УДК 004.054

ОСОБЛИВОСТІ ПОБУДОВИ МЕТОДУ ВИЯВЛЕННЯ НЕКОРЕКТНОЇ ТА ЗАСТАРІЛОЇ ІНФОРМАЦІЇ НА ВЕБ-РЕСУРСАХ

Дивак М.П.[1)], Ковбасістий А.В.[2)]
Тернопільський національний економічний університет
[1)] д.т.н., професор; [2)] магістрант

I. Постановка проблеми

Розвиток інтернет технологій сприяв появі великого об'єму веб-ресурсів із інформацією, яку можна отримати з пошукових серверів. Разом із тим веб павутина містить ресурси, які в свою чергу наповненні застарілою або некоректною інформацією. Переважно такі вади мають сайти-візитки, які використовують багато організацій та установ.

Сайт-візитка - це повноцінний віртуальний ресурс із унікальним дизайном та програмною частиною, що керується й може повністю змінюватися самим клієнтом.

За основу для перевірки інформації веб-наповнення сайту, наприклад, можна використати внутрішні бази даних (БД) відділу кадрів або бухгалтерії, в яких зберігається оновлена та завжди актуальна інформація про працівників.

Один із прикладів організації такої інформації наведено в таблиці 1.

Таблиця 1.

Фрагмент таблиці представлення штату кафедри

Name	Position	Academic title
Коваленко Іван Петрович	Старший викладач	кандидат техн. наук
Бойко Василь Іванович	Доцент	кандидат техн. уаук
Ткаченко Тарас Юрійович	Викладач	

Разом із тим на веб-ресурсах часто інформація є відмінною від актуальної. Тож актуальним є метод автоматичного виявлення та усунення некоректної та застарілої інформації на сайтах-візитках із використанням внутрішньо організаційних БД.

Для цього необхідно також розробити засоби автоматичного аналізу html коду веб-сторінки.

II. Особливості реалізації методу

Як відомо, усі веб-сторінки складаються із двох частин - заголовка і тіла. Тема містить необхідну інформацію для ідентифікації сторінки, а в тілі відображається весь контент. Після заголовку <! DOCTYPE> починається веб-сторінка, позначена тегом <html>, а кінець веб-сторінки позначається тегом </ html>. Між ними розміщуються заголовок і тіло: пара тегів <head> і </ head> визначають заголовок, а <body> і </ body> - тіло документа. Тіло документа може містити будь-яку кількість будь-яких тегів (або взагалі не міситити їх), які будуть представляти інформацію користувачам.

Спочатку для реалізації методу необхідно здійснити приведення інформації представленої html кодом до єдиного стандарту.

У роботі використаємо парсинг сайтів - послідовний синтаксичний аналіз інформації, розміщеної на інтернет-сторінках, для приведення до єдиного стандарту вмісту веб-сторінок та баз даних.

Якісний парсинг обов'язково складається з трьох основних етапів:

збір інформації в первісному вигляді. Автоматичний парсинг - копіювання коду конкретної сторінки з наступним витяг з неї контенту (необхідних даних). Досить часто коди необхідних сторінок витягають зі спеціальної бібліотеки cURL.

фаза вилучення і подальшої зміни формату інформації. Отримавши код сторінки, парсер для вилучення користується, так званими, «регулярними виразами». Якщо є необхідність, на цьому етапі можна зробити перетворення даних в конкретний формат

генерація результату - це остання фаза парсингу. Під час неї виводиться або записується інформація, отримана в процесі попередньої фази. У переважній більшості випадків інформація відразу ж переводиться в необхідний формат з подальшим записом в базу.

Після здійснення парсингу генеровані результати необхідно опрацювати, щоб надати інформації вигляду придатного для подальшого використання. Конкретний формат залежить від того, як в подальшому будуть оброблятися зібрані дані. Доволі часто з парсенного контенту, за допомогою XML, формується RSS-потік, який зручний для використання даних, без процедури рерайтингу. Іноді результат парсингу поміщають в CSV-файл, оскільки цей текстовий формат дуже простий у подальшій обробці, легко конвертується в SQL-запити і без проблем відкривається в Excel. У особливих випадках потрібно, щоб кінцеві дані були представлені у вигляді електронних таблиць XLS.

Наведемо загальну схему реалізації методу виявлення некоректної та застарілої інформації на веб-ресурсах (рисунок 1)

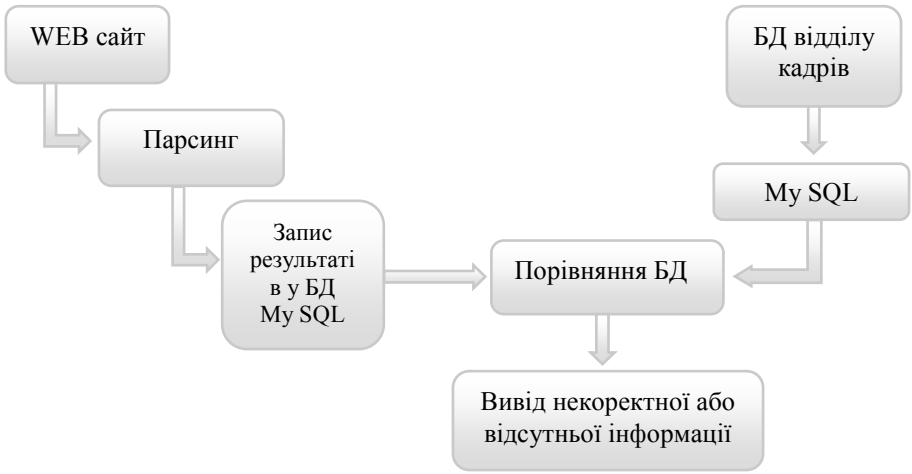

Рисунок 1 - Схема реалізації методу

Висновок

Показано, що через великий об'єм інформації на веб-ресурсах часто дані подані на них є застарілими та некоректними. Серед них найбільше вад містять сайти-візитки, які використовують практично усі організації та установи. Тому основою для перевірки достовірності інформації доцільно було використати внутрішні бази даних організацій.

Наведено основні процедури реалізації методу приведення інформації з html коду до єдиного стандарту, а також загальну схему реалізації методу виявлення застарілої та некоректної інформації.

Список використаних джерел

1. Pasichnyk N. Mathematical modeling of the Website quality characteristics in dynamics / N.Pasichnyk, M.Dyvak, R.Pasichnyk// Journal of Applied Computer Science / Technical University Press – Lodz, Poland, 2014. – Vol. 22 № 1. – PP. 171-183.
2. Пасічник Н. Метод та алгоритм побудови структур тематичних Веб-сайтів на основі онтологічного підходу. / Н.Р. Пасічник, М.П. Дивак // Наукові праці Донецького національного технічного університету, серія «Інформатика, кібернетика та обчислювальна техніка», 2012. – Вип 15 (203). – С. 184-189.
3. Пасічник Н. Метод формування онтологічного контенту на основі аналізу інформації спеціалізованих Веб-сайтів. / Н.Р.Пасічник // Вісник Хмельницького національного університету:Технічні науки, 2012.-Вип. №5.-С. 241-244.