УДК 681.3.06

В.М. ДУБОВОЙ, О.Ю. КРАКОВЕЦЬКИЙ, О.В. ГЛОНЬ

МЕТОД ОЦІНКИ ПОДІБНОСТІ ВЕБ-СТОРІНОК

Вінницький національний технічний університет,
вул. Хмельницьке шосе, 95, м Вінниця, Україна, 21021,
тел.: +380 (432) 43-78-96, E-mail: dub@faksu.vstu.vinnica.ua

Аннотація. Наведено метод оцінювання Інтернет-сторінок, та їх алгоритм очищення від інформаційного шуму, а також запропоновано алгоритм видалення веб-сторінок з даними, що дублюються.

Аннотация. Приведен метод оценивания Интернет-страниц, и их алгоритм очищения от информационного шума, а также предложен алгоритм удаления веб-страниц с данными, которые дублируются.

Ключові слова: подібність веб-сторінок, c-means, VIPS, критерій Пірсона χ^2, кластеризація веб-сторінок.

ВСТУП

Проблема вибору необхідного і, головне, корисного матеріалу в Інтернеті серед безлічі документів є досить актуальною. Пошукові системи використовують різноманітні коефіцієнти (Google – PageRank [1, 2], Яндекс – ТІЦ [2] тощо) для сортування результатів пошуку. Таким чином, «розкручені» сайти будуть завжди знаходитися на вершині списку, хоча дуже часто частина цих сайтів містить ідентичну інформацію, що призводить до втрачання часу користувачами на «візуальне» фільтрування результатів пошуку. Наприклад для пошукового запиту *«Иллюстрированный самоучитель по MathCAD 12»* [3] з перших 10 ресурсів 5 мали ідентичну інформацію в тій чи іншій мірі. Таким чином, користувач втрачає до 50% часу для перегляду даних ресурсів. Потрібно зазначити, що інформація на цих сайтів дублюється не в повній мірі, а лише основна її частина (контент), все інше - дизайн, посилання, рекламні блоки відрізняються. Таким чином актуальними задачами є визначення основного контенту сторінок та оцінка їх подібності.

ОГЛЯД ЛІТЕРАТУРИ

Поняття подібності веб-сторінок в літературі безпосередньо не розглядається. Натомість існує ряд робіт, присвячених задачам кластеризації текстової інформації [4-6], що оперують поняттям «відстанні» між об’єктами, що по своїй суті схоже з поняттям подібності. Для обчислення відстані використовують міри Ейлера, Меланхобіса, критерій Пірсона тощо. Наприклад, відстань можна знаходити за допомогою критерію χ^2 Пірсона

$$\chi^2 = n_1 n_2 \sum_{i=1}^{k} \frac{1}{m_{1,i} + m_{2,i}} \left(\frac{m_{1,i}}{n_1} - \frac{m_{2,i}}{n_2} \right)^2, \qquad (1)$$

де n_1 - об’єм першого тексту, n_2 - об’єм другого тексту, $m_{1,i}$ - частота i-ї ознаку в першому тексті, $m_{2,i}$ - частота i-ї ознаки в другому тексті [7]. Значення $m_{i,j}$ (матриця подібності) обчислюється як кількість входжень i-ї ознаки в j-у документі: $m_{i,j} = count(f_{i,j} \in p_j)$. Відповідно, коефіцієнт подібності можна обчислювати як $S_{p_i, p_2} = 1/\chi^2$.

Даний підхід має ряд недоліків: значення відстанні може відрізнятися для однакового степеня дублювання даних за рахунок різної довжини текстів, значення відстанні достатньо великі навіть для невеликих текстів, тобто необхідно вводити понижаючі коефіцієнти і, накінець, основний недолік в тому,

© В.М. ДУБОВОЙ, О.Ю. КРАКОВЕЦЬКИЙ, О.В. ГЛОНЬ, 2008

що отримане значення відстані (і відповідно й подібності) неможливо оцінити експертним шляхом, тобто визначити чи отримана відстань є великою чи малою. Для оцінки експерним шляхом необхідно мати повну матрицю відстаней.

АКТУАЛЬНІСТЬ ТА МЕТА РОБОТИ

Велика кількість даних та технологій «розкрутки» сайтів призводить до необхідності розробки методів та алгоритмів фільтрації результатів пошуку з метою покращення ефективності пошуку в Інтернеті.

Метою даної роботи є вирішення задачі очищення веб-сторінок від інформаційного шуму та розробка методу оцінки подібності веб-сторінок.

