УДК 519.7: 378.147

Н.Р. ПАСІЧНИК
Тернопільський національний економічний університет

МЕТОД ФОРМУВАННЯ ОНТОЛОГІЙНОГО КОНТЕНТУ НА ОСНОВІ АНАЛІЗУ ІНФОРМАЦІЇ СПЕЦІАЛІЗОВАНИХ ВЕБ-САЙТІВ

У статті запропоновано метод автоматизованого формування та підтримки соціально значимих онтологій, який уможливлює зменшення завантаженості експертів. Створено алгоритм формування онтологійного наповнення на основі структурного аналізу наповнення тематичних веб-сторінок з відсівом не ключових термінів за частотним принципом.

The article offers method of the automated generation and support of socially meaningful ontologies, which makes the decrease of experts' overload possible. The algorithm of generation of ontologic 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.

Ключові слова: спеціалізований Веб-сайт, онтологійне наповнення, структурний аналіз, частотний відсів.

Вступ

Розроблення веб-сайту розпочинається із проектування його структури, яку доцільно розбивати на типову семантичну компоненту та компоненту, що подає особливості об'єкта, що репрезентується даним сайтом. Формування першої зі згаданих компонент, оскільки вона є спільною для цілої множини тематичних сайтів, можна формалізувати [1].

Для підвищення інтересу веб-спільноти до окремого сайту його інформаційне наповнення повинно відповідати критеріям актуальності та унікальності. Виходячи із аналізу основних типів наповнення сайтів, серед його компонент відзначимо інформаційні ресурси, які складали основу технології веб 1.0 [2, 3]. Серед цих ресурсів виділяємо публікації редакторів програмного продукту, які повинні представляти актуальну інформацію як про об'єкт, який представляє веб-сайт, так і про пов'язані з ним питання, що цікавлять певний сегмент веб-спільноти. Створення такого контенту вимагає копітких зусиль розробників, актуальність роботи яких швидко знижується.

В цьому плані більш продуктивним є підхід із застосуванням он-лайн-сервісів, які надають користувачам різноманітні інформаційні послуги, як у здійсненні певних непростих функцій при роботі з веб, так і при аналізі тематичного наповнення веб-простору. Такий підхід є одним із базових технології веб 2.0 [2, 3]. Функціональні сервіси спрощують виконання певних Web-функцій своїм користувачам, а синтезуючі сервіси наповнюють слабо структуровані актуальні поняття змістом, що встановлюється на основі інформації, представленої на тематичних Веб-сайтах. Залежно 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. Якщо статичні статистичні сервіси можна формувати на основі сформованих баз даних, то синтезуючі сервіси працюють з масивами даних, що динамічно міняють свою структуру. Для їх організації необхідно створювати та поновлювати тематичні онтології. Тому розробка таких сервісів здійснюється в рамках технологій Semantic Web, тобто технологій, об'єднаних умовною назвою Веб 3.0. Розробці методів наповнення таких онтологій присвячена дана робота.

Постановка задачі

Понятійні онтології ефективно використовуються для аналізу формалізованих характеристик понять, що зазнають динамічних структурних змін. Структура загальних понять визначається через систему пов'язаних часткових понять з врахуванням багатомовності та синонімічності можливих мовних реалізацій. Тому необхідно спроектувати структури даних, які дозволяли б створювати, зберігати та модифіковувати онтології.

Після розробки онтологічних структур можна здійснювати їх наповнення та використання. Однак створення онтологій на основі лише експертних суджень є достатньо трудомістким з важко контрольованою прийнятністю для окремих користувацьких спільнот та суб'єктивізмом у проведенні їх реструктуризації. Поряд із цим зовсім виключити втручання експертів можна лише у випадку, коли варіанти онтологійних 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 з мінімізацією зусиль залучених експертів. Перевірка ефективності розробленого методу вимагає його програмної реалізації, а отже і формалізацію відповідного алгоритму.

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.

Структура понятійних онтологій

При розгляді деталізації загального поняття отримуємо деревоподібну структуру. При мовній реалізації кожного із понять цієї структури знову отримуємо дерево, гілками якого служать синонімічні слова або фрази, які описують поняття. Такі зв'язки зручно описувати за допомогою наступної сукупності

кортежів

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

$$O\_Phr =< IdCn, IdLg, IdPh, IdBs, IdFm, Id \Pr Bs >, \quad (2)$$

$$O\_Bs =< IdLg, IdBs, WBase >, \quad (3)$$

$$O\_Frm =< IdLg, IdFm, WForm > . \quad (4)$$

Перший з кортежів визначає деревоподібну структуру загального поняття, деталізуючи його на основі ідентифікаторів часткових $IdCn$ понять та визначаючи зв'язки за допомогою вказівника на ідентифікатор батьківського поняття $\Pr Cn$. При цьому загальне поняття відрізняється від часткових відсутністю батьківського $(\Pr Cn = NULL)$. Окрім формалізованих понять онтології, що відбираються з використанням формалізованих процедур, використовуються пояснюючі мета-поняття, які вводяться безпосередньо експертом і маркуються логічним значенням $True$ атрибуту $Meta$.

