ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

УДК 681.3.06

В. М. Дубовой, д. т. н., проф.;
О. Ю. Краковецький, асп.;
О. В. Глонь, к. т. н., доц.

ФАКТОРНИЙ АНАЛІЗ ОЦІНКИ ВАЖЛИВОСТІ ІНФОРМАЦІЙНИХ БЛОКІВ САЙТІВ

Розглянуто проблему оцінки важливості інформаційних блоків сайтів, теоретичні основи досліджень, запропоновано ряд факторів, які потенційно можуть впливати на важливість блоків. Проведено факторний аналіз та отримано рівняння регресії для оцінки важливості інформаційних блоків сайтів.

Вступ

Швидкі темпи розвитку Інтернету як сховища різноманітної гіпертекстової інформації спричинює великий попит на автоматичні системи пошуку. До таких систем відносяться Google, MSN, Yahoo, Rambler, Яндекс та інші. Кожна з цих пошукових систем має свій власний алгоритм сканування всесвітньої мережі, визначення релевантності сторінок щодо пошукового запиту, визначення так званого «рейтингу» термів тощо.

Крім необхідної (релевантної) інформації кожен сайт може містити велику кількість додаткової інформації — елементи дизайну, навігацію, рекламні банери тощо, що майже завжди є безкорисною для користувача. Проте дане наповнення також враховується пошуковими системами, що в кінцевому рахунку приводить до погіршення основних статистичних характеристик ресурсу, що в результаті може привести до погіршення результатів пошуку. Ефективним варіантом вирішення даної проблеми є розділення веб-сторінок на блоки та врахування при аналізі лише тих, які безпосередньо відносяться до теми пошуку. В цій роботі розглянуто факторний аналіз оцінки важливості блоків, який може використовуватися для фільтрації основного наповнення в системах обробки, аналізу та пошуку інформації.

Перші роботи, що були повязані з пошуком на основі блоків, присвячені аналізу веб-сторінок в рамках одного веб-сайту для знаходження так званого «шаблону» [1], тобто загальних частин для всіх сторінок. Основною ідеєю цих робіт була теза, що «шаблон веб-сайту містить лише загальну інформацію і елементи дизайну» [2]. В своїй роботі [3, 10] Лін і Хо розробили систему InfoDiscover, яка ділить веб-сторінку на блоки (по тегам <table>), для яких потім підраховуються їх ентропії. В роботах [2, 4] було проведено дослідження, які показали, що відкидання інформаційного шуму потенційно може підвищити результати таких задач Data Mining, як класифікація і кластеризація даних. Інший клас робіт присвячений поділу конкретної веб-сторінки на інформаційні блоки. В роботі [5] запропоновано метод відсікання рекламних блоків на основі списку рекламних хостів. Даний метод є досить простим, проте він вимагає постійного слідкування за новими рекламодавцями. В роботі [6] запропоновано метод побудови ієрархічного М-дерева для виділення таких загальних блоків, як меню, верхня, нижня, центральна частини сторінки. Нарешті, в [7, 8] запропоновано відповідно метод VIPS для розділення сторінки на блоки та модель для їх оцінки. Дана модель є досить ефективною, проте вона не використовує ряд ознак вмісту блоків, які потенційно можуть вплинути на оцінку важливості блоків.

Стрімкий розвиток інформаційних технологій аналізу текстової інформації, підвищення вимог до автоматичних аналітичних та пошукових систем, а також ряд робіт по даній тематиці свідчать про *актуальність досліджень*.

Метою роботи є проведення факторного аналізу для оцінки важливості блоків сайтів.

Теоретичні основи побудови моделі

Релевантність блоку — це міра відповідності результатів пошуку і задачі, що поставлена в пошуковому запиті; визначає, наскільки повно блок відповідає критеріям пошукового запиту [9].

Під *важливістю блоку* будемо вважати певну числову безрозмірну величину, яка характеризує, наскільки вміст блоку є корисним для користувача (є релевантним по відношенню до пошукового запиту).

