УДК

А.М. Пелещишин
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

МЕТОДИ ВИЗНАЧЕННЯ ТА ОПТИМІЗАЦІЇ ТЕМАТИКИ САЙТУ

Розглянуто проблеми визначення та оптимізації тематики Веб-сайту з огляду на його популярність та інші критерії ефективності. Запропоновано підхід до вирішення проблеми, що базується на запитах до пошукових машин.

This paper considers main problems of web site thematic definitions and optimization. Some approaches to resolve these problems are proposed.

Постановка проблеми у загальному вигляді

Актуальність проблеми визначення тематики інформаційних ресурсів World Wide Web є задачею, яка обумовлена рядом факторів, найважливішими з яких є:

- високий користувацький попит на сервіси пошуку, класифікації та аналізу інформаційних ресурсів WWW;
- потреба власників сайтів у точному відображенні тематики сайту в сервісах пошуку, класифікації та аналізу інформаційних ресурсів WWW.

Задача визначення тематики інформаційних ресурсів Інтернету неодноразово розглядалася як з теоретичної точки зору, так і зі спробами реального впровадження. Проте дослідження у даній сфері мають односторонній характер – це автоматизоване (частково чи повністю) визначення тематики сайту для використання надалі в алгоритмах пошуку інформації в WWW та її аналізу.

Отже, основні дослідження у даній сфері проводяться дослідницькими групами, що працюють над створенням чи вдосконаленням глобальних інформаційних сервісів – пошукових систем, каталогів, систем Інтернет-реклами, порталів.

Аналіз останніх досліджень

World Wide Web є активним середовищем, яке складається з мільйонів сайтів, кожен з яких має окремого власника і, відповідно, власні цілі, які не завжди збігаються з цілями вказаних глобальних проектів.

Як наслідок, результати, що отримуються автоматизованими сервісами, не є повноцінним описом тематики сайту. Не випадково гаслом найбільшого каталогу сайтів ODP (Open Directory Project) є "**Humans do it better**" ("люди роблять це краще") [1].

Значна кількість досліджень, дотичних до задачі визначення тематики сайту, проводиться фахівцями з Інтернет-реклами, просування сайту в Інтернет, оптимізації сайтів під пошукові машини. Проте ці дослідження мають лише практичний характер і часто навіть розглядаються (в першу чергу власниками глобальних сервісів) як ворожі чи шкідливі для глобального середовища.

Стає очевидним, що розв'язання складних задач з опрацювання інформаційних ресурсів WWW є можливим лише за умови врахування як інтересів звичайних користувачів та глобальних сервісів (які беруть на себе місію з представлення інтересів користувачів), так і власників сайтів, які власне і формують WWW. Як наслідок, у спільноти WWW виникає певне розуміння щодо спільних цілей, які стоять перед надавачами та користувачами різних послуг. З'являється чітке розмежування двох альтернативних підходів до популяризації сайтів їхніми власниками серед користувачів WWW:

- **Технології спаму** ("чорна оптимізація") – спам пошукових машин і каталогів, спам інтерактивних сервісів, заплутування та скерування навігації користувача, поштовий спам, дезінформація користувачів.

• **Технології оптимизації тематики сайту** (“біла оптимізація”) – вибір оптимальної тематики сайту, вдосконалення відображення тематики сайту, дотримання правил Інтернет-спільноти.

Даний поділ підтверджується в появі організацій [2], що забезпечують єдині “правила гри” для глобальних сервісів навігації (пошукових машин, каталогів, систем Інтренет-реклами) та звичайних сайтів, у проведенні спільних конференцій та нарад між фахівцями з “білої оптимізації” та представниками глобальних сервісів [3], складанні спеціальних “кодексів білої оптимізації” [4, 5], організації служб виявлення фактів “чорної оптимізації”.

Виділення з невирішених раніше частин загальної проблеми

Розглядається проблема подання та оптимізації тематики сайту з точки зору власника сайту. Ця проблема для власника сайту є критично важливою, і від успішного її вирішення значною (а деколи і вирішальною) мірою залежить успішність Веб-проекту.

Проте ця задача для власників сайтів має розв’язуватися з врахуванням усталених норм глобального середовища та існуючих у ньому правил та обмежень (“біла оптимізація”). За точку відліку при розв’язанні задачі треба брати існуючі в WWW методи визначення тематики сайту користувачами сайту та глобальними сервісами, що забезпечують навігацію користувачів по WWW. Сама задача для власника сайту формулюється як **побудова сайту, який правильно, точно та ефективно відображає вибрану оптимальну тематику для досягнення поставлених перед сайтом цілей в умовах існуючого глобального середовища.**

Розглянемо детальніше сформульовану задачу. Основними тезами в цьому формулюванні є:

• сайти можуть неправильно чи неточно відображати тематику, що є бажаною;

• тематика сайту може відображатися неефективно (особливо з врахуванням активного конкурентного середовища WWW);

• вибрана тематика сайту може бути неоптимальною для досягнення цілей, що були поставлені перед сайтом, вона може потребувати уточнення чи модифікації;

• тематика має визначатися в умовах реально функціонуючого глобального середовища, з визначеними “правилами гри” та високою інертністю, що виключає реальну можливість появи та впровадження власниками принципово нових методів і служб визначення тематики сайту.

Цілі статті

Головними цілями цієї статті є:

• Дослідження та формалізація де-факто існуючих методів опису тематики сайту в середовищі World Wide Web;

• Побудова формальних підходів до оптимізації тематики сайту, що базуються на методах опису тематики;

• Верифікація підходів до оптимізації сайту на конктретному реальному прикладі.

Основний матеріал

Однією з базових задач, що постають ще на ранніх етапах побудови сайту, є чітке визначення тематики сайту. Від визначення показників тематичності сайту залежить більшість найважливіших показників ефективності сайту системного класу, зокрема його популярність, рівень та якість подання в World Wide Web. Ці показники, у свою чергу, є лише відображенням рівня успішності виконання основної задачі, що ставиться власниками сайту (наприклад, економічна віддача від сайту, суспільний резонанс тощо).

Від визначення тематики сайту значною мірою можуть залежати також і технічні рішення, що будуть використовуватися при побудові сайту, зокрема програмні засоби сайту, мережева та серверна платформи сайту.

Помилки при визначенні тематики сайту на етапі його розробки та впроваження в World Wide Web мають системний характер. Виправлення таких помилок після початку функціонування сайту може бути достатньо витратним та трудомістким і фактично вимагати повернення до ранніх етапів

побудови сайту. У деяких випадках імовірною є необхідність у повному рестарті проекту (включно з відмовою від доменного імені сайту).

Серед помилок щодо визначення тематики сайту доцільно виділити два основні типи:

- Помилки розробників щодо вибору тематики сайту взагалі (наприклад, замість сайту про українську історію розробляється сайт про історію Східної та Центральної Європи);
- Помилки розробників щодо відображення тематики на самому сайті та в його оточенні в глобальній системі World Wide Web (наприклад, сайт про українську історію ідентифікується користувачами та автоматизованими сервісами WWW як сайт про історію Східної та Центральної Європи).

Зазначимо, що правильне та ефективне визначення тематики сайту повинне базуватися на методах відображення тематики сайту та пов'язаних з ними методах визначення системних показників сайту взагалі.

Дослідимо основні методи відображення тематики сайту та відповідні їм методи уточнення і корегування тематики сайту.

Під **тематикою сайту (або реальною тематикою сайту)** будемо розуміти тематику сайту, що відображається сайтом та його оточенням. **Тематику сайту**, що закладається у нього власниками, вважатимемо **бажаною**.

Як уже було сказано вище, бажана тематика сайту не завжди збігається з реальною.

Навігаційні методи визначення тематики сайту

Оскільки тематика сайту збігається з тематикою його аудиторії, методи опису тематики сайту повинні базуватися на моделі аудиторії сайту та поведінки її членів.

Основними методами, що базуються на моделі тематичної зацікавленості аудиторії сайту, є методи, що орієнтуються на визначення можливих шляхів навігації по WWW, якими може потрапити на сайт користувач.

Виділимо такі методи навігації користувача по WWW:

- Пряма навігація користувача (заходи на сайт без використання гіперпосилань шляхом набирання адреси в рядок броузера).
- Перехід користувача з каталогу сайтів на вибраний сайт;
- Перехід користувача з тематичних рубрик порталів на вибраний сайт;
- Перехід користувача з з результатів видачі пошукової системи на вибраний сайт;
- Перехід користувача з інших сайтів (відмінних від вказаних вище типів) на вибраний сайт за гіперпосиланнями;
- Перехід з глобальних систем, відмінних від World Wide Web, що функціонують в мережі Інтернет (системи конференцій, системи розсилки електронної пошти тощо).

Визначення тематики сайту на основі рубрик порталів та каталогів

Опис тематики сайту на WWW-порталах та каталогах базується на рубриках. Як і в питанні визначення аудиторії сайту, в питанні визначення тематичних рубрик сайту є два аспекти – реальні рубрики, які описують тематику сайту та ті рубрики, які, на думку власників сайту, повинні описувати його тематику.

Методів організації рубрик може бути сформульовано достатньо багато, і вони можуть достатньо сильно відрізнятися між собою.

Проте при визначенні тематики сайту необхідно базуватися на усталених у системі World Wide Web головних методах побудови тематичних рубрикаторів. Виділимо такі типи організації каталогів та порталів:

- Одно- або дворівневі каталоги (з фіксованою кількістю рівнів);
- Багаторівневі ієрархічні каталоги;
- Багаторівневі ієрархічні каталоги зі взаємозв'язками між рубриками;
- Фасетні каталоги.

Тематику сайту фактично доводиться описувати на основі кожного з методів рубрикації зокрема.

Повна назва розділу є головним фактором, за допомогою якого каталоги визначають тематику сайту. Проте каталоги можуть додавати інформацію про тематику сайту й іншими способами, серед них головними є:

- Текстовий опис сайту в каталозі;
- Сайти-сусіди за розділом.

Текстовий опис описує тематику сайту за допомогою ключових слів (див. далі).

Сайти-сусіди за розділом є важливими для випадків, коли вміст каталога використовується спеціалізованими WWW-сервісами та клієнтськими програмами (зокрема додатками до браузерів), які допомагають користувачу у навігації за принципом "За даною темою дивіться також...".

Одно- та дворівневі каталоги

Рубрикатори з фіксованою кількістю рівнів (як правило, не більше двох) є найпростішим методом організації WWW-каталогів. Такі каталоги є характерними для порталів та інших сайтів з великою підбіркою корисних посилань.

Крім цього, ще одним (і вкрай важливим) видом рубрикації, що доцільно описувати як однорівневий каталог, є рубрикація статей на сайтах новин та інших Інтернет-ЗМІ.

При описі тематики сайту розділами каталогів даного типу достатньо визначити головну тему сайту (наприклад, "Політика", "Наука", "Технології") для однорівневого опису або головну та уточнену теми (наприклад, "Наука/Математика", "Технології/Інтернет").

Багаторівневі ієрархічні каталоги

Багаторівневні ієрархічні каталоги організовані на базі деревоподібної структури розділів. Кожен розділ каталога може містити підрозділи. У каталогах такого типу не регламентується глибина вкладень розділів.

Кожен розділ такого каталога може містити як підрозділи, так і безпосередньо посилання на сайти. Повна назва розділу каталога формується за класичним принципом:

"Повна назва" = "Назва надрозділу" + "коротка назва розділу"

Сайт, розміщений у певному розділі, вважається таким, що повна назва розділу його описує найточніше – достатньо повно і водночас детально. Якщо опис за допомогою назви розділу є недостатньо деталізованим, то сайт повинен бути переміщеним у більш детальний підрозділ розділу. Якщо опис сайту занадто детальний і не охоплює усю тематику сайту, то сайт повинен бути переміщений вище.

