УДК 127.0.0.1

ОЦІНКА РЕЛЕВАНТНОСТІ ІНФОРМАЦІЙНИХ БЛОКІВ САЙТІВ ЗА ДОПОМОГОЮ МЕТОДУ SEORANK

Дубовий В.М., Краковецький О.Ю.

Анотація. Запропоновано метод оцінки релевантності інформаційних блоків сайтів SeoRank, що грунтується на використанні правил пошукової оптимізації SEO та дозволяє підвищити вірогідність оцінки важливості інформаційних блоків.

Аннотация. Предложен метод оценки релевантности информационных блоков сайта SeoRank, который использует правила поисковой оптимизации SEO и позволяет увеличить вероятность оценки важности информационных блоков.

Abstact. In this paper method for evaluating relevance of web-page’s information blocks is proposed. It is based on search engine optimization rules and allows to increase probability of evaluation importance of information blocks.

Ключові слова: релевантність, інформаційні блоки сайтів, метод SEORANK, веб-ресурси

Вступ

Проблема створення ефективних технологій пошуку інформації в мережі Інтернет є актуальною в зв’язку з зростанням кількості веб-ресурсів та збільшенням інформаційних потреб Інтернет-користувачів. На сьогоднішній день вимоги до пошукових технологій надзвичайно високі і вимагають від останніх швидкої адаптації, розробки нових методів та удосконалення існуючих для підвищення ефективності пошуку. До сучасних задач інформаційного пошуку можна віднести задачі збільшення швидкості індексування, пошуку у реальному часі, відео-, аудіо- та графічного контенту, знаходження контенту, що дублюється, визначення першоджерел, знаходження основного контенту в автоматичному режимі та інші.

Для того, щоб веб-ресурс індексувався пошуковими системами та мав високі позиції при видачі результатів пошуку по деяким ключовим словам, він повинен відповідати певним вимогам, які відомі в літературі під назвою SEO (Search Engine Optimization) [1]. Ці правила накладають на розробників веб-ресурсів зобов’язання щодо оформлення веб-сторінок та подання матеріалу в вебі. Переважна більшість сайтів відповідають основним правилам SEO, мають чітку структуру - заголовки, абзаци та заповнені метатеги. Останні, як правило, описують блок з основним контентом і мало пов’язані з іншими інформаційними блоками. Таким чином, характеристика блоку, розрахована на основі правил SEO, може підвищити вірогідність оцінки типу блоків. Це питання детально розглядається в [2, 3]. В роботі запропоновано метод SeoRank для визначити релевантності інформаційних блоків веб-сторінки щодо її основного змісту, що представлений на веб-сторінці у вигляді інформації у метатегах. На відміну від існуючих методів оцінки релевантності веб-сторінок (наприклад, PageRank [4, 5]), запропонований метод SeoRank не розглядає релевантність інформаційних блоків щодо конкретних пошукових запитів і не враховує зовнішні параметри, такі як взаємозв’язки між ресурсами, фізичну доступність ресурсу, відповідність веб-стандартам тощо, а дає можливість оцінити інформаційні блоки в межах конкретної веб-сторінки.

Актуальність

Сучасні веб-сайти, крім основного контенту, містять велику кількість іншої інформації, що з точки зору користувачів є інформаційним шумом. Часто основний контент дублюється на різних веб-ресурсах, що, зазвичай, відрізняються лише графічним оформленням. Це стає причиною того, що в пошуковій видачі присутні різні веб-сторінки з однаковим основним контентом, що є негативною ситуацією для користувачів. Таким чином, розробка методів очищення веб-сторінок від інформаційного шуму та визначення основного контенту є **актуальними задачами**. Основою розв’язання цих задач є оцінювання релевантності інформаційних блоків сайтів до пошукового запиту.

Мета

Метою статті є розробка методу оцінки релевантності інформаційних блоків сайтів на основі критерію SeoRank.

Відповідно до мети досліджень формулюються такі задачі:

1) Аналіз основних правил пошукової оптимізації;
2) Розробка методу оцінки релевантності інформациних блоків веб-сайтів на основі правил пошукової оптимізації.

Розв’язання задачі

Основною метаінформацією веб-сторінки є її назва (title), ключові слова (meta keywords) та короткий опис (meta description). Як можна побачити з рис. 1, назва сторінки [7] «Goodbye, Google Buzz - PCWorld» відповідає заголовку основного контенту «Goodbye, Google Buzz», а ключові слова «google» і “social networks” в повній мірі присутні в основному тексті статі.

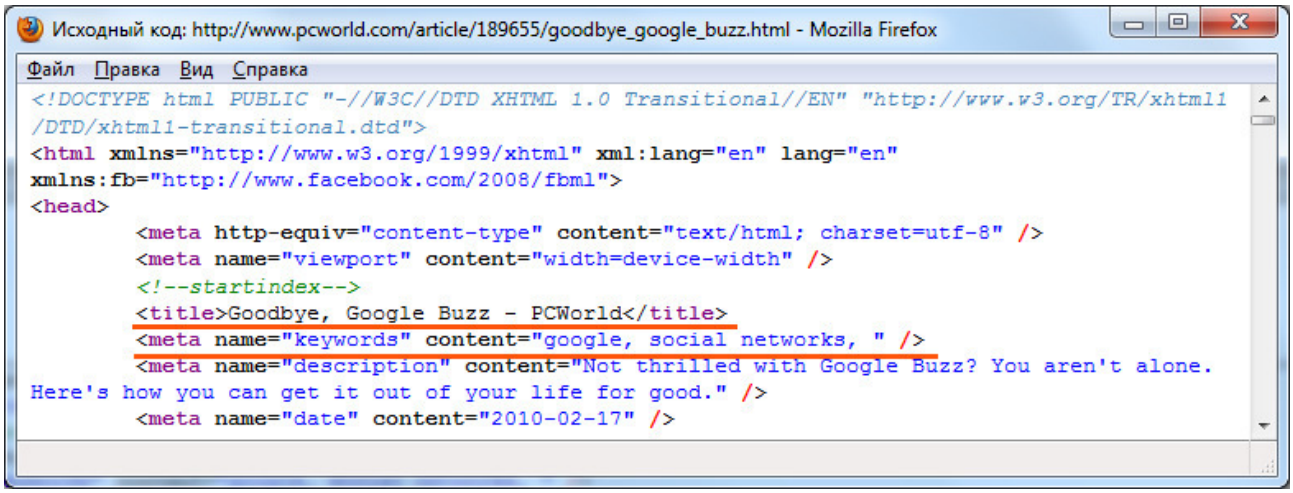

Рисунок 1. Мета інформація веб-сторінки з сайту pcworld.com

Рисунок 2 - Знімок екрану сайту pcworld.com з виділеним блоком, що є основним контентом

В [8] наведено основні параметри, що впливають на позицію сайту в пошуковій видачі Google. В таблиці 1 наведено список основних параметрів, їх важливість та тип.

Відкинувши зовнішні параметри (наприклад, кількість посилань на сторінку, адресу сторінки тощо), які не можливо оцінити на основі аналізу лише зададої веб-сторінки, та проаналізувавши інші параметри, пропонується використовувати такі характеристики для розрахунку SeoRank:

- релевантність заголовку веб-сторінки (<title>) до тексту інформаційного блоку r_1 – відношення кількості унікальних входжень слів з заголовку до загальної кількості слів з заголовку;

- релевантність ключових слів веб-сторінки (<meta keywords>) до тексту інформаційного блоку r_2 – відношення кількості унікальних входжень ключових слів до загальної кількості ключових слів;
- релевантність заголовків веб-сторінки (<headers>) до тексту інформаційного блоку r_4 – відношення кількості унікальних входжень слів з заголовків (<h1>-<h6>) до загальної кількості слів з заголовків;
- релевантність слів з опису веб-сторінки (<meta description>) до тексту інформаційного блоку r_3 – відношення кількості унікальних входжень слів з опису до загальної кількості слів з опису.

Таблиця 1

Основні параметри, що впливають на пошукову видачу

Назва параметру	Важливість	Значення важливості (%)	Тип параметру
Назви вхідних посилань	дуже важливо	73%	зовнішній
Кількість і якість вхідних посилань	дуже важливо	71%	зовнішній
Кількість доменів, що посилаються на сайт	дуже важливо	67%	зовнішній
Використання ключових слів в метатегу «Title»	дуже важливо	66%	внутрішній
Рівень довіри до сайту	дуже важливо	66%	зовнішній
Ключові слова в тексті	дуже важливо	66%	внутрішній
Ключове слово – перше слово в заголовку	дуже важливо	63%	внутрішній
Ключове слово в назві домену	дуже важливо	60%	зовнішній
Ключове слово в заголовках H1	важливо	49%	внутрішній
Інформація в метатегу «Description»	мало важлива	13%	внутрішній

Не дивлячись на те, що інформація в метатегах «Description» має низьку важливість для пошукової видачі, вона використовується для показу сніпетів – короткого опису веб-сторінки. Саме тому інформація в даному метатегу є також важливою в нашому випадку, так як вона гарантовано описує основний контент веб-сторінки (рисунки 3, 4).

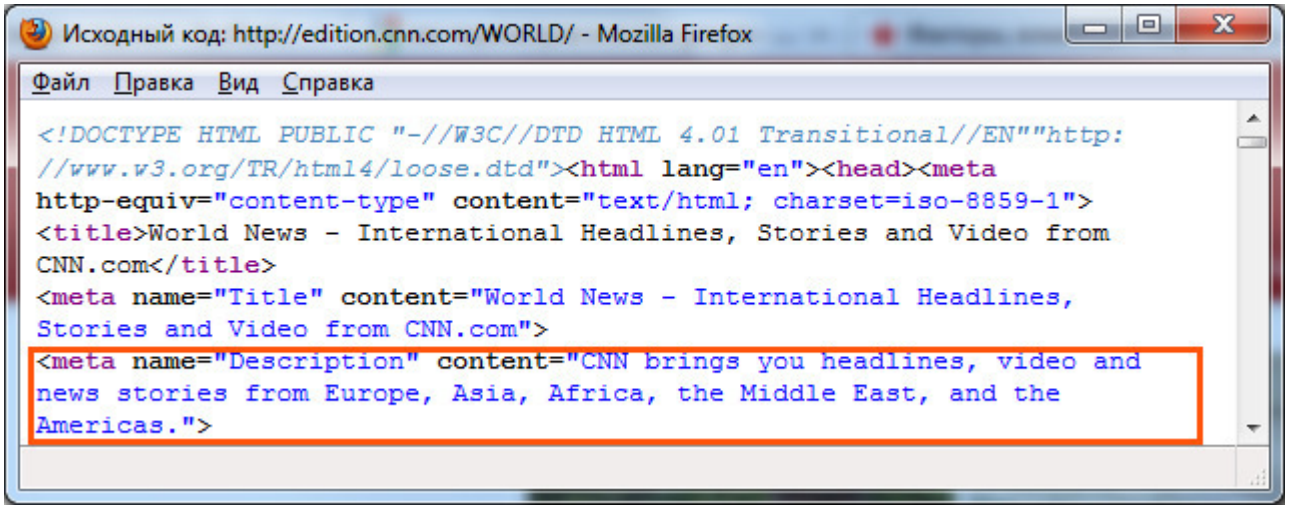

Рисунок 3 – Метатег *Description* веб-сторінки сайту CNN

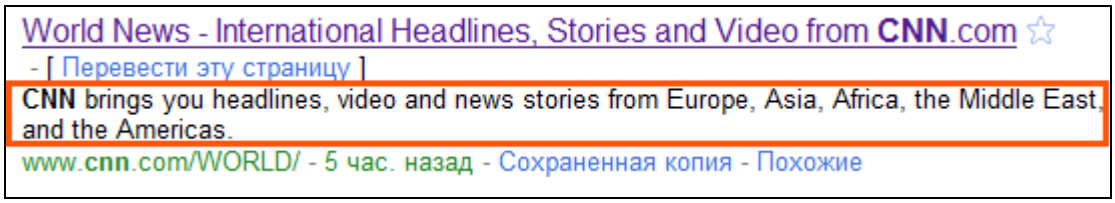

Рисунок 4 – Сніпет веб-сторінки CNN в видачі пошукової системи Google

Пропонується обчислювати SeoRank за допомогою адитивного виразу

$$SeoRank = \sum_{i=1}^{4} \alpha_i r_i ,$$

де r_i - значення параметра,

α_i - вага параметра, причому $\sum_{i=1}^{4} \alpha_i = 1$.

Ваги параметрів можна налаштовувати декількома способами:

- автоматично за допомогою нейроної мережі, де навчальними даними виступає набір типових веб-сайтів. Користувач може самостійно обрати список сайтів, що дає змогу адаптувати критерій SeoRank до особистих вподобань користувача;
- задати коефіцієнти самостійно на основі власного досвіду.

Ефективність SeoRank досліджувалась за допомогою наступної методики: було проаналізовано ряд статей відомих електронних ЗМІ, серед яких PCWorld, BBC, ComputerWorld, після чого було виділено основний контент та інші інформаційні блоки. Всього було завантажено 320 статей, на основі яких було отримано 1224 інформаційних блоків. Для кожного з блоків був розрахований критерій SeoRank. Отримані результати аналізу зведено в таблицю 2.

Таблиця 2

Результати дослідження SeoRank для α=0.25

Середнє значення SeoRank для блоку з основним контентом	Середнє значення SeoRank для всієї веб-сторінки	Середнє значення SeoRank для блоків з не основним контентом	Процент блоків з основним контентом, в яких значення SeoRank максимальне в межах веб-сторінки
0.91	0.84	0.42	94%

Як бачимо з таблиці 2, середнє значення SeoRank для основних блоків вище, ніж аналогічне значення для сторінки загалом, і набагато перевищує значення SeoRank для блоків з не основним контентом. В останній колонці наведено процент блоків, що є основним контентом, для яких значення SeoRank є макисмальним в порівнянні з іншими інформаційними блоками веб-сторінки. Таким чином, можна зробити висновок, що SeoRank є ефективною оціночною характеристикою та може використовуватися для підвищення вірогідності оцінки важливості інформаційних блоків веб-сайтів.

Висновки

1. Проаналізовано основні правила пошукової оптимізації, що можуть бути використані для визначення типів інформаційних блоків.
2. Запропоновано метод оцінки релевантності інформаційних блоків сайтів SeoRank, що грунтується на використанні правил пошукові оптимізації SEO та дозволяє підвищити вірогідність оцінки важливості інформаційних блоків.
3. Ефективність розробленого методу підтверджено на реальних даних.

Список літератури

1. Cristian Darie, Jaimie Sirovich. Professional Search Engine Optimization with ASP.NET: A Developer's Guide to SEO / Wrox. – 2007. - 410 p.

2. В.М. Дубовой, О.Ю. Краковецький, О.В. Глонь. Метод очищения веб-страниц от информационного шума. - 2008. Режим доступу: http://www.nbuv.gov.ua/portal/natural/Oeiet/2008_2/16pdf/10.pdf.
3. В.М. Дубовой, О.Ю. Краковецький, О.В. Глонь. Факторний аналіз оцінки важливості інформаційних блоків сайтів // Вісник Вінницького політехнічного інституту. - 2008. - №6. - С. 103-107.
4. Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web // Technical Report. Stanford InfoLab. - 1999. Режим доступу: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.
5. Amy N. Langville. Google's PageRank and Beyond: The Science of Search Engine Rankings // Princeton University Press. - 2006. - 234 p.
6. Оптимізація для пошукових систем // Вікіпедія – вільна енциклопедія. Режим доступу: http://uk.wikipedia.org/wiki/Оптимізація_для_пошукових_систем.
7. Goodbye, Google Buzz // Інформаційний сайт PCWorld. – 2010 р. Режим доступу: http://www.pcworld.com/article/189655/goodbye_google_buzz.html.
8. 34 фактора, влияющие на ранжирование сайта в Google SERP // Агентство интернет маркетинга StarMarketing. – 2010 р. Режим доступу: http://star-marketing.com.ua/seo/34-faktora-vliyayushhix-na-ranzhirovanie-sajta-v-google-serp/.

Відомості про авторів

Дубовой Володимир Михайлович – завідувач кафедри комп’ютерних систем управління Вінницького національного технічного університету, Хмельницьке шосе, 95, м. Вінниця, 21021, e-mail: sveta.vova.dub@gmail.com, Вінницький національний технічний університет;

Краковецький Олександр Юрійович – аспірант кафедри комп’ютерних систем управління Вінницького національного технічного університету, Хмельницьке шосе, 95, м. Вінниця, 21021; тел. (063) 26-55-367: e-mail: Alex.Krakovetskiy@gmail.com, Вінницький національний технічний університет.