УДК 004.773.2

СТРУКТУРУВАННЯ ІНФОРМАЦІЙНОГО НАПОВНЕННЯ ДЛЯ ПОКРАЩЕННЯ РАНГУ ВЕБ-ФОРУМУ

А.М. Пелещишин
Доктор технічних наук, професор*
Контактний тел.: (032) 258-25-38
E-mail: apele@ridne.net

Ю.О. Сєров
Кандидат технічних наук, асистент*
*Кафедра інформаційних систем та мереж**
Контактний тел.: (032) 258-25-38
E-mail: syerov@ridne.net

К.О. Слобода*
Кафедра прикладної лінгвістики**
E-mail: sloboda@misto.ridne.net
**Національний університет „Львівська політехніка”
вул. Бандери, 12, Львів, Україна,79013

У статті запропоновано методи структурування інформаційного наповнення Веб-форуму, методи структурування дискусій Веб-форуму відповідно до тематики розділів, семантичного та організаційного структурування інформаційного наповнення Веб-форуму та описано вплив появи небажаного інформаційного наповнення на ранг

Ключові слова: позиціонування, ІН, тематичне структурування, семантичне структурування, організаційне структурування

В статье предложены методы структурирования информационного наполнения веб-форума, методы структурирования дискуссий Веб-форума в соответствии с тематикой разделов, семантического и организационного структурирования информационного наполнения Веб-форума и описано влияние появления нежелательного информационного наполнения на ранжирование

Ключевые слова: позиционирование, ІН, тематическое структурирование, семантическое структурирование, организационное структурирование

In this article methods of Web-forum’s content structuring are developed. Methods of Web-forum’s threads structuring according to their theme are proposed, methods of semantics structuring are developed, and methods of organizational structuring are suggested, influence of objectionable content is described

Keywords: positioning, content, thematic structuring, semantic structuring, organizational structuring

1. Вступ

World Wide Web є цілісною системою глобального масштабу, елементи якої тісно взаємодіють між собою. Хоча кількість Веб-форумів зростає щоденно, значна їх частина може не мати жодного зацікавленого відвідувача протягом тривалого часу, а велика частина – незначну кількість відвідувачів, що робить існування Веб-сайту беззмістовним [4]. Отже, невдале позиціонування сайту в глобальному інформаційному середовищі є одною з найважливіших причин невдачі реалізації Інтернет-проекту.

Результати досліджень з позиціонування Веб-форумів є важливими для широкого кола фахівців з організації та побудови Веб-форумів, як такі, що повинні бути використані в переважній більшості реальних Веб-проектів та забезпечувати їхню успішність та ефективність.

2. Постановка проблеми

Попри велику кількість наявних нині Веб-форумів, лише окремі є ефективними. Переважна більшість Веб-форумів є неефективними і не виправдовують сподівань власників. Така ситуація, насамперед, пояснюється відсутністю формалізованих методів та засобів побудови ефективних Веб-спільнот.

Надзвичайно важливим завданням, що постає в процесі структурування інформаційного наповнення (ІН) Веб-форуму є відстеження появи небажаного ІН, оскільки його поява призводить до викривлення тематики та, відповідно, спотворення семантичного ядра, що в свою чергу призводить до зниження ефективності позиціонування Веб-форуму в глобальному інформаційному середовищі World Wide Web. Тому доцільно паралельно з розробленням методів структурування ІН розробляти і методи виявлення небажаного ІН.

3. Аналіз останніх досліджень

Завдання підвищення ефективності позиціонування Веб-форумів в глобальній інформаційній системі World Wide Web є надзвичайно актуальним. Останніми дослідженнями вирішення проблеми покращання позиції Веб-форумів є праці професора А.М. Пелещишина про позиціонування сайтів в глобальному інформаційному середовищі (2007), І.С. Ашманова та А.А. Іванова про просування сайтів у пошукових системах (2007), Н.В. Євдокімова про основи оптимізації ІН (2007).

В цих працях розглянуто загальні методи покращання позиції сайту, але не розглянуто методів структурування ІН, зокрема методів структурування відповідно до тематики розділів Веб-форуму, семантичного та організаційного структурування ІН Веб-форуму.

4. Цілі статті

1. Розроблення методів структурування ІН Веб-форуму.
2. Розроблення методів структурування дискусій відповідно до тематики розділів Веб-форуму.
3. Розроблення методів семантичного структурування ІН Веб-форуму/
4. Розроблення методів організаційного структурування ІН.
5. Опис впливу появи на Веб-форумі небажаного ІН на ефективність позиціонування.

5. Основна частина

1. Структурування дискусій відповідно до тематики розділів Веб-форуму.

Уведемо формальний опис Веб-форуму з погляду його семантики [5].

Кожен розділ Веб-форуму містить повідомлення та дискусії, що стосуються однієї або декількох предметних областей.

Нехай тематика ***Theme*** – множина тем форуму:

$\boldsymbol{Theme} = \left\{\mathrm{Theme}_i\right\}_{i=1}^{N^{(Tm)}}$, де $N^{(Tm)}$ – кількість елементів множини ***Theme***, усі теми форуму.

Кожна з тем форуму описується множиною ключових слів:

$\boldsymbol{Keyword} = \left\{\mathrm{Keyword}_i\right\}_{i=1}^{N^{(Kw)}}$, де $N^{(Kw)}$ – кількість елементів множини ***Keyword***, тобто кількість ключових слів.

Визначимо теми розділів Веб-форуму як множину пар ключових слів та вагових коефіцієнтів, які задають важливість ключового слова для цієї теми:

$$\mathrm{Theme}_i = \left\{\left\langle \mathrm{Keyword}_j, w_{ij}\right\rangle\right\}_{j=1}^{N^{(KwTheme_i)}}, \quad (1)$$

де $\mathrm{Keyword}_j \in \boldsymbol{Keyword}$ – ключове слово з множини ключових слів;

w_{ij} – вага ключового слова $\mathrm{Keyword}_j$ у темі Theme_i;

$N^{(KwTheme_i)}$ – кількість ключових слів у темі Theme_i.

Для кожного повідомлення необхідно визначити ключові слова із множини ключових слів. Поставлені у відповідність повідомленню ключові слова утворюють множину ключових слів повідомлення, яку позначатимемо як $\mathrm{KeywordList}(\mathrm{Post}_i)$, $\mathrm{KeywordList}(\mathrm{Post}_i) \subseteq \boldsymbol{Keyword}$.

Сукупність усіх ключових слів повідомлень дискусії з урахуванням їх кількостей утворює сукупність ключових слів дискусії, яку визначатимемо як:

$$\mathrm{KeywordList}(\mathrm{Thread}_i) = \left\{\left\langle \mathrm{Keyword}_j, n_{ij}\right\rangle\right\}_{j=1}^{N^{(KwThread_i)}}, \quad (2)$$

де $\mathrm{Keyword}_j \in \bigcup_{\mathrm{Post}_k \in \mathrm{Thread}_i} \mathrm{KeywordList}(\mathrm{Post}_k)$ – ключові слова, які трапляються у повідомленнях дискусії Thread_i;

n_{ij} – кількість повідомлень дискусії Thread_i, у яких виявлене ключове слово $\mathrm{Keyword}_j$;

$N^{(KwThread_i)}$ – кількість ключових слів дискусії Thread_i.

Тоді, враховуючи (1), (2), міра відповідності дискусії Thread_i до теми Theme_p визначена як

$$\mu(\mathrm{Thread}_i, \mathrm{Theme}_p) = \frac{\sum_{j=1}^{N^{(KwThread_i)}} w_{pj} n_{ij}}{\sum_{j=1}^{N^{(KwTheme_p)}} w_{pj}}, \quad (3)$$

де $w_{pj} = w(\mathrm{Theme}_p, \mathrm{Keyword}_j)$ – вага ключового слова $\mathrm{Keyword}_j$ у темі Theme_p;

$n_{ij} = n(\mathrm{Thread}_i, \mathrm{Keyword}_j)$ – кількість повідомлень дискусії Thread_i, у яких виявлене ключове слово $\mathrm{Keyword}_j$.

Умова тематичної структурованості Веб-форуму – міра відповідності дискусії до теми розділу (3), в якому вона знаходиться, є не меншою, ніж міра відповідності цієї дискусії до тем інших розділів Веб-форуму. Нехай дискусія Thread_i знаходиться в розділі Part_p, $\mu(\mathrm{Thread}_i, \mathrm{Theme}_p)$ – міра відповідності дискусії Thread_i до тематики Theme_p розділу Part_p, тоді ІН є тематично структурованим, якщо для всіх дискусій Thread_i та всіх розділів Part_j з тематикою Theme_j виконується умова $\mu(\mathrm{Thread}_i, \mathrm{Theme}_p) \geq \mu(\mathrm{Thread}_i, \mathrm{Theme}_j)$.

2. Організаційне структурування ІН Веб-форуму

Початкову структуру ІН Веб-форуму створює адміністратор. Залежно від обраної тематики адміністратор створює назви розділів. У процесі функціонування часто виникають ситуації нерівномірного приросту кількості дискусій у різних розділах Веб-форуму. Певні розділи стають популярнішими і кількість дискусій у них збільшується швидше, ніж в інших. Така ситуація спричинює зниження якості структури ІН Веб-форуму і ускладнює процес пошуку інформації. Тоді виникає потреба зміни структури ІН Веб-форуму з метою рівномірного розподілу дискусій між розділами.

Алгоритм організаційного структурування Веб-форуму полягає у виконанні таких кроків:

1. Виконується перевірка усіх розділів Веб-форуму на відповідність умові збалансованості.

2. Якщо певний розділ замалий, тобто кількість дискусій розділу Part_i менша ніж n, то адміністратору пропонується на цьому ж рівні утворити новий розділ

у якому об'єднати дискусії цього розділу з дискусіями іншого, близького за тематикою. Для цього потрібно назвати новий розділ, після чого об'єднати множини дискусій та ключових слів тематик двох вихідних розділів.

3. Якщо певний розділ завеликий, тобто кількість дискусій розділу $Part_i$ більша ніж $2n$, то адміністратору Веб-форуму пропонується рівнем вище утворити два нові розділи за умови, що наявна кількість розділів на вищому рівні не перевищує максимально допустиму. Якщо наявна кількість розділів на вищому рівні перевищує допустиму, то поділ потрібно здійснити на одному з вищих рівнів, на якому ця умова виконується. Процес поділу розділу відбувається шляхом поділу множини ключових слів тематики цього розділу, і як наслідок, дискусій цього розділу на дві підмножини.

4. Застосовуються кроки 1-3 до тих пір, поки хоч один розділ Веб-форуму порушуватиме умову збалансованості.

3. Семантичне структурування ІН Веб-форуму

Семантичне ядро – це список ключових слів і словосполучень, які відображають тематику Веб-форуму і використовуються в його інформаційному наповненні [3]. Для семантичного структурування ІН Веб-форуму доцільно поділити списки ключових слів та фраз за семантичними ознаками на відповідні підтеми, в кожну з яких додавати відповідні ключові слова та фрази. Основними показниками семантичного ядра форуму є обсяг семантичного ядра та частотність запитів.

Ключовим словом у глобальному інформаційному середовищі вважається як одне слово природної мови, так і словосполучення з кількох слів [2]. Тому доцільно розділити два поняття - ключове слово (одне слово, або логічно нерозривне коротке словосполучення) та ключова фраза (сукупність ключових слів та словосполучень). Отже, семантичне ядро форуму визначається його тематикою і є множиною ключових слів усіх тем форуму.

Обсяг семантичного ядра залежить від кількості ІН форуму.

Частотність запиту - це міра його популярності, яка показує скільки разів користувачі вводили його в пошукову систему.

На етапі створення семантичного ядра проводиться аналіз ключових слів і фраз, які використовуються або ймовірно можуть використовуватися користувачами для знаходження інформації яка відповідає тематиці форуму. Сторінки форуму, оптимізовані під певні ключові слова будуть стабільно отримувати постійних відвідувачів. До семантичного ядра сайту необхідно внести всі ключові слова та вирази, які користувачі задають в якості запитів пошуковим машинам.

На цьому етапі оцінюється популярність кожної ключової фрази, ступінь її відповідності інтересам цільової аудиторії і на основі цих даних робиться висновок про включення фрази в семантичне ядро форуму.

4. Вплив небажаного ІН на позиціонування Веб-форуму

Поява на Веб-форумі небажаного ІН призводить до спотворення його тематики, а відповідно семантики і семантичного ядра та суттєво знижує ефективність його позиціонування. Небажаним для Веб-форуму інформаційним наповненням є ненормативна лексика, вислови, що ображають людей на расовому, етнічному, релігійному ґрунті, ІН, яке порушує авторські права, шкідливе програмне забезпечення, інформація, що може спричинити виникнення конфліктів тощо. Поява у дописах користувачів небажаного ІН призводить до появи небажаних ключових слів в семантичному ядрі Веб-форуму. Пошукові системи індексують ці ключові слова і включають їх до опису загальної тематики ресурсу. Це призводить до незворотного процесу подальшої появи іншого небажаного ІН.

Для виявлення в інформаційному наповненні Веб-форуму небажаного ІН для власників або адміністраторів необхідно фільтрувати його. Особливу увагу слід приділити зовнішнім посиланням на Веб-форум з інших ресурсів, а також модерувати прикріплені користувачами файли. Основним завданням фільтрації ІН Веб-форуму є виявлення в дописах користувачів слів, попередньо визначених в списку заборонених та автоматична заміна цих слів на набір символів без певного значення або напис «цензуровано».

Висновок

У статті представлено методи структурування ІН Веб-форуму відповідно до тематики розділів, семантичного структурування ІН Веб-форуму, організаційного структурування ІН Веб-форуму, описано вплив появи на Веб-форумі небажаного ІН на ефективність позиціонування Веб-форумів.

Література

1. Ашманов И. С. Продвижение сайта в поисковых системах / И. С. Ашманов, А. А. Иванов – М. : Вильямс, 2007. –304 с.
2. Пелещишин А. М. Позиціонування сайтів у глобальному інформаційному середовищі : Монографія. – Львів : Видавництво Національного університету «Львівська політехніка», 2007. – 260 с.
3. Пелещишин А.М., Слобода К.О. Модифікація контенту для ефективного позиціонування форуму в середовищі WWW / Збірник наукових праць Людина. Комп'ютер. Комунікація: за ред. Ф.С.Бацевича. – Львів: Видавництво Національного університету «Львівська політехніка», 2010. – С.84–87.
4. Сєров Ю. О. Аналіз комунікативних процесів у Веб-спільнотах середовища Веб 2.0 / Ю. О. Сєров, А. М. Пелещишин, К. О. Слобода // Східно-Європейський журнал передових технологій. – Харків, 2009. – №1/2 (37) /2009. – С.38–41.
5. Peleschyshyn A. Web-Forum Semantics and Thematic Modelling / A. Peleschyshyn, R. Kravets, Yu. Syerov, K. Sloboda // Proceedings of the 5th International Scientific and Technical Conference «CSIT'2010». – Lviv: Publishing House Vezha&Co., 2010. – P.90–92.