Зазначимо також, що у каталогах даного типу не обов'язково присутні рубрики, що відповідають кожній можливій комбінації ознак сайту, отже, необхідно шукати серед існуючих такий розділ, який найповніше описує тематику сайту.

Фасетні каталоги

Фасетні каталоги володіють найпотужнішими засобами класифікації ресурсів WWW порівняно з іншими видами каталогів. Фасетні каталоги оргнаізовані за принципом кількох незалежних класифікацій сайту, що здійснюються одночасно за кількома ознаками.

Фасетні каталоги забезпечують багатоаспектний опис ресурсів WWW та можливість існування достатньо компактної структури рубрик, які охоплюють усі можливі тематики сайтів. Фактично фасетний каталог є багатовимірним гіперкубом інформації, вимірами якого є класифікатори, а наповненням кожної атомарної комірки – окремі сайти.

Зазначимо, що при організації класифікаторів фасетного каталогу допускається існування як лінійних, так і ієрархічних класифікаторів (зокрема, для опису географічних характеристик сайту та тематичного спрямування сайту). Це, безсумнівно, ускладнює організацію каталогу та користування ним, проте забезпечує достатньо потужні механізми класифікації за складними ознаками. Кожен окремо ієраріхічний класифікатор можна вважати простим ієрахічним каталогом. Лінійні класифікатори можна розглядати як однорівневі каталоги.

Отже, рубрикатор фасетного каталогу є декартовим добутком рубрикаторів кількох незалежних ієрархічних та лінійних каталогів відносно невеликого розміру та побудованих за незалежними класифікаційними ознаками.

Визначення тематики сайту на основі ключових фраз та пошукових запитів

Ключові слова є традиційним методом визначення тематики текстової інформації як в електронному середовищі, так і в звичайних публікаціях (зокрема наукових). Відповідно і в мережі Інтренет цей метод зберігає свою актуальність у дещо модифікованому вигляді.

Найважливішим методом визначення тематики сайту на основі навігаційного підходу є використання пошукових запитів. Це викликано тим, що домінуючим сьогодні методом навігації користувача по системі World Wide Web є переходи на сайти з результатів пошуку інформації пошуковими машинами. За своєю суттю пошукові запити являють собою набори ключових слів з можливими додатковими обмеженнями.

При формуванні методів опису тематики сайту ключовими фразами важливо враховувати методи опису тематики сайту через його аудиторію. У такому разі модель тематики сайту, що базується на ключових фразах, близька до моделі тематики сайту, що базується на аудиторії сайту.

Кожна ключова фраза відповідає певній групі користувачів сайту. Відповідно, для кожної ключової фрази визначається ряд спеціальних величин, які відповідають аналогічним показникам груп користувачів сайту.

Для кожного пошукового запиту визначимо міру корисності, що відображає математичне сподівання міри досягення власниками певної цілі щодо відвідувача, який потрапив на сайт за ключовим словом.

$$\mathrm{Uf}(Kw_i) = \sum_{j=1}^{N_{Tr}} \mathrm{Pr}^{(\mathrm{Uf})}\left(\mathrm{U}(Qs_i), Tr_j\right)\mathrm{Uf}\left(Tr_j^{(\mathrm{Uf})}\right)$$

де $\mathrm{Uf}(Qs_i)$ – корисність і-го пошукового запиту для сайту; $\mathrm{Uf}\left(Tr_j^{(\mathrm{Uf})}\right)$ – корисність j-ї цілі сайту; $\mathrm{U}(Qs_i)$ – множина користувачів, що потрапила на сайт за запитом Qs_i; $\mathrm{Pr}^{(\mathrm{Uf})}\left(\mathrm{U}(Qs_i), Tr_j\right)$ – імовірність досягнення j-ї цілі сайту щодо користувача сайту, який потрапив на сайт за запитом Qs_i.

Цей показник відповідає показнику корисності групи користувачів сайту, де групування відбувається за навігаційною ознакою – ключовій фразі, яка привела на сайт користувача.

Аналогічно для окремого пошукового запиту визначимо міру відповідності, що відображає математичне сподівання міри досягення користувачами певної цілі щодо сайту, на який він потрапив за даною ключовою фразою.

$$\mathrm{It}(Kw_i) = \sum_{j=1}^{N_{Tr}} \mathrm{Pr}^{(\mathrm{It})}\left(\mathrm{U}(Qs_i), Tr_j\right)\mathrm{Uf}\left(Tr_j^{(\mathrm{It})}\right)$$

де $\mathrm{Uf}(Qs_i)$ – корисність і-го пошукового запиту для сайту; $\mathrm{It}\left(Tr_j^{(\mathrm{It})}\right)$ – корисність j-ї цілі сайту; $\mathrm{U}(Qs_i)$ – множина користувачів, що потрапила на сайт за запитом Qs_i; $\mathrm{Pr}^{(\mathrm{Uf})}\left(\mathrm{U}(Qs_i), Tr_j\right)$ – імовірність досягнення користувачем сайту, що потрапив на сайт за запитом Qs_i j-ї цілі щодо сайту.

Даний показник відповідає показнику зацікавленості групи користувачів сайту.

Оптимізація опису та подання тематики сайту

Комплексна задача оптимізації сайту поєднує в собі задачі:

- визначення оптимальної тематики сайту;
- ефективного та якісного подання вибраної тематики на сайті.

Розв'язання цих задач вимагає введення ряду додаткових супутніх характеристик до кожного з показників тематики сайту.

Такими супутніми характеристиками є:

- Кількісні характеристики, що описують розмірність та якість аудиторії, зацікавленої в певній тематиці;
- Кількісні характеристики, що описують конкурентне середовище сайтів за тематикою;

• Порядкові характеристики, що описують позиціонування сайтів за тематикою;

• Кількісні характеристики, що описують витратність позиціонування сайту за певною тематикою.

Першою окресленою проблемою з визначення тематики сайту власниками є проблема неправильного чи неточного подання тематики в різних моделях навігації потенційної аудиторії сайту.