На рис. 1 показана тестова сторінка, на якій відмічений єдиний блок, який не є інформативним «шумом» з погляду користувача. Все інше наповнення може вважатися мало важливим (блоки *«навігація»*, *«участь»*, *«панель інструментів»* або взагалі неважливим (верхній та нижній блоки, блок *«пошук»*). Розбиття сторінки на блоки здійснюється за допомогою методу VIPS [7].

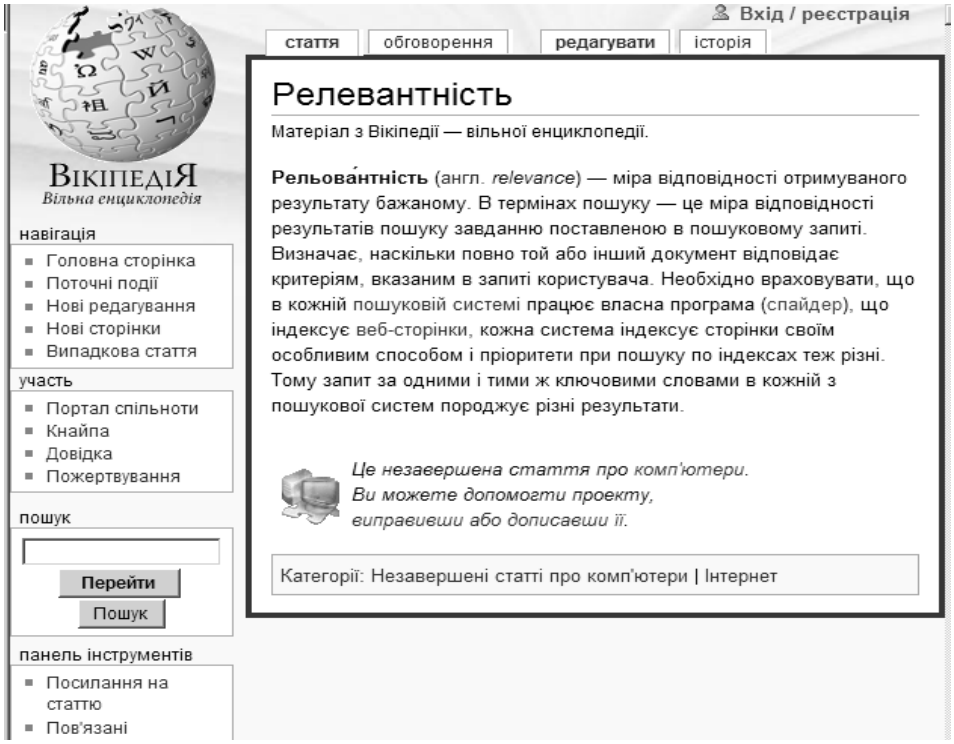

Рис. 1. Тестова сторінка і виділений важливий блок

Проаналізувавши велику кількість веб-ресурсів, виділено ряд основних типів блоків з точки зору важливості їх наповнення для користувача. Їх опис подано в таблиці 1.

Кожний блок характеризується набором властивостей, які прямо або опосередковано вказують на дані, що вони містять. Ці властивості є комбінацією морфологічних, семан-тичних та статистичних характеристик, а також враховують семантику мови HTML, що є основою для всіх веб-ресурсів.

Таблиця 1

Типи блоків

№	Наповнення блоку	Важливість блоку з точки зору інформативності
1	Інформаційний «шум» (рекламні блоки, «шапки» сайтів, інформація про сайт, форми для введення інформації, пошуку)	Неважливий
2	Блоки з навігацією, меню, зміст, список суміжних тем, список посилань на додаткові ресурси, анонси	Мало корисний
3	Основне наповнення	Корисний

В таблиці 2 подано основні ознаки кожного з типів блоків.

Таблиця 2

Ознаки, які вказують на тип блоку

Тип блоку	Ознаки
1	— велика кількість зображень; — мала кількість речень або їх відсутність; — наявність великої кількості stop-words, таких слів як «contact», «copyright», «advertice», «help», «sign», «All rights reserved» тощо; — наявність flash, gif, silverlight контенту; — велика кількість елементів керування, таких як текстові блоки, кнопки, прапорці (описуються тегами <input> та <textarea>), випадані списки (тег <select>).
2	— велика кількість гіперпосилань; — велика кількість тексту; — наявність таких тегів, як , , .
3	— наявність графічного, мультимедійного, відео контенту — велика кількість речень — велика кількість слів, що входять у речення

Властивості блоків

Список властивостей, які будуть розраховуватися для кожного блоку, наступний: кількість слів *WordsNum*; кількість речень *SentNum*; кількість слів, що входять у речення *WordsInSentsNum*; кількість посилань *LinksNum*; кількість слів, що є посиланнями *WordsAsLinksNum*; кількість зображень *ImgsNum*; кількість зображень, що є посиланнями *ImgsAsLinksNum*; довжина тексту блоку *InnerTextLength*; довжина гіпертексту блоку *InnerHtmlLength*; кількість об'єктів мультимедія *MediaObjectsNum*; кількість об'єктів керування *ControlsNum*; кількість заголовків *HeadersNum*; кількість елементів-списків *ListItemsNum*; кількість слів, що входять у списки *WordsAsListItemsNum*; величина шрифт *FontSize* і насиченість *FontWeight*.

Крім властивостей, перелічених вище, розглянемо декілька додаткових властивостей.

Середня довжина речення. Було проведено дослідження середньої довжини речень в залежності від типу блоків і знайдено, що 89,6 % речень блоків, що є основним наповненням, містять в середньому від 8 до 30 слів, в той час як даний показник містять лише 1,34 % блоків, шо не є основним блоком. Безпосередньо довжина речення на важливість блоку не впливає. Це означає, що блок, середня довжина речень якого 15 слів не обовязково є більш інформативним, ніж блок з середньою довжиною речень, що становить, наприклад, 12 слів. Основним важливим фактом є входження у визначений діапазон. Тому дана властивість буде розраховуватися за таким принципом:

$$SentAvgLengthRatio = \begin{cases} 1, \textit{ if } SentAvgLength \in [8;30], \\ 0 \textit{ otherwise.} \end{cases} \quad (1)$$

Stop-words. Stop-words — це слова або словосполучення, які містяться в блоках, які розташовані вверху або внизу сторінки. Наприклад, 95 % ресурсів внизу сторінки містять такі слова як «copyright», «all right reserved», символ «©» тощо. Інтуїтивно, велика кількість таких слів буде вказувати на дані типи блоку.

Відносні властивості блоків

Крім абсолютних значень властивостей, важливими також є їх відносні коефіцієнти, складені на основі попередніх: коефіцієнт входження слів у речення *WordsInSentRatio*; коефіцієнт слів, що є посиланнями *WordsAsLinksRatio*; коефіцієнт зображень, що є посиланнями *ImgsAsLinksRatio*; коефіцієнт слів, що є stop-words *StopWordsRatio*; коефіцієнт слів, що є списками *WordsAsListItemsRatio*.

Відносна властивість блоку розраховується за формулою

$$Ratio(prop) = \frac{WordsNum(prop)}{WordsNum}, \quad (2)$$

де *prop* — конкретна відносна властивість.

Методика проведення факторного аналізу

Для проведення досліджень було розроблено програмне забезпечення, яке зображено на рисунку 2.

Дослідження проводилися за наступною методикою: для кожного з 10 пошукових запитів з різних областей було завантажено по 20 перших веб-ресурсів, виданих пошуковою системою Google. Кожна сторінка була проаналізована за допомогою методу VIPS та розбита на інформаційні блоки.

Для кожного блоку було обчислено всі властивості, розглянуті вище. Потім експерт визначив інформативність блоків за трирівневою системою.

В результаті вище перелічених дій було отримано вибірку, що складається з близько 500 записів, інформацією про 21 фактор та експертну оцінку про важливість конкретного блоку. На основі обчислених характеристик був проведений факторний аналіз за допомогою вбудованих засобів математичного пакету SPSS 16.0.

