Самойлик А.В., Ушеренко Р.Б., Шафрай А.Ю.
ННК "ІПСА" НТУУ "КПІ"

Аналіз Web-сайтів інформаційних Web-порталів

Задача аналізу Web-сайтів виникає у багатьох випадках. По-перше, це перевірка працездатності сайтів, що входять до порталу. По-друге, це аналіз характеристик Web-сайтів для оцінки діяльності організацій, наприклад, наукової діяльності університетів, як це робить Webometrics [1]. Webometrics – це світова система оцінювання рейтингу вищих навчальних закладів по науковій роботі. При цьому рейтинг університету в системі залежить від того, наскільки вимоги Webometrics забезпечуються на внутрішніх Web-сайтах.

Програмна реалізація система аналізу Web-сайтів включає наступні підсистеми:

1. Робот обходу Web-сайтів і збирання інформації.
2. Програма обрахунку характеристик згідно з критеріями аналізу.
3. Програма формування результатів аналізу.
4. Програма розсилання результатів аналізу Web-майстрам сайтів, та на сайт http://webometr.ntu-kpi.kiev.ua.

Робот обходу враховує уявлення Web-сайтів як направленого графа. Однією з основних функцій робота є збирання текстової інформації. Крім надання даних для аналізу текстовий масив дозволяє реалізувати таку функцію як пошукова система на заданій множині Web-сайтів. Ця система, на відміну від традиційних пошукових систем, чітко орієнтована на джерела, де знаходиться потрібна інформація, і не видає зайвої інформації.

У системі використовується два набори характеристик. Перший набір пов'язаний з аналізом працездатності Web-сайтів. При цьому аналізується не тільки працездатність на цей час, але й порівняння з попереднім аналізом. У якості основних параметрів розглядаються:

1. Загальна працездатність (працює чи ні сайт).
2. Загальний обсяг сайту та його основних розділів.
3. Останній термін оновлення сайту.
4. Наявність застарілої інформації.

Другий набір пов'язаний з оцінкою параметрів, що враховує Webometrics:

Розмір (Size) – число сторінок.

Видимість (Visibility) – число унікальних зовнішніх зв'язків.

Цінні файли (Rich files) – файли у форматах, які зазвичай використовують автори для представлення та поширення власних робіт.

Цитованість (Scholar) – Google Scholar дозволяє оцінити цитованість для кожної наукової установи.

Кожний з цих параметрів має різну вагу при оцінці рейтингу. Нажаль, не всі параметри, що використовує Webometrics можуть бути проаналізовані. Програмними засобами аналізуються "Розмір" та "Цінні файли", але вони у рейтингу мають найбільшу вагу. При аналізі "Розмір" оцінюється не тільки кількість сторінок, але й їх об'єм. При аналізі Цінних файлів по рекомендаціям Webometrics оцінюються файли (кількість і об'єм) у форматах .pdf, .doc, .ps та .ppt і підраховується як загальна кількість на сайт, так і на їх основних розділах. Також надається порівняння з попереднім аналізом.

Подальше вдосконалення системи аналізу пов'язано з розширенням кількості параметрів включаючи аналіз мультимедійних файлів, семантичний аналіз текстової інформації.

Література

1. http://www.webometrics.info/methodology.html