Виправлення цієї помилкової ситуації має проводитися власниками сайту в таких напрямках:

• Зміна текстів сайту;

• Зміна структури та технічної реалізації сайту;

• Зміна оточення сайту;

• Зміна параметрів цільової та контактної реклами.

Другу проблему розглянемо детальніше.

Модифікація тематики сайту

Часто досягнення цілей, що поставлені перед сайтом, вимагає не тільки виправлення помилок та вдосконалення подання тематики сайту, але й зміни тематики сайту загалом.

Така ситуація виникає, коли:

• тематика сайту не повністю відповідає визначеним цілям сайту;

• тематика сайту не користується достатньою популярністю серед користувачів WWW;

• тематика сайту є висококонкурентною, що ставить під сумнів досягення визначених цілей.

Виправити цю ситуацію можна:

• узагальненням тематики сайту;

• зміщенням тематики сайту;

• уточненням тематики сайту.

Тематика сайту традиційно модифікується на основі експертних оцінок та досвіду. Проте ця задача може бути частково формалізованою та описаною на основі відповідних математичних апаратів. Така формалізація дозволяє будувати формальні алгоритми оптимізації тематики сайту та може використовуватися при розробці спеціалізованих програмних засобів проектування та реінжинірингу сайтів.

Ключовою проблемою, що постає при оптимізації тематики сайту, є отримання необхідної базової інформації – без вхідних даних застосування формальних підходів втрачає сенс. Тому побудова методів отримання необхідних даних є важливою задачею, що входить у загальну комплексну проблему визначення та оптимізації тематики сайту.

Математична модель оптимізації тематики сайту

Нехай:

$Th^{(Alt)} = \{Th^{(j)}\}_{j=1}^{N_{Th}^{(Alt)}}$ – множина альтернативних тематик сайту.

ц $Th^{(j)} = \{Th_i^{(j)}\}_{i=1}^{N_{Th_j}}$ – тематика сайту – множина тем, що описують тематику сайту. Зазначимо, що описувати можна будь-яким зазначеним вище методом.

$\mathrm{Uf}\left(Th_i^{(j)}\right)$ – корисність і-ї теми для сайту;

$\mathrm{Fr}\left(Th_i^{(j)}\right)$ – частота звернення до і-ї теми користувачами WWW;

$$\mathrm{Uf}\left(Th^{(j)}\right) = \sum_{i=1}^{N_{Th_j}} \mathrm{Fr}\left(Th_i^{(j)}\right)\mathrm{Uf}\left(Th_i^{(j)}\right)\mathrm{Pr}^{(Jump)}\left(Th_i^{(j)}\right),$$

де $\mathrm{Pr}^{(Jump)}\left(Th_i^{(j)}\right)$ – імовірність переходу на сайт користувача, що звернувся до і-ї теми.

Величина $\mathrm{Pr}^{(Jump)}\left(Th_i^{(j)}\right)$ залежить від характеристик глобального конкурентного середовища, у якому функціонує сайт. Користувач може потрапити не на конкретний сайт, а на сайт конкурентів за цією темою. Методи визначення цієї величини залежать від вибраного методу опису тематики сайту.

У такому разі задача оптимізації тематики сайту визначається так.

$$\mathrm{Uf}\left(Th^{(j)}\right)\xrightarrow[Th^{(j)}\in Th^{(Alt)}]{} Max$$

Далі розглянемо часткові випадки оптимізації тематики сайту для кожного з визначених методів опису тематики. Зазначимо також, що крім функції корисності відвідувача в виразах може використовуватися функція зацікавленості користувача.

Оптимізація аудиторії сайту

Нехай:

$Gr^{(Alt)}=\{Gr^{(j)}\}_{j=1}^{N_{Gr}^{(Alt)}}$ – множина альтернативних аудиторій сайту.

$Gr^{(j)}=\left\{Gr_i^{(j)}\right\}_{i=1}^{N_{Gr_j}}$ – аудиторія сайту – множина відвідувачів, що можуть бути корисними для сайту.

$\mathrm{Uf}\left(Gr_i^{(j)}\right)$ – корисність і-ї групи відвідувачів для сайту;

$$\mathrm{Uf}\left(Gr^{(j)}\right)=\sum_{i=1}^{N_{Th_j}}\left\|Gr_i^{(j)}\right\|\mathrm{Uf}\left(Gr_i^{(j)}\right)\mathrm{Pr}^{(Jump)}\left(Gr_i^{(j)}\right),$$

де $\mathrm{Pr}^{(Jump)}\left(Th_i^{(j)}\right)$ – імовірність переходу на сайт користувача, що звернувся до і-ї теми.

У такому разі задача оптимізації тематики сайту визначається так.

$$\mathrm{Uf}\left(Gr^{(j)}\right)\xrightarrow[Gr^{(j)}\in Gr^{(Alt)}]{} Max\,.$$

Визначити конкретні величини у даних виразах, зокрема величини $\mathrm{Pr}^{(Jump)}\left(Gr_i^{(j)}\right)$, є складно та часто не може бути здійснено безпосередньо.

Оптимізація розділів каталогів

Оптимізація розділів каталогів, що відображають тематику сайту, є близькою до загальної моделі оптимізації тематики сайту.

Нехай:

$Ct^{(Alt)}=\{Ct^{(j)}\}_{j=1}^{N_{Ct}^{(Alt)}}$ – множина альтернативних наборів категорій сайту.

