УДК 004.773.2

[1]**О.Ю. Тимовчак-Максимець,** асп.
[2]**А.М. Пелещишин,** д.т.н., доц.

РОЗРОБЛЕННЯ ФОРМАЛІЗОВАНИХ ЗАПИТІВ ДЛЯ ВИЯВЛЕННЯ РЕЛЕВАНТНИХ ВЕБ-ФОРУМІВ ТА ДИСКУСІЙ

Національний університет «Львівська політехніка»
[1]E-mail: linoks@gmail.com
[2]E-mail: apele@ridne.net

Описано способи виявлення веб-форумів, які релевантні об'єкту шуканого споживацького досвіду. Дано означення споживацького досвіду та його об'єкта. Розроблено формалізовані запити для виявлення релевантних веб-спільнот та їхніх дискусій.

Ключові слова: веб-спільнота, веб-форум, глобальні пошукові системи, дискусія, пошуковий запит, споживацький досвід.

Постановка проблеми

Розроблення формалізованих запитів на виявлення релевантних веб-форумів та дискусій є частиною вирішення більшої пошукової задачі, що полягає у виявленні в дискусіях веб-форумів вражень або міркувань споживачів стосовно певного об'єкта чи сутності, які формуються у них на основі знань, отриманих від взаємодії з навколишнім світом. Такі міркування або враження будемо називати споживацьким досвідом.

Споживачами можуть бути самі учасники веб-спільнот або інші особи, на досвід яких посилаються учасники спільноти.

Комунікативні веб-спільноти, такі, як веб-форуми, виступають платформою для публікування та обміну досвідом щодо продуктів, послуг, осіб тощо, які виступають об'єктами досвіду.

Об'єктом споживацького досвіду (далі об'єкт досвіду) є предмет або сутність, з яким споживач певним чином вступає у взаємодію і стосовно якого/якої формує своє враження або міркування.

Релевантність веб-спільнот та їх дискусій визначається відносно об'єкта шуканого споживацького досвіду.

Виявлення дискусій на веб-форумах, які можуть містити споживацький досвід стосовно певного визначеного об'єкта, є нетривіальною задачею з огляду на такі причини:

– відбір сторінок, які належать веб-спільнотам, не є тривіальним, оскільки потребує виділення відповідних недвозначних характеристик цих сторінок, за якими їх можна відрізнити від сторінок інших типів;

– формулювання заголовків дискусій на веб-форумах не є регламентоване, тобто не має жодних обмежень окрім відповідності тематичній гілці форуму, тому стосунок до об'єкта шуканого споживацького досвіду не завжди очевидний;

– обговорення в дискусії може переходити у площину тематик, які є суміжними або дотичними до основної, а тому пошук релевантних дискусій повинен відбуватися в межах усієї предметної області об'єкта досвіду.

Ураховуючи ці особливості необхідно дослідити способи виявлення веб-форумів або веб-спільнот та дискусій, які можуть містити споживацький досвід про певний об'єкт.

Аналіз останніх досліджень

Сучасні комп'ютерно-лінгвістичні дослідження користувацького інформаційного наповнення у вебі зосереджені на застосуванні обчислювальних методів для виявлення суб'єктивності в тексті, що отримали назву «opinion mining» («добування думок», «добування оцінкових суджень») – виявлення суджень користувачів про певний предмет або сутність.

У зарубіжних працях активно досліджуються дві проблеми у сфері виявлення оцінкових суджень: класифікації та видобування [1].

Класифікація речень або уривків текстів відбувається за:

– характером: об'єктивні/суб'єктивні [2; 3];

– сутністю емоцій: позитивні/негативні та їх ступенем [4];

–типом емоцій: [5; 6].

Методи та засоби видобування оцінкових суджень досліджуються в роботах [7; 8].

Способи виявлення релевантних спільнот

На першому етапі виявлення релевантних спільнот здійснюється пошук та класифікація релевантних веб-спільнот.

Релевантними спільнотами будемо називати ті веб-форуми, які мають хоча б одну дискусію в тій предметній області, в якій знаходиться об'єкт досвіду.

За результатами пошуку формується перелік веб-форумів, на яких здійснюватиметься пошук досвіду стосовно певного об'єкта.

Перелік спільнот формується такими способами:

– пошуком у веб шляхом використання можливостей та ресурсів пошукових систем;

– аналізом діяльності з позиціонування та реклами веб-спільноти;

– відкритим суспільним доповненням переліку;

– аналізом та пропозиціями експерта щодо редагування переліку.

