УДК 519.688

К.т.н., доцент Петрашенко А.В., студент Саядян С.Г.

Національний технічний університет України
«Київський політехнічний інститут»

ЗАСТОСУВАННЯ ФОРМАТУ MARKDOWN ДЛЯ АНАЛІЗУ ГІПЕРТЕКСТОВИХ ДОКУМЕНТІВ

Abstract

Andriy V. Petrashenko, assoc. prof., PhD; Surik Saiadian, student
The using of the Markdown format to analyze hypertext documents.

This paper presents an automatic tag-tree independent approach to detect central article of hypertext document. Comparing to other existing techniques, this approach provides better results even when the HTML structure is far different from layout structure. The ways for further research are proposed as well.

Вступ

У світі щодня виникає незмірна кількість подій, які висвітлюються в новинах на web-порталах. На сьогоднішній день існує велика кількість web-порталів, причому зберігається тенденція до її зростання. У зв'язку з цим виникає необхідність в автоматичних або автоматизованих системах пошуку та впорядкування новин [4]. Однією з основних задач таких систем, без якої вони не можуть функціонувати, є задача збору новин. Під збором новин розуміють пошук так званої центральної статті на web-порталах, що власне містить саму новину, та видалення іншої непотрібної інформації, такої як меню навігації, банери тощо.

У даній статті пропонується метод пошуку центральної статті гіпертекстового документу за допомогою мови розмітки Markdown. Використання даного методу дозволить підвищити ймовірність точного вилучення центральної статті з гіпертекстових сторінок.

Постановка задачі

Задача полягає в створенні та дослідженні алгоритму виділення центральної статті з гіпертекстових сторінок Web-ресурсів ЗМІ.

Опис алгоритму

Markdown - це полегшена мова розмітки. Вона була створена, як легка для читання і зручна у публікації мова розмітки. Основне застосування дана мова знайшла в різних системах обробки тексту, текстових редакторах тощо. Markdown широко використовується для написання статей та електронних листів з форматуванням. Основними елементами мови є: текст з виділенням, програмний код, списки, заголовки, цитати, посилання та зображення.

Традиційно для аналізу гіпертекстових документів використовується DOM-модель, яка дозволяє програмам і скриптам отримати доступ до вмісту HTML-документу [1-3]. При цьому виконується обхід по DOM-дереву і проводиться аналіз вмісту документу. Кожен з вузлів такого дерева може являти собою елемент, атрибут, текстовий, графічний або будь-який інший об'єкт. Вузли зв'язані між собою відносинами «батько - син».

Звичайно DOM-дерево містить в собі весь гіпертекстовий документ, що призводить до ускладнення навігації по дереву в процесі пошуку змістовної частини документа. До того ж для виконання навігації необхідні великі обчислювальні ресурси та великі обсяги пам'яті.

Дану задачу можна вирішити шляхом перетворення HTML-документу в документ з мовою розмітки Markdown з подальшим аналізом отриманого документу. Найчастіше, перетворення сторінки в формат Markdown виконується за один прохід. Завдяки тому, що тільки конкретні елементи гіпертекстового документа потрапляють в документ Markdown, вже на етапі перетворення в документ потрапляє лише інформаційна частина гіпертекстового документу. Всі посилання на інші джерела, а також посилання на зображення виділяються у вигляді окремого блоку, який додається в кінець Markdown-документу, а до тексту потрапляють тільки виноски на посилання.

Існує досить велика кількість бібліотек для перетворення гіпертекстових документів з формату HTML в формат Markdown, але найбільш широко використовується бібліотека «html2text».

Таким чином, алгоритм виділення центральної статті можна представити у вигляді послідовного виконання двох етапів:

1. Перетворення гіпертекстового документу у формат Markdown.
2. Видалення «інформаційного шуму», який залишився після першого етапу.

Розглянемо детально кожен етап.

На першому етапі під час перетворення HTML документу в Markdown документ вже здійснюється відсіювання «інформаційного

шуму». Видаляються різні елементи HTML сторінок, такі як скрипти, елементи управління: меню і кнопки, банери, допоміжні блоки. На виході отримуємо документ з основним контекстом і невеликими домішками інформаційного шуму, які будуть фільтруватися на другому етапі.