$Ct^{(j)}=\left\{Ct_i^{(j)}\right\}_{i=1}^{N_{Ct_j}}$ – тематика сайту – множина категорій, що описують тематику сайту. $\mathrm{Uf}\left(Ct_i^{(j)}\right)$ – корисність і-ї категорії для сайту (корисність користувачів, що потрапляють на сайт з даної категорії);

$\mathrm{Fr}\left(Ct_i^{(j)}\right)$ – частота звернення до і-ї категорії користувачами WWW;

$$\mathrm{Uf}\left(Ct^{(j)}\right)=\sum_{i=1}^{N_{Ct_j}}\mathrm{Fr}\left(Ct_i^{(j)}\right)\mathrm{Uf}\left(Ct_i^{(j)}\right)\mathrm{Pr}^{(Jump)}\left(Ct_i^{(j)}\right),$$

де $\mathrm{Pr}^{(Jump)}\left(Ct_i^{(j)}\right)$ – імовірність переходу на сайт користувача, що звернувся до і-ї теми. При певному спрощенні можна вважати:

$$\mathrm{Pr}^{(Jump)}\left(Ct_i^{(j)}\right)=\frac{\mathrm{Pr}\left(Ct_i^{(j)}\right)}{Count\left(Site,Ct_i^{(j)}\right)},$$

де $Count\left(Site,Ct_i^{(j)}\right)$ – кількість сайтів та підкатегорій у категорії каталогу; $\mathrm{Pr}\left(Ct_i^{(j)}\right)$ – імовірність потрапляння користувача в даний розділ каталогу.

Для ієрархічних багаторівневих каталогів величина $\mathrm{Pr}\left(Ct_i^{(j)}\right)$ для розділу може визначатися рекурсивно, на основі відповідної величини для надрозділу (розділу вищого рівня):

$$\Pr^{(Jump)}\left(Ct_i^{(j)}\right)=\frac{\Pr\left(Parent\left(Ct_i^{(j)}\right)\right)}{Count\left(Site,Parent\left(Ct_i^{(j)}\right)\right)},$$

де $Parent\left(Ct_i^{(j)}\right)$ – розділ вищого рівня.

Задача оптимізації тематики сайту визначається так

$$\mathrm{Uf}^*\left(Ct^{(j)}\right)\xrightarrow[Ct^{(j)}\in Ct^{(Alt)}]{}Max.$$

Множина можливих наборів категорій для сайту є обмежена правилами світових каталогів. Зокрема, існує обмеження на кількість категорій, у яких може бути представлений сайт:

$$\left\|\mathrm{Ct}^{(j)}\right\|\le \mathrm{Const},\ \text{де Const} - \text{ невелике число} \ge 1.$$

Крім того, часто обмеження базуються на комерційній природі деяких каталогів. Замість величини $\mathrm{Uf}\left(Ct^{(j)}\right)$ у задачі оптимізації повинна розглядатися величина

$$\mathrm{Uf}^*\left(Ct_i^{(j)}\right)=\mathrm{Uf}\left(Ct_i^{(j)}\right)-Price\left(Ct_i^{(j)}\right),$$

де $Price\left(Ct_i^{(j)}\right)$ – ціна розміщення сайту в розділах каталогу. Як правило, вона обчислюється як добуток ціни за реєстрацію в одному розділі на кількість розділів, у яких зареєстровано сайт.

Оптимізація пошукових запитів

Оптимізація тематики сайту, що визначається пошуковими запитами (SEO – Search Engine Optimization), є одним з найважливіших етапів оптимізації тематики сайту.

Нехай:

$Qs^{(Alt)}=\{Qs^{(j)}\}_{j=1}^{N_{Qs}^{(Alt)}}$ – множина альтернативних комплектів пошукових запитів, що ідентифікують тематику сайту.

$Qs^{(j)}=\left\{Qs_i^{(j)}\right\}_{i=1}^{N_{Qsj}}$ – тематика сайту – множина пошукових запитів, що описують тематику сайту.

$\mathrm{Uf}\left(Qs_i^{(j)}\right)$ – корисність і-ї пошукової фрази для сайту;

$\mathrm{Fr}\left(Qs_i^{(j)}\right)$ – частота використання і-ї пошукової фрази користувачами WWW;

$$\mathrm{Uf}\left(Qs^{(j)}\right)=\sum_{i=1}^{N_{Qsj}}\mathrm{Fr}\left(Qs_i^{(j)}\right)\mathrm{Uf}\left(Qs_i^{(j)}\right)\Pr^{(Jump)}\left(Qs_i^{(j)}\right),$$

де $\Pr^{(Jump)}\left(Qs_i^{(j)}\right)$ – імовірність переходу на сайт користувача, що звернувся до і-ї теми.

Величина $\Pr^{(Jump)}\left(Qs_i^{(j)}\right)$ залежить від характеристик глобального конкурентного середовища, у якому функціонує сайт. Користувач може потрапити не на конкретний сайт, а на сайт конкурентів за певною темою. Цю величину можна визначити так:

$$\Pr^{(Jump)}\left(Qs_i^{(j)}\right)=\Pr^{(Jump)}\left(\mathrm{Pos}\left(Site,Qs_i\right)\right)\mathrm{Pos}\left(Site,Qs_i\right),$$

де $\mathrm{Pos}\left(Site,Qs_i\right)$ – позиція, яку обіймає сайт пзао запитом Qs_i у результатах пошукових машин; $\Pr^{(Jump)}\left(Pos\right)$ – імовірність переходу користувача за посиланням, що обіймає позицію Pos у результатах роботи пошукових машин.

Функція розподілу густоти імовірності $\Pr^{(Jump)}\left(Pos\right)$ є достатньо складною та залежить від багатьох факторів (зокрема від якості пошуку та якості автоматизованого формування анотацій). Реальний вигляд цієї функції сьогодні власниками пошукових машин не розголошується.

У такому разі задача оптимізації тематики сайту визначається так:

$$\mathrm{Uf}\left(Qs^{(j)}\right)\xrightarrow[Qs^{(j)}\in Qs^{(Alt)}]{}Max.$$

Множина можливих наборів пошукових фраз обмежується сучасними алгоритмами визначення релевантності сторінки запиту. Одна й та ж сторінка не може бути одночасно високорелевантна багатьом пошуковим фразам з абсолютно різними ключовими словами. Проте сторінка може бути релевантна різним фразам, що складаються з одних ключових слів. У такому разі до одного набору пошукових фраз можна відносити комплекти різних комбінацій ключових слів, якими є релевантна сторінка.