ОЧИЩЕННЯ ВЕБ-СТОРІНОК ВІД ІНФОРМАЦІЙНОГО ШУМУ

Алгоритм виділення основного контенту зі сторінки полягає в наступному:

1. На вхід подається веб-сторінка, яка ділиться на окремі інформаційні блоки.
2. На основі регресійної моделі оцінки важливості інформаційних блоків сайтів.

$$F = 1.739 + 0.033\cdot ImgsNum - 0.062\cdot ImgsAsLinksNum + 0.087\cdot ImgsAsLinksRatio + 0.002\cdot LinksNum - 0.006\cdot WordsAsLinksNum + 0.291\cdot WordsAsLinksRatio + 0.012\cdot SentNum + 1.523\cdot SentAvgLengthRatio - 0.005\cdot WordsInSentsNum + 1.75\cdot WordsInSentsRatio - 0.164\cdot StopWordsNum - 8.22\cdot StopWordsRatio + 0.004\cdot WordsNum - 0.003\cdot ListItemsNum - 0.002\cdot HeadersNum - 0.14\cdot ControlsNum - 0.456\cdot MediaObjectsNum + 3.712\cdot ContentRatio + 0.849\cdot WordsAsListsRatio + 0.105\cdot FontSize + 0.002\cdot FontWeight,$$

для кожного блоку розраховуються числові значення важливості.

3. За допомогою нечіткого методу кластеризації c-means [6] (на основі числових значень оцінки важливості) блоки діляться на три кластери (відповідно до трьохрівневої системи оцінки важливості).
4. На виході отримуємо сторінку, до якої входять лише ті блоки, які були ідентифіковані як важливі.

Розглянемо роботу алгоритму на реальному прикладі. В якості тестової сторінки було взято статтю інформаційного агенства CNN [8]. Вона була розбита на інформаційні блоки, а кожен з блоків був оцінений за допомогою регресійної моделі. В результаті отримано 7 блоків, які були кластеризовані за допомогою нечіткого методу кластеризації c-means. Результати наведено в таблиці 1.

Таблиця 1.

Результати нечіткої кластеризації

№ блоку № класт.	4	5	7	1	3	2	6
I	0.0449	0.0133	0.0881	0.3291	**0.9963**	**0.9844**	0.0000
II	0.0059	0.0015	0.0056	0.0139	0.0005	0.0033	**1.0000**
III	**0.9492**	**0.9852**	**0.9063**	**0.6569**	0.0032	0.0122	0.0000

В таблиці 2 наведено порівняння кластеризованих результатів і результатів, отриманих за допомогою експерта (1 – інформація неважлива, 2 – мало важлива, 3 – основний контент).

Таблиця 2.

Порівняння результатів оцінки важливості блоків

№	Оцінка за регресійною моделю	Експ. оцінка	Кластер. оцінка
1	7.88679307143625	1	1
2	13.5747065165397	2	2
3	38.154141209372	3	3
4	1.1426199376947	1	1
5	2.44260051472691	1	1
6	15.1783915874052	2	2
7	6.23407392941833	1	1

Таким чином, на основі результатів можна побачити, що запропонований алгоритм успішно справляється з поставленою задачею.

МЕТОД ОЦІНКИ ПОДІБНОСТІ ВЕБ-СТОРІНОК

Під *подібністю* веб-сторінок будемо розуміти степінь дублювання даних, що вони містять

$$Similarity(page_1,\ page_2) = data_1 \cup data_2. \tag{2}$$

Під *коефіцієнтом подібності* будемо розуміти числову характеристику подібності, причому повному дублюванню даних буде відповідати значення 1, відповідно 0 – документам, що повністю відрізняються.

Нехай p_1 і p_2 - дві веб-сторінки, для яких необхідно визначити їх подібність, f_1 і f_2 - набір унікальних ознак відповідних веб-сторінок. Під унікальними ознаками матимемо на увазі набір ознак, що не повторюються в межах однієї веб-сторінки. В якості таких ознак будемо використовувати елементи інформаційного наповнення, а саме слова, речення, посилання та графічний контент (в принципі, використання інших елементів також допустиме). Загальний набір ознак F (2) утворюється шляхом об'єднання набору ознак веб-сторінок, що порівнюються. Ознаки, що повторюються, також видаляються з набору

$$F = f_1 \cup f_2,\ count(F) \le count(f_1) + count(f_2) \tag{3}$$

Коефіцієнт дублювання даних i -ї веб-сторінки в j -й будемо обчислювати за допомогою наступного виразу

$$\delta_{p_i(p_j)} = \frac{count(f_i) + count(f_j) - count(F)}{count(f_i)} \tag{4}$$

Загальний коефіцієнт даних, що дублюється в веб-сторінках p_1 і p_2 будемо обчислювати за допомогою виразу

$$\delta_{p_i,\ p_j} = \frac{count(f_i) + count(f_j) - count(F)}{count(f_i) + count(f_j)}, \tag{5}$$

а коефіцієнт подібності відповідно як

$$S_{p_i,\ p_j} = 2\delta_{p_i,\ p_j}, \quad S_{i,j} \in [0, 1]. \tag{6}$$

Відповідно відстань між веб-сторінками може бути обчислена за допомогою виразу

$$D_{p_i,\,p_j} = \frac{1}{S_{p_i,\,p_j}} \tag{7}$$

Таким чином, коефіцієнт подібності враховує невідповідність розмірів веб-сторінок та їх величини, а відстань між сторінками є певним чином обмеженою, так як коефіцієнт подібності може змінюватися лише в діапазоні від 0 до 1. Таким чином вирішується проблема розмірності відстані і подібності.