Кожен із перелічених способів є достатньо ужитковим, оскільки дає змогу виявити ті спільноти, які могли не потрапити до переліку іншим способом через певні обмеження.

Усі способи можна застосовувати паралельно або в певних комбінаціях залежно від об'єкта пошуку для виявлення якомога більшої кількості релевантних спільнот, тобто для забезпечення більшої повноти переліку.

Кожен зі способів має особливості, які визначають ефективність та доцільність застосування того чи іншого способу в різних умовах та обставинах.

Застосування існуючих глобальних пошукових систем, таких, як Google, Yahoo, для виявлення релевантних спільнот на практиці є найпродуктивнішим способом, що дає змогу виявити найбільшу кількість спільнот. Такий спосіб є практично універсальним, тобто ефективним у більшості випадків.

Глобальні пошукові системи є менш ефективними в тих випадках, коли, наприклад, об'єкт пошуку є складний, неоднозначний і/або специфічний, оскільки інформація про такий об'єкт повинна надходити зі спеціалізованого джерела з високим рівнем достовірності, тоді як пошукові системи повертають результати запиту до всіх проіндексованих спільнот. Прикладами таких об'єктів пошуку є медичне лікування, юридичні послуги тощо.

Аналіз діяльності з позиціонування та реклами спільнот полягає в аналізі банерної реклами, контекстної реклами (наприклад, Google AdSense), аналізі каталогів, зокрема платних, тощо. Цей спосіб дозволяє виявити спільноти, які не мають високої відвідуваності та не є популярними через певні обставини, такі, як незначна тривалість існування спільноти, складні умови реєстрації, обмеження кола учасників певними вимогами.

Аналіз реклами та іншої діяльності з позиціонування спільнот є додатковим, тобто використовується для доповнення переліку.

Наприклад, такий спосіб є ефективним для доповнення переліку отриманого за допомогою пошукових систем, тобто для виявлення сторінок спільнот, які не потрапили до результатів пошуку глобальними системами.

Причинами, які пояснюють відсутність деяких веб-сторінок у результатах пошукових систем, є непроіндексованість веб-сторінки, невисокі позиції веб-сторінки в пошукових системах тощо.

Цей спосіб доцільно застосовувати у випадках, коли об'єкт пошуку є новинкою й інформація про нього з'явилася недавно.

До сторінок-новинок будемо відносити ті сторінки, які ще не проіндексовані глобальними пошуковими системами. На практиці термін індексування нових сторінок може становити від кількох тижнів до півроку.

Аналіз діяльності з позиціонування також застосовується, коли необхідно отримати якомога повнішу інформацію про об'єкт.

Відкрите суспільне доповнення переліку веб-спільнот відбувається за умови винесення цього переліку для суспільного доступу, наприклад, на певному он-лайн ресурсі. У таких випадках, особи, які, наприклад, зацікавлені у просуванні спільноти (підвищенні авторитетності, збільшенні аудиторії форуму тощо), можуть вносити інформацію про спільноту безпосередньо до переліку. Цей спосіб є додатковим і застосовується для доповнення переліку спільнот. Такий спосіб доцільно використовувати для виявлення альтернативної і/або неординарної точки зору.

Можливість суспільного доповнення переліку зазвичай спричиняє появу великої кількості інформаційного шуму (спам, додавання нерелевантних спільнот тощо), тому одержані результати потребують додаткової верифікації в експертів.

Цей спосіб доповнення переліку спільнот є тривалим у часі, тому малоефективний у тих випадках, коли необхідно оперативно виявити релевантні спільноти. Цей спосіб ефективний для моніторингу популярності певної тематики, поширення інформації у віртуальних спільнотах тощо.

У результаті використання одного або кількох згаданих способів виявлення вебсторінок спільнот отримуємо «сирий» нефільтрований перелік веб-сторінок. Цей перелік аналізується експертами, які приймають рішення щодо редагування цього переліку (доповнення, вилучення спільнот зі списку тощо).

До аналізу експерта слід вдаватися в тих випадках, коли авторитетність спільноти відіграє ключову роль.

Експерт аналізує показники відвідуваності, активності учасників спільноти, досліджує тематику спільноти, тематику гілок форуму тощо. На основі цього аналізу експерт визначає відповідність рівня авторитетності спільноти поставленому завданню і приймає рішення про включення або вилучення спільноти з переліку. Наприклