УДК 004.62
Ю. Горбуляк
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ОГЛЯД МЕТОДІВ МАЙНІНГУ WEB-КОНТЕНТУ

UDC 004.62
Yu. Horbuliak

SURVEY OF THE METHODS OF WEB-CONTENT MINING

Майнінг веб-контенту – це пошук, сканування та видобування (отримання) тексту, відео, графіків та зображень із веб-документів. Дані вмісту відповідають набору фактів, які веб-сторінка була розроблена для передачі користувачам. Більшість даних, доступних в Інтернеті, є неструктурованими даними.

Автоматичне виявлення інформації в Інтернеті є складним через відсутню структуру джерел інформації у всесвітній мережі. Допомогою в пошуку інформації є традиційні пошукові системи, такі як Google, Bing, Yahoo або AltaVista. Але проблема в тому, що вони не надають структурної інформації шляхом категоризації, фільтрації чи інтерпретації документів.

У аналізі веб-контенту використовуються два типи підходів: підхід на основі бази даних і підхід на основі агента. Підхід на основі бази даних намагається розробити методи організації напівструктурованих даних, що зберігаються в Інтернеті, у більш структуровані колекції інформаційних ресурсів. Тоді для аналізу цих колекцій можна використовувати стандартні механізми запитів до бази даних і методи аналізу даних. Підхід до баз даних можна розділити на два підтипи: багаторівневі бази даних і системи веб-запитів.

Підхід на основі агента використовує так звані веб-агенти для збору відповідної інформації зі всесвітньої мережі. Веб-агент – це програма, яка відвідує веб-сайт і фільтрує інформацію, яка цікавить користувача. Існують три підтипи підходу на основі агента: агенти інтелектуального пошуку, фільтрація/категоризація інформації та персоналізовані веб-агенти.

Текстовий документ – це форма неструктурованих даних. Більшість даних, які доступні в Інтернеті, є неструктурованими даними. Дослідження застосування методів аналізу даних до неструктурованих даних відоме як отримання знань з текстів. Для отримання інформації з неструктурованих даних використовується веб-шаблон зіставлення. Він відстежує ключові слова та фрази, а потім з’ясовує зв’язок ключових слів у тексті. Коли є великий обсяг тексту, ця техніка дуже корисна. Отримання інформації перетворює неструктурований текст у більш структуровану форму. Спочатку з витягнутих даних видобувається інформація, потім за допомогою різних правил виявляється пропущена інформація. Видобута інформації, що робить неправильні прогнози щодо даних, відкидається.

Також при опрацюванні WEB-контенту можливе застосування аналізу структурованих даних для власне отримання структурованих даних з веб-сторінок. Прикладами структурованих даних є дані у вигляді списку, таблиць чи дерева.

Веб-джерела зберігають велику кількість даних, котрі можуть бути чітко видобуті за допомогою методів веб-майнінгу при умові коректного підбору цих методів. Видобуток веб-контенту виявився дуже корисним у сферах електронної комерції, соціальних мереж тощо. Задачі, пов'язані з пошуком інформації пошуковими системами, виявилися великою проблемою. Майнінг веб-контенту вирішує ці задачі. Незважаючи на те, що доступно багато різних методів отримання різноманітних типів даних в Інтернеті, існує потреба в подальшому покращенні ефективності та результативності отримання потрібної інформації з Інтернету.