Центральна стаття є цілісною одиницею, яка розташована в певній області документа, тому для виділення центральної статті з Markdown документу нам необхідно знайти її початок та кінець. На другому етапі проводиться пошук верхньої та нижньої меж, за якими відбувається обробка документу.

Після аналізу близько 200 сторінок з різних джерел, було виявлено, що заголовок центральної статті практично завжди зустрічається в заголовку гіпертекстового документу. Таким чином якщо провести пошук по Markdown документу за словами і словосполученнями, які зустрічаються у заголовку гіпертекстового документу, можна знайти початок центральної статті в Markdown документі. Також, в результаті аналізу було з'ясовано, що заголовок центральної статті в більшості випадків виділено тегом «заголовок» ('#') або «виділення» ('*'). Якщо брати до уваги два вищеописаних правила, то з високою ймовірністю можна визначити верхню межу центральної статті в Markdown документі.

Задача пошуку нижньої межі центральної статті гіпертекстового документа більш складна. Необхідно провести аналіз кожного абзацу після заголовку центральної статті. Аналіз проводиться за наступними критеріями:

1. Ключові слова, які визначаються за базою ключових слів, що була отримана в результаті аналізу гіпертекстових сторінок з різних джерел. Ці слова в переважній більшості випадків зустрічаються після закінчення статті. Підраховуючи коефіцієнт співвідношення загальної кількості слів з кількістю ключових слів у абзаці, необхідно визначити абзац, який можна вважати нижньою межею центральної статті.
2. Середня довжина речення в абзаці.
3. Відсоток стоп-слів у реченнях абзацу.

Таким чином, за цими критеріями визначається нижня межа. Результатом оцінки за кожним критерієм є цифра від нуля до одиниці. Кожен критерій має свій коефіцієнт важливості. У результаті для кожного абзацу вираховується сума добутків коефіцієнтів важливості на результати оцінки за критеріями:

$$S = \sum_{i=1}^{n} k_i R_i ,$$

де S - оцінка абзацу, k_i - коефіцієнт важливості i-го критерію, R_i - результат оцінки за i-тим критерієм, та

$$\sum_{i=1}^{n} k_i = 1.$$

Якщо ця сума перевищує граничне значення, то цей абзац вважається нижньою межею.

Після одержання обох меж документу (верхньої та нижньої), відкидаються частини Markdown документу до верхньої межі, та після нижньої. В результаті маємо отримати центральну статтю гіпертекстового документу. Далі результат зберігається до бази даних для подальшої обробки.

Висновки

Формат Markdown можна успішно використовувати для аналізу гіпертекстових документів. З десяти випадкових web-сторінок, які не використовувалися при формуванні критеріїв, у семи були успішно знайдені та вилучені центральні статті.

Подальше проведення аналізу якомога більшої кількості різних web-сторінок, корегування коефіцієнтів критеріїв і граничного значення нижньої межі та доповнення бази ключових слів, які зустрічаються після закінчення статті дозволить підвищити ефективність знаходження центральної статті документу.

Подальше поліпшення алгоритму, наприклад, може бути виконано за допомогою введення додаткового критерію, який би враховував оцінки попереднього і наступного абзаців.

Література

1. *Adelberg, B.*, NoDoSE: A tool for semiautomatically extracting structured and semistructured data from text documents, In Proceedings of ACM SIGMOD Conference on Management of Data, 1998, pp. 283-294.
2. *Ashish, N. and Knoblock, C. A.*, Semi-Automatic Wrapper Generation for Internet Information Sources, In Proceedings of the Conference on Cooperative Information Systems, 1997, pp. 160-169.
3. *Buttler, D., Liu, L., and Pu, C.*, A Fully Automated Object Extraction System for the World Wide Web, In International Conference on Distributed Computing Systems, 2001.
4. *Крейнес М. Г.* Технология смыслового поиска информации в сетевых информационных ресурсах / // Искусственный интеллект. №2 – Донецк. –2000. – С. 114-124.