Рис. 2. Головне вікно програми для аналізу веб-ресурсів

Аналіз отриманих результатів

В якості результату було отримано таку регресійну модель оцінки важливості інформаційних блоків:

$$F = 1{,}739 + 0{,}033 \cdot ImgsNum - 0{,}062 \cdot ImgsAsLinksNum + 0{,}087 \cdot ImgsAsLinksRatio + \\ + 0{,}002 \cdot LinksNum - 0{,}006 \cdot WordsAsLinksNum + 0{,}291 \cdot WordsAsLinksRatio + \\ + 0{,}012 \cdot SentNum + 1{,}523 \cdot SentAvgLengthRatio - 0{,}005 \cdot WordsInSentsNum + \\ + 1{,}75 \cdot WordsInSentsRatio - 0{,}164 \cdot StopWordsNum - 8{,}22 \cdot StopWordsRatio + \\ + 0{,}004 \cdot WordsNum - 0{,}003 \cdot ListItemsNum - 0{,}002 \cdot HeadersNum - \\ - 0{,}14 \cdot ControlsNum - 0{,}456 \cdot MediaObjectsNum + 3{,}712 \cdot ContentRatio + \\ + 0{,}849 \cdot WordsAsListsRatio + 0{,}105 \cdot FontSize + 0{,}002 \cdot FontWeight.$$

Основні властивості блоків, які впливають на важливість, наведено в таблиці 3.

Таблиця 3

Основні властивості блоків, що впливають на важливість

Назва властивості	Коефіцієнт рівняння регресії
StopWordsRatio	—8,22
ContentRatio	3,712
WordsInSentsRatio	1,750
SentAvgLengthRatio	1,523
WordsAsListsRatio	0,849

Висновки

В цій роботі розглянуто проблему оцінки важливості інформаційних блоків веб-сайтів. Запропоновано статистичні характеристики блоків, проведено факторний аналіз та отримано регресійну модель оцінки важливості блоків. Отримано властивості блоків, які найбільшим чином впливають на важливість.

СПИСОК ЛІТЕРАТУРИ

1. Bar-Yossef Z., Rajagopalan S. Template Detection via Data Mining and its Applications // Proceedings of Eleventh World Wide Web conference (WWW 2002). — May, 2002. Honolulu, Hawaii, USA. — 2002. — P. 580—591.

2. Lan Yi, Bing Liu. Web Page Cleaning for Web Mining through Feature Weighting // Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), August, 2003. — Acapulco, Mexico. — 2003. — P. 43—50.

3. Shian-Hua Lin, Jan-Ming Ho. Discovering Informative Content Blocks from Web Documents // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD'02). — 2002. — P. 588 — 593.

4. Yi L., Liu B. Eliminating Noisy Information in Web Pages for Data Mining // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2003), August, 2003. — Washington DC, USA. — 2003. — P. 296 — 305.

5. Gupta S., Kaiser G., Neistadt D. and Grimm P. DOM-based Content Extraction of HTML Documents // Proceedings of the welfth World Wide Web conference, May 2003. — Budapest, Hungary. — 2003. — P. 207—214.

6. Kovacevic M., Diligenti M., Gori M., Milutinovic V. Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification // Proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02), December, 2002. — Maebashi City, Japan. — 2002. — P. 250.

7. Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma. Block-based Web Search // Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, July 25 — 29, 2004. — Sheffield, UK. — 2004. — P. 456—463.

8. Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma. Learning Block Importance Models for Web Pages // Proceedings of the 13th international conference on World Wide Web, May 17 — 22, 2004. — New York, USA. — 2004. — P. 203—211.

9. Релевантність: матеріал з Вікіпедії — вільної енциклопедії. — Режим доступу: http://uk. wikipedia. org/wiki/Релевантність.

Рекомендована кафедрою комп'ютерних систем управління

Надійшла до редакції 21.10.08
Рекомендована до друку 20.11.08

Дубовой Володимир Михайлович **—** завідувач кафедри; ***Краковецький Олександр Юрійович*** **—** аспірант; ***Глонь Ольга Віталіївна*** **—** доцент.

Кафедра комп'ютерних систем управління, Вінницький національний технічний університет