кожного входження основи-кандидата в відношення AS вибираються словоформи, що формують елемент списку, який містить дану основу. Атрибут Phn основи-кандидата у відношенні BF повинен бути невизначеним .
6. Після сформування контексту він подається експертові для аналізу. Після відбору експертом елементів для онтології, елементи списку, жодна компонента якого не була відібрана, помічаються значенням атрибуту $Phn := 1$ у відношенні BF для виключення їх повторної подачі в контексті іншого терміна.
7. Відібрані для онтології концепти поміщаються в її структуру на основі згенерованих анотацій. Перехід на пункт 2.

4. Чисельні експерименти

На основі запропонованого методу досліджено перші стадії процесу побудови онтології кваліфікаційних вимог до Веб-програміста, який спеціалізується на PHP програмуванні. Цю діяльність можна розглядати як надання високо технологічних програмістських послуг, особливості виконання яких можна описати онтологією поняття “PHP програміст” (“PHP programmer”). Для побудови онтології, значимої для софтверних українських компаній вибрано множину сайтів, що спеціалізуються на пропозиціях вакантних посад на підприємствах України, зокрема ”rabota.ua”, “jobs.ua”, “work.ua” і містять спеціальні розділи вакансій в сфері IT. Серед цих сайтів для проведення перших експериментів вибрано сайт ”rabota.ua” та множину ключових слів, яка складається з єдиного елемента $KWS = \{"PHP"\}$.

Серед 20 перших сторінок, що описують вакансії по даному запиту лише 10 містили спискові структури із входженням ключового слова ”PHP”. Фрагмент сторінки однієї із вакансій наведено на рис. 1. В даній сторінці вимоги до кандидата на заміщення посади сформовані у вигляді списку.

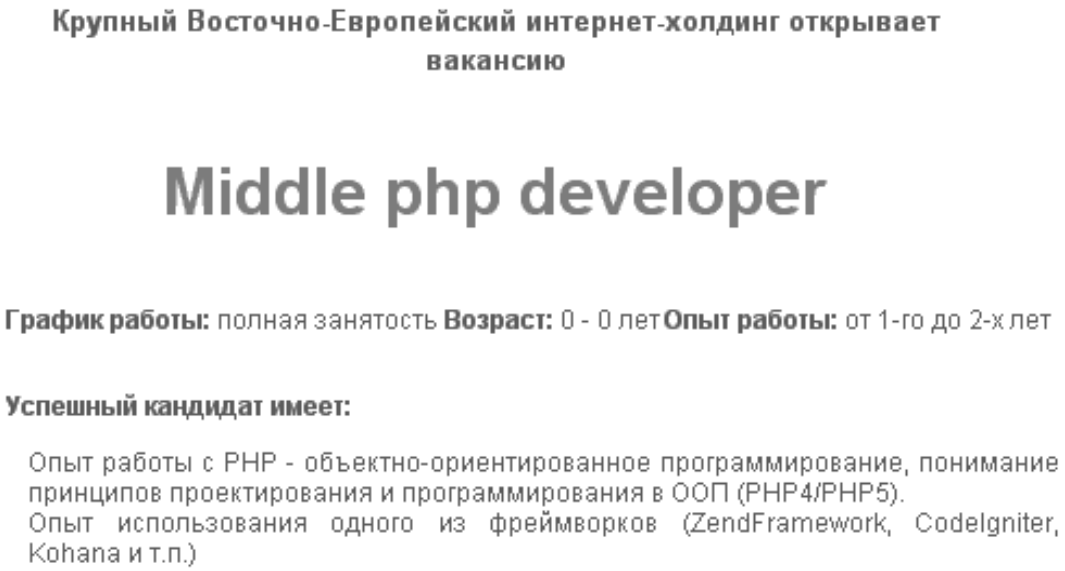

Рис. 1. Фрагмент сторінки однієї із вакансій

На рисунку 2 представлено фрагмент відношення *AS* змодельованого в електронних таблицях. В даній моделі для наочності замість посилання на батьківські елементи багатослівного терміну просто нумеруються, основи представлені своїми ідентифікаторами в стовпчику *IdWrd* , а словоформи подані безпосередньо в стовпчику *Wrd* . На рисунку 3 наведено фрагмент списку термінів, який пропонувався для включення в онтологію разом із відповідними контекстами. Маркер "+" позначає, що вони були дійсно включені в онтологію.

IdPg	IdLst	IdIt	IdWrd	Wrd	
2	1	1	75	Опыт	1
			76	работы	2
			4	PHP	3
		2	77	объектно	1
			78	ориентированное	2
			79	программирование	3
		3	80	понимание	1
			81	принципов	2
			82	проектирования	3
		4	83	программирования	1
			48	ООП	2
		5	4	PHP4	
		6	4	PHP5	

Рис.2. Модель фрагменту відношення AS засобами електронних таблиць

Php +				
	Strong	knowledge	PHP	
	Опыт	работы	с	PHP
	PHP4			
	Опыт	разработки	на	PHP5
	хорошее	знание	PHP	
	их	реализации	в	php5.3
MySql +				
	Strong	knowledge	MySql	

Рис.3. Фрагмент списку термінів із контекстами для їх включення в онтологію

Зразок анотації першого онтологічного поняття подано на рисунку 4.

HTML
HTML (від англ. Hypertext Markup Language - мова розмітки гіпертексту) - це стандартна мова розмітки документів у Всесвітній павутині. Всі веб-сторінки створюються за допомогою мови HTML (або XHTML). Мова HTML інтерпретується браузером і відображається у вигляді документа, зручному для людини. HTML є додатком SGML (стандартної узагальненої мови розмітки) і відповідає міжнародному стандарту ISO 8879. HTML-документ є текстовим файлом розмічений за допомогою спеціальних (текстових) команд. Текстовий формат представлення веб-документів був вибраний виходячи з основних вимог до веб-документу: простота, можливість безпосередньої інтерпретації в будь-якій операційній системі, мінімальний розмір файлу, зручність редагування і інтерпретації. Мова розмітки гіпертекстових документів HTML дозволяє визначити різні типи елементів (у оригіналі element), що забезпечують функціональність документа: текстові фрагменти із заданими параметрами форматування, списки, таблиці, зображення, гіперпосилання і т.д. Елементи HTML оголошуються за допомогою команд розмітки, званих тегами (від англійського tag - ярлик). Всі HTML-теги, що зустрічаються в тексті документа інтерпретуються браузером при відображенні документа.

Рис.4. Зразок анотації онтологічного поняття

Сама онтологія, побудована на основі автоматизованого аналізу п’яти Веб-сторінок, подана на рисунку 5, де метапоняття відображені курсивом. Як бачимо, навіть при аналізі незначного числа слабоформалізованих вимог онтологія включає базові напрямки аналізованої спеціалізації. Для їх позначення експертом введено відповідні мета терміни. Окрім 9 термінів онтології також відібрано 8 фонових термінів, які марковані за допомогою атрибуту *Phn* відношення *BF* .

Онтологія поняття

- ***PHP програміст***
  - ***Технології опису зовнішнього виду документа***
    - **HTML** - мова розмітки документа
    - **CSS** - технологія опису зовнішнього виду документа
    - **DHTML** - стандарт роботи з HTML документом, як з об'єктною структурою
  - ***Генерація HTML сторінок на стороні сервера***
    - **OOP** - парадигма програмування, яка розглядає програму як множину об'єктів, що взаємодіють між собою
    - **PHP** - скриптова мова програмування HTML сторінок на стороні сервера
  - ***Сервери баз даних***
    - **MySQL** - компактний багатопотічний сервер баз даних
  - ***Програмування Веб-броузерів***
    - **JavaScript** - мова програмування Веб-броузерів для маніпулювання HTML елементами
      - **Ajax** - технологія Javascript асинхронного обміну даними між броузером і сервером
  - ***Управління проектами***
    - **GIT** - система контролю версій розробки проекту

Рис.5. Онтологія, побудована на основі автоматизованого аналізу п’яти Веб-сторінок

Всього експертові для перегляду із врахуванням контексту було подано 23 стрічки, з яких відібрано 9 понять. При перегляді повних текстів 5 аналізованих сторінок експерт повинен був би переглянути біля 200 стрічок. Тобто вдалося принаймі на порядок зменшити завантаженість експерта і частково зняти інформаційну зашумленість даних.

5. **Висновки**

У статті розглянуто один із можливих шляхів формування онтологій шляхом аналізу зашумленої слабо структурованої інформації спеціалізованих Веб-сайтів. В основу автоматизованого методу формування онтологійного наповнення покладено структурний аналіз тематичних сторінок спеціалізованих Веб-сайтів відсів фонових термінів за частотним критерієм та залучення експерта для остаточного відбору термінів та структурування онтології.

У результаті проведених досліджень отримано такі наукові та практичні результати. Вперше запропоновано формування метод формування онтологій шляхом аналізу зашумленої слабо структурованої інформації тематичних Веб-сторінок спеціалізованих Веб-сайтів. Це уможливлює формалізацію процедури побудови та підтримки онтологій вимог до високотехнологічних продуктів та послуг, значимих для певних сегментів Веб-спільноти. Ефективність запропонованого методу та алгоритму підтверджено при аналізі початкового етапу структурування онтології “PHP програміст”, значимої для працедавців софтверних компаній України.

Література

1. Пасічник Н.Р., Дивак М.П. Формалізм в постановці задачі створення якісного сайту. // Наукові праці ДонНТУ. Серія „Інформатика, кібернетика та обчислювальна техніка.- 2011. Вип 14 (188).- С.325-329.

2. Глибовец Н. Н., Шыпович Л. О. Становление технологии WEW 3.0. [Електрон. ресурс]. - Режим доступу: http:// dspace. nbuv.gov. ua/ dspace/ handle/ 123456789/ 18769

3. Анатольев А. Г. Перспекивы развития Веб-технологий. [Електрон. ресурс]. - Режим доступу: www.4stud.info/web-programming/lecture9.html