АЛГОРИТМ ВИДАЛЕННЯ ВЕБ-СТОРІНОК З ДАНИМИ, ЩО ДУБЛЮЮТЬСЯ

Даний алгоритм може використовуватися для фільтрування результатів веб-пошуку будь-якої пошукової системи.

Параметри алгоритму:

- кількість сторінок результатів пошуку n , що необхідно проаналізувати (зазвичай використовуються результати, що знаходяться на перших двох сторінках результатів пошуку, тому оптимальним є значення 15-20 ресурсів);
- максимальний допустимий коефіцієнт подібності $S_{\max}$, при якому сторінка вважаться такою, що дублюються і не включається в рузультуючий набір;
- метод видалення сторінок: *inline*, коли сторінки можуть видалятися при першому знайденому недопустимому дублюванні чи *postprocessed*, коли спочатку аналізуються всі сторінки, і лише потім приймаються рішення про видалення тієї чи іншої сторінки.

Принцип видалення веб-сторінки з результуючого набору:

$$\begin{cases} p_i \ if \ \delta_{p_i(p_j)} > \delta_{p_j(p_i)}, \\ p_j \ otherwise. \end{cases}$$

Покроковий опис алгоритму:

1. Користувач вводить пошуковий запит, пошукова система видає результати – набір сторінок P .
2. Аналізуються перші n веб-сторінок: вони розбиваються на блоки, оцінюються за допомогою регресійної моделі, інформаційний шум «відсікається». На виході отримуємо n веб-сторінок, які складаються лише з основного контенту.
3. Обчислюються коефіцієнти подібності між сторінками. Всі сторінки, що дублюються, видаляються з результуючого набору (в залежності від методу видалення сторінок).
4. На виході отримуємо набір сторінок, інформація в яких не дублюється.

ВИСНОВКИ

В даній роботі запропоновано алгоритм очищення веб-сторінок від інформаційного шуму, запропоновано метод оцінки подібності веб-сторінок та алгоритм видалення веб-сторінок з даними, що дублюються. Розроблені метод та алгоритми можуть бути реалізовані як надбудова до сучасних рішень і ефективно використовуватися в пошукових системах.

Подальша робота над покращенням розроблених методів та алгоритмів буде проводитися в

таких напрямках:

1. Аналіз додаткових ознак веб-сторінок для обчислення подібності та відстані.
2. Реалізація багатопотокового аналізу для зменшення часу аналізу.
3. Розробка методу автоматичного визначення максимально допустимого коефіцієнта подібності S_{max} для конкретного пошукового запиту.

СПИСОК ЛІТЕРАТУРИ

Стаття в Вікіпедії про PageRank // Режим доступу: http://ru.wikipedia.org/wiki/PageRank.

1. Что такое тИЦ? Что такое PR? // Режим доступу: http://fogmaker.net/post_1192898532.html.
2. Результати пошуку «Иллюстрированный самоучитель по MathCAD 12» в Google // Режим доступу:http://www.google.com.ua/search?hl=ru&q=Иллюстрированный+самоучитель+по+MathCAD+12&btnG=Поиск&meta=.
3. Дюк В. А., Самойленко А. П. Data Mining: учебный курс. - СПб.: Питер, 2001. – 368 с.
4. Барсегян А. А. Методы и модели анализа данных: OLAP И Data Mining. - СПб.: BHV, 2004. – 336 с.
5. Чубукова И. А. Data Mining. БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий - ИНТУИТ.ру. – 2006. – Режим доступу: http://www.intuit.ru/department/database/datamining/. – Заголовок з екрану.
6. Бормашов Дмитрий Александрович. Кластерный анализ текстов // Томск: Томск. гос. ун-т. Факультет информатики, 2006.- 43 с.
7. Стаття інформаційного агенства CNN // Режим доступу: http://edition.cnn.com/2008/WORLD/weather/09/06/hurricane.ike/index.html.

Надійшла до редакції 05.10.2008р.

ДУБОВОЙ ВОЛОДИМИР МИХАЙЛОВИЧ, д.т.н, професор – завідувач кафедри комп'ютерних систем управління Вінницького національного технічного університету.

КРАКОВЕЦЬКИЙ ОЛЕКСАНДР ЮРІЙОВИЧ – аспірант кафедри комп'ютерних систем управління Вінницького національного технічного університету.

ГЛОНЬ ОЛЬГА ВІТАЛІЇВНА – к.т.н., доцент кафедри комп'ютерних систем управління Вінницького національного технічного університету.