Детальніший аналіз наборів можливих фраз вимагає оцінки позиціонування сайту в глобальному середовищі та детальнішої класифікації запитів користувачів.

Оптимізація параметрів реклами сайту

Оптимізація тематики сайту, що визначається параметрами контекстної та цільової реклами, є одним з найважливіших етапів оптимізації тематики сайтів, просування яких в WWW має комерційний характер.

Оптимізація цільової реклами (реклами, що базується на прямому визначенні аудиторії сайту) є уточненим варіантом оптимізації аудиторії сайту, що розглядалася вище. Уточнення стосуються величин $\Pr^{(Jump)}\left(Th_i^{(j)}\right)$ – імовірність переходу на сайт користувача, що звернувся до і-ї теми, та $\mathrm{Fr}\left(Th_i^{(j)}\right)$ – частота звернення до і-ї теми користувачами WWW. Дані величини в середовищі показу цільової реклами є доступними та визначаються на основі статистики сайтів, що здійснюють показ реклами.

Замість функції $\mathrm{Uf}\left(Th^{(j)}\right)$ використовується функція $\mathrm{Uf}^*\left(Th^{(j)}\right)$

$$\mathrm{Uf}^*\left(Th_i^{(j)}\right)=\mathrm{Uf}\left(Th_i^{(j)}\right)-Price\left(Th_i^{(j)}\right),$$

де $Price\left(Th_i^{(j)}\right)$ – ціна реклами за тематикою $Th_i^{(j)}$.

У загальному випадку ціна реклами має декілька складових:

$$Price\left(Th_i^{(j)}\right)=\left(Price_{Show}\left(Th_i^{(j)}\right)+Price_{Click}\left(Th_i^{(j)}\right)\right)\left\|Gr_i^{(j)}\right\|,$$

де $Price_{Show}$ – затрати на показ реклами за вибраною темою у розрахунку на окремого користувача; $Price_{Click}$ – затрати на перехід користувача на сайт з реклами.

Оптимізація контекстної реклами (реклами, що використовує контекст пошуку та обробки інформації користувачем в WWW) базується на тих же підходах, що і оптимізація тематики на основі пошукових запитів.

Основними відмінностями цієї задачі від задачі оптимізації пошукових запитів є:

- Простіше визначення обсягів аудиторії сайту та імовірності потрапляння на сайт;
- Наявність додаткових параметрів, що відображають комерційну суть реклами.

У такому разі задача має такий вигляд:

$$\mathrm{Uf}^*\left(Qs^{(j)}\right)\xrightarrow[Qs^{(j)}\in Qs^{(All)}]{}Max$$

$$\mathrm{Uf}^*\left(Qs^{(j)}\right)=\sum_{i=1}^{N_{Qs_j}}\mathrm{Fr}\left(Qs_i^{(j)}\right)\mathrm{Uf}^*\left(Qs_i^{(j)}\right)\Pr^{(Jump)}\left(Qs_i^{(j)}\right)$$

$$\mathrm{Uf}^*\left(Qs_i^{(j)}\right)=\mathrm{Uf}\left(Qs_i^{(j)}\right)-Price\left(Qs_i^{(j)}\right),$$

де $Price\left(Qs_i^{(j)}\right)$ – затрати на перехід користувача на сайт за рекламним оголошенням, що відповідає пошуковому запиту $Qs_i^{(j)}$.

Оптимізація оточення сайту

Оптимізація тематики сайту, що визначається через оточення сайту, є близькою за підходами до оптимізації тематичної реклами. Проте ця задача є складніша та загальніша. Основними складностями, що відрізняють цю задачу від аналогічних, є:

• складність отримання інформації;

• об'єктивність процесу формування оточення сайту: власники сайту мають обмежені можливості щодо формування оточення.

Отже, повноцінне формальне розв'язання даної задачі є практично неможливим, хоча це не виключає можливості цілеспрямованої роботи з покращання тематики сайту, через його оточення.

Важливим фактором, що впливає на задачу оптимізації оточення сайту є тісний зв'язок даної задачі зі задачею оптимізації сайту через пошукові фрази.

Структура уточненої функції корисності $\mathrm{Uf}^*\left(Site_i^{(j)}\right)$ у виразах відповідної задачі оптимізації має такий характер:

$$\mathrm{Uf}^*\left(Site_i^{(j)}\right)=\mathrm{Uf}\left(Site_i^{(j)}\right)-\mathrm{Price}\left(Site_i^{(j)}\right)+\mathrm{Uf}_{Search}\left(Site_i^{(j)}\right)$$

де $\mathrm{Price}\left(Site_i^{(j)}\right)$ – затрати на розміщення посилання на сайт; $\mathrm{Uf}_{Search}\left(Site_i^{(j)}\right)$ – вплив даного посилання на відображення тематики сайту через пошукові запити.

Зазначимо, що корисність уточненого посилання та його складові можуть бути від'ємними величинами. Це може бути викликано такими факторами:

• Наявністю посилань, що небажано модифікують тематику сайту та її подання через пошукові запити;

• Наявністю посилань, за якими на сайт потрапляють відвідувачі, що руйнують спільноту сайту та завдають шкоди репутації сайту.

Методи оптимізації тематики сайту

Результатом розв'язання задачі оптимізації сайту є побудова нової тематики сайту, яка більшою мірою задовольняла б потреби власників сайту та сприяла досягненню поставленої перед ним мети. Як правило, нова оптимізована тематика сайту є похідною від первісної і утримується одним з таких методів:

• узагальнення тематики;

• зміщення тематики;

• уточнення тематики.

Ці методи можуть використовуватися для побудови множини можливих тематик сайту, що потім будуть розглядатися як допустимі в задачі оптимізації.

Такий підхід дозволяє уникнути множин великої розмірності та розробити достатньо формалізовані підходи та автоматизовані засоби підбору можливої тематики сайту.

Нижче детальніше розглянемо кожен з наведених методів.

Узагальнення тематики

Метод узагальнення тематики сайту використовується у випадку, коли первісна тематика сайту є непопулярною серед користувачів WWW – тоді існування сайту може себе не виправдовувати, і сайт повною мірою не виконуватиме поставлених перед ним завдань.

Узагальнення тематики сайту в означеному випадку доцільно здійснювати окремо для кожного з методів визначення тематики.

Узагальнення аудиторії сайту

У випадку, якщо обсяги аудиторії сайту є меншими за певну контрольну величину, доцільно проводити узагальнення аудиторії.

Як показник обсягу аудиторії доцільно використовувати сумарну зважену корисність аудиторії:

$$\mathrm{Uf}\left(Gr^{(j)}\right)=\sum_{i=1}^{N_{Th_j}}\left\|Gr_i^{(j)}\right\|\mathrm{Uf}\left(Gr_i^{(j)}\right)<Const$$

або просто сумарний обсяг аудиторії:

$$\mathrm{Uf}\left(Gr^{(j)}\right)=\sum_{i=1}^{N_{Th_j}}\left\|Gr_i^{(j)}\right\|<Const\ .$$

Узагальнення аудиторії здійснюється шляхом розширення діапазона допустимих значень показників, що її характеризують. Це зокрема:

- географічні та мовні показники;
- часові показники (періоди доби, тижня, місяця, року);
- соціальні показники.

Уточнення тематики

Метод уточнення тематики сайту використовується у випадку, коли первісна тематика сайту є неефективною для власників сайту або є занадто висококонкурентною в глобальному середовищі WWW. У такому разі існування сайту може себе не виправдовувати, і сайт повною мірою не виконуватиме поставлених перед ним завдань.

Тематику сайту в означеному випадку доцільно уточнювати окремо для кожного з методів визначення тематики.

Уточнення аудиторії сайту

У випадку, якщо корисність аудиторії сайту або імовірність потрапляння представника аудиторії на сайт є меншою за певну контрольну величину, доцільно проводити уточнення аудиторії.

Як показник необхідності уточнення аудиторії доцільно використовувати такі критерії:

$$\mathrm{Uf}\left(Gr^{(j)}\right)=\sum_{i=1}^{N_{Th_j}}\left\|Gr_i^{(j)}\right\|\mathrm{Uf}\left(Gr_i^{(j)}\right)\mathrm{Pr}^{(Jump)}\left(Gr_i^{(j)}\right)<Const$$

або

$$\mathrm{Uf}\left(Gr^{(j)}\right)=\sum_{i=1}^{N_{Th_j}}\left\|Gr_i^{(j)}\right\|\mathrm{Pr}^{(Jump)}\left(Gr_i^{(j)}\right)<Const\ .$$

Уточнення аудиторії здійснюється шляхом звуження діапазона допустимих значень показників, що її характеризують. Це зокрема:

- географічні та мовні показники;
- часові показники (періоди доби, тижня, місяця, року);
- соціальні показники.

Зміщення тематики

Зміщення тематики сайту може застосовуватися як альтернатива узагальненню та уточненню тематики сайту у випадку такої необхідності. Зміщення тематики з метою розширення аудиторії сайту може відбуватися в тих же випадках, що і узагальнення. Аналогічно зміщення з метою звуження аудиторії може відбуватися за тих же умов, що і звуження тематики сайту.

У більшості випадків зміщення тематики сайту відбувається як альтернатива наведеним методам, яка дозволяє зберегти високі показники корисності та зацікавленості аудиторії, є реальною для ефективного подання тематики в конкурентному середовищі і водночас дозволяє уникнути небажаного звуження аудиторії.

Приклади оптимізації тематики сайту

Оптимізація сайту art.ridne.net

Наводяться результати процесу оптимізації тематики та її подання для сайту art.ridne.net, який здійснювався згідно з наведеними вище алгоритмами.

Мета сайту була визначена як:

1. Продаж творів сучасного українського мистецтва через мережу Інтернет.
2. Експозиція для Інтернет-спільноти сучасного українського мистецтва.

Як аналоги в середовищі WWW було вибрано існуючі Інтернет-магазини, що торгують творами мистецтва.

Як аналоги поза середовищем WWW було вибрано галереї сучасного мистецтва.

Задача вибору критерію ефективності виявилася достатньо складною. Функція корисності в даному випадку повинна була враховувати два фактори:

1. Прибуток від продажу корисності (перша компонента мети).
2. Сам факт відвідування Інтернет-галереї користувачем Інтернету (друга компонента мети).

Проте, успішність продажу унікальних творів мистецтва визначається більшою мірою характеристиками, які не могли контролюватися власниками сайту (ціна роботи, якість роботи, популярність митця, якість електронних копій). Цей факт, а також прогнозоване розміщення реклами на сторінках сайту (прибуток від якої є пропорційним до числа відвідувачів сайту) спричинили рішення критерєм ефективності використовувати кількість відвідувачів сайту.

З точки зору визначення тематики час існування можна поділити на такі періоди:

1. Вибір оптимальної тематики не проводився.
2. Тематику було визначено як “українське мистецтво”.
3. Тематику було розширено на англійський відповідник “ukrainian art”.
4. Тематику було зміщено як “art for sale”.
5. Тематику було розширено як “modern art”.

На кожному з етапів проводилися роботи з покращання подання тематики на сайті та серед його глобального оточення в системі WWW.

Нижче наводяться діаграми (рис. 1, 2), які ілюструють покращання показників ефективності тематики сайту. Дані наводяться за період від 1 серпня 2001 року до 1 лютого 2004 року. Групування здійснено помісячно.

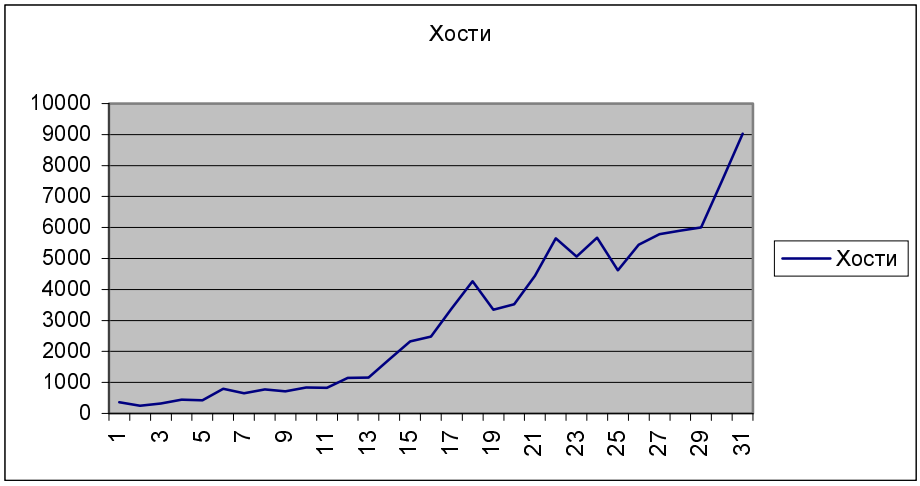

Рис. 1. Динаміка зміни кількості відвідувачів сайту (“хостів”)

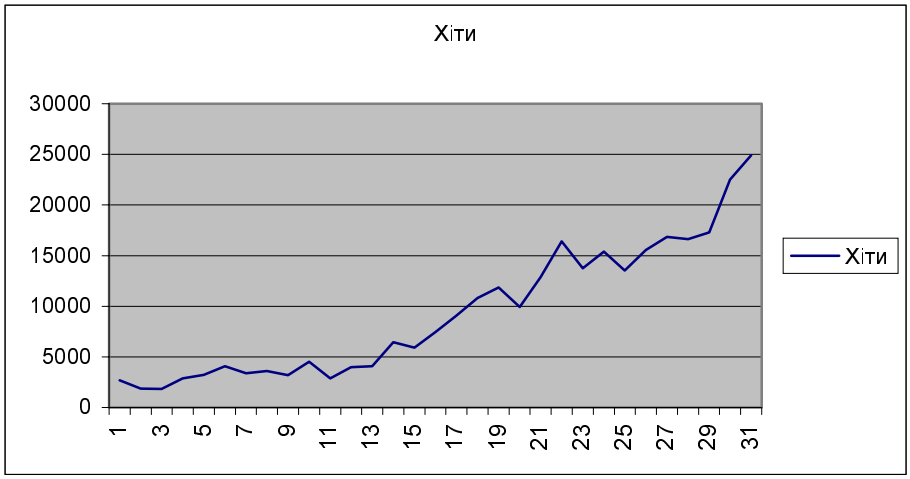

Рис. 2. Динаміка зміни кількості запитів до сайту (“хітів”)

Висновки

Тематика є однією з найважливіших характеристик сайту. Тематика сайту має описуватися на основі існуючих реалій Інтернету та WWW, зокрема правил опису та визначення тематики сайту глобальними пошуковими та каталожними сервісами. Важливим аспектом опису тематики сайту є побудова моделі аудиторії сайту. Аудиторія сайту і є головним показником тематики сайту.

Моделювання тематики сайту на основі аудиторії часто є достатньо складною задачею, яка може бути розв'язаною лише за допомогою введення додаткових елементів моделі. Таким елементом є модель навігації користувача по WWW, яка приводить його на сайт.

Отже, для опису тематики сайту доцільно використовувати навігаційні методи, кожен з яких відповідає типовому сценарію навігації користувача в WWW. Тобто, отримано такі методи опису тематики сайту:

- визначення тематики сайту на основі рубрик порталів та каталогів;
- визначення тематики сайту на основі пошукових запитів;
- визначення тематики сайту на основі WWW-оточення.

Тематика сайту, визначена на основі наведених методів, може потребувати оптимізації для повнішого досягнення поставлених перед сайтом цілей.

1. ODP Social Contract. http://dmoz.org/socialcontract.html. 2. Organization of Search Engine Optimization Professionals. http://www.seopros.org/ 3. Поисковая оптимизация и продвижение сайтов в Интернете. http://www.optimization.ru/ 4. Кодекс оптимизатора http://www.searchengines.ru/pages.php?page=code. 5. Хартия оптимизаторов. http://charter.seolab.ru/ 6. Report a Spam Result http://www.google.com/contact/spamreport.html. 7. B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, and J. Wiltshire. Discovery of aggregate usage profiles for Web personalization. // Proceedings of the WebKDD 2000 Workshop at the ACM SIGKKD 2000, Boston, August 2000. 8. Flake G. , Lawrence S., Giles C. Efficient identification of web communities. // Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000), Boston, MA, 2000. ACM Press. 9. Flake G., Lawrence S., Giles C., Coetzee F. Self-Organization of the Web and Identification of Communities. IEEE Computer, 35(3), 66–71, 2002 http://webselforganization.com/ 10. Gillet S., Kapor M. Self-govering Internet: Coordination by Design. Massachusetts Institute of Technology. Center for Coordination Science. Technical Report. 1997., 25p., http://ccs.mit.edu/CCSWP197/CCSWP197.html. 11. J. Srivastava, R. Cooley, M. Deshpande, P-T. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explorations, (1) 2, 2000. 12.Pennock D., Flake G., Lawrence, Glover E., Giles C. Winners don't take all:Characterizing the competition for links on the web. Proceedings of the National Academy of Sciences, Volume 99, Issue 8, pp. 5207–5211, April, 2002. http://modelingtheweb.com/ 13. Usabiity in Russia. http://www.usability.ru/articles.htm. 154. Кирсанов Д. Веб-дизайн. –Спб., 1999 – С.360. 14. Лебедев А. Юзабилити. 2000р. http://www.design.ru/kovodstvo/paragraphs/45.html. 15. Пелещишин А.М. Методи та алгоритми моделювання Web-систем // Вісник ДУ "Львівська політехніка". 2000. – №406. – С.199–211. 16. Пелещишин А.М., Гулка Т.Б. Інформаційна система аналізу діяльності Web-вузла // Вісник НУ "Львівська політехніка". – 2001. – №438. – С.115–120.