Кортеж O_Phr забезпечує мовні реалізації понять за допомогою окремих слів або словосполучень і представляє також деревоподібну структуру. Він включає ідентифікатори мови реалізації $IdLg$, фрази $IdPh$, основи $IdBs$ та форми $IdFm$ слова. Аналогічно попередньому зв'язки слів у фразах та фраз з поняттям визначаються за допомогою вказівника на ідентифікатор батьківської основи слова поняття $Id \Pr Bs$. Кортежі основ O_Bs та форм O_Frm слів служать для збереження відповідних словесних представлень $WBase$, $WForm$.

Метод формування онтологійного наповнення

Для формування онтологійного 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 онтологійного наповнення може розглядатися як інформаційний 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. Однією з ключових умов відсіву сторонньої інформації є включення ключових слів з множини KWS, що характеризують вибране загальне поняття в множину LIS значень елементу It деякого списку Lst аналізованої веб-сторінки Pg:

$$KWS \subset LIS_{It,Lst,Pg} \quad (5)$$

Ця умова є достатньо жорсткою, тому в множину ключових слів включаємо мінімальну кількість елементів. Виконання умови (5) хоча б для одного елементу списку приводить до включення всього списку в аналізовану структуру. В цю структуру включаються також списки, які містять елементи зі сформованої онтології.

Аналізована структура також є деревоподібною і має наступний вид:

$$AS =< IdPg, IdLst, IdIt, IdBs, IdFm, Id \Pr Bs > \quad (6)$$

Вона багато в чому еквівалентна структурі (2), однак містить ідентифікатори сторінки, вибраного списку на ній а також елемента самого списку, оскільки він може складатися зі декількох слів. Для виділення елементів списку використовуються роздільники, які утворюють спеціальну множину сепараторів:

$$SS = \{";",",","–","/","or","and","и","или","і","або",":","(",")","*"\}, \quad (7)$$

а також, звичайно, теги $< li >$ $< /li >$ елементів списку. Словоформи вибираються безпосередньо із елемента списку і розпізнаються за допомогою відношення форм O_Frm або поповнюють його. Основи розпізнаних словоформ вибираються із відношення основ O_Bs, а якщо вони не розпізнані, то будуються за допомогою відкидань елементів, що входять в множину закінчень En. Після ідентифікації основи слова відбувається її пошук у відношенні частот основ, структура якого задається наступним кортежем

$$BF =< IdBs, BsFr, IdLPg, Phn > \quad (8)$$

Якщо основа знайдена і номер поточної сторінки не співпадає з номером останньої врахованої, то індекс частоти $BsFr$ збільшується на 1, а номер поточної сторінки заноситься в поле $IdLPg$. При такому підході на частоту основи впливає її використання лише на різних сторінках. Якщо основа не знайдена, вона заноситься у відношення частот основ із індексом частоти рівним 1 та номером поточної сторінки.

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. Для кожного входження основи в відношення AS вибираються словоформи, що формують

елемент списку, який містить дану основу. Цей елемент включається в контекст, якщо аналогічного елемента в ньому ще немає а також він не входить в перелік фонових контекстів PhC. Допускається невідповідності лише у так званих стоп-словах.

Після сформування контексту він подається експертові для аналізу. Після відбору експертом елементів для онтології, елементи списку, жодна компонента якого не була відібрана, запам'ятовуються для виключення їх повторної подачі в контексті іншого терміна. З цією метою формується спеціальна структура фонових контекстів

$$PhC =< IdC, IdBs, Id \Pr Bs > \quad (9)$$

Також фіксуються часто згадувані основи, які не ввійшли в жоден відібраний термін. Вони маркуються як фонові і не будуть служити основами наступних контекстів.

Алгоритм формування онтологійного наповнення

На основі наведених теоретичних положень сформуємо алгоритм автоматизованого формування онтологійного 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 з множини KWS до кожного сайту з множини StL та формуємо множини $HTML$ кодів веб-сторінок.

3. Якщо сторінка містить список, хоча б один елемент якого містить слова з 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 з множини SS, а елементарні терміни розбиваються на слова. Основи відібраних слів заносяться у відношення BF, а якщо вони вже там зареєстровані з сторінки, що не співпадає з 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 у відношення AS вибираються словоформи, що формують елемент списку, який містить дану основу. Цей елемент включається в контекст, якщо аналогічного елемента в ньому ще немає, а також він не входить в перелік фонових контекстів PhC.

6. Після сформування контексту він подається експертові для аналізу. Після відбору експертом елементів для онтології, елементи списку, жодна компонента якого не була відібрана, запам'ятовуються у відношенні PhC для виключення їх повторної подачі в контексті іншого терміна.

7. Основи, які не ввійшли в жоден відібраний термін онтології, а також основи, що ввійшли в онтологію, маркуються як фонові. Перехід до пункту 2.

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 вирано 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 зі входженням ключового слова "PHP".

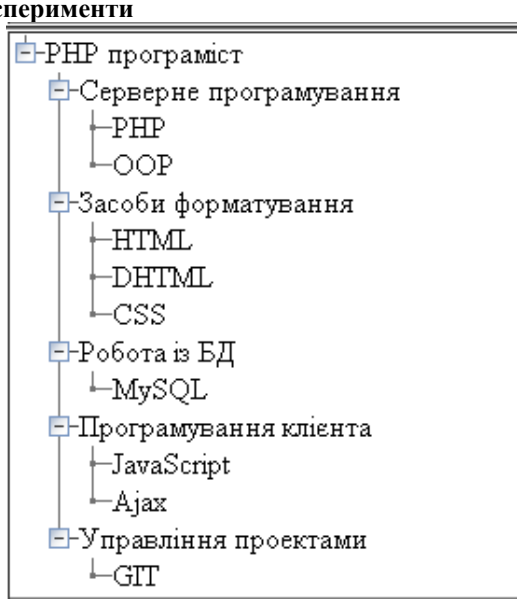

Рис. 1. Онтологія, побудована на основі автоматизованого аналізу п'яти Веб-сторінок

Сама онтологія, побудована на основі автоматизованого аналізу п'яти Веб-сторінок, подана на рисунку 1. 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. Сюди були віднесені наступні терміни: "Понимание", "Программирование",

“Проектирование”, “Knowledge”, “Опыт ”, “Работа”, “Хорошее”, “Знание”. Всього експертові для перегляду з врахуванням контексту було подано 23 стрічки, з яких відібрано 17 понять. При перегляді повних текстів 5 аналізованих сторінок експерт повинен був би переглянути біля 200 стрічок. Тобто вдалося принаймні на порядок зменшити завантаженість експерта і частково зняти інформаційну зашумленість даних.

За критерієм результативності відбору важливої інформації ([кількість відібраних термінів]/[кількість переглянутих стрічок]) ефективність роботи експерта зросла від значення 0.085 до 0.739.

Висновки

У статті розглянуто один з можливих шляхів формування онтологійного наповнення 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 наступніі наукові та практичні результати. Вперше запропоновано формування онтологійного наповнення шляхом аналізу зашумленої слабко 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. Формалізм в постановці задачі створення якісного сайту / Н.Р. Пасічник, М.П. Дивак // Наукові праці ДонНТУ. Інформатика, кібернетика та обчислювальна техніка. – 2011. – Вип. 14 (188). – С. 325– 329.

2. Глибовец Н.Н. Становление технологии WEW 3.0. [Електрон. ресурс] / Н. Н. Глибовец, Л. О. Шыпович. – 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

Надійшла 24.9.2012 р.
Рецензент: д.фіз-мат.н. Боднар Д.І.

УДК 621.317

К.Л. ГОРЯЩЕНКО
Хмельницький національний університет

ОГЛЯД КЛАСИЧНИХ МОДЕЛЕЙ ПРОВІДНИКОВИХ РЕГУЛЯРНИХ ЛІНІЙ ПЕРЕДАЧІ

В статті розглядаються класичні моделі провідникових регулярних ліній передачі, що застосовуються в загальному аналізі провідникових ліній. Показано, що відомі прості моделі є спрощеними і не дають можливості врахування частото-залежних параметрів лінії у первинних та вторинних параметрах лінії.

The article deals with the classical model of regular conductor transmission lines used in the overall analysis of conductor lines. The famous simple model is simplified and does not allow for incorporation of frequency-dependent parameters of the line in the primary and secondary parameters line.

Ключові слова: провідникова регулярна лінія, проста модель лінії.

Вступ

За останні роки суттєво зросла зацікавленість у моделюванні провідникових ліній із застосуванням різного програмного забезпечення. Причини такого зростання є цілком очевидні. Високі капітальні витрати, невідновлювана ізоляція та висока вартість на обслуговування та заміну кабельних мереж у випадку руйнування ізоляції або провідникового осердя. Крім того, необхідність визначення поточного стану лінії, її параметрів, тенденцій в роботі та інші – всі ці фактори стимулюють створення та дослідження математичних моделей провідникових ліній. Кожна з моделей має бути максимально адекватна до досліджуваної реальної лінії.

Постановка задачі

Математична модель провідникової лінії повинна відповідати визначеному типу лінії або окремої групи ліній. Під відповідністю моделі приймають відтворення характеристик цієї кабельної лінії. Це дає можливість визначати особливості будови, взаємодію елементів між собою, взаємодію між елементами та зовнішнім середовищем, визначати умови роботи із сигналом, що розповсюджується в лінії та обов’язково надавати можливість проводити аналіз роботи цієї лінії, так і повинна дозволяти проаналізувати роботу та зміну характеристик різноманітних з’єднань, що утворені відрізками кабелю та підключені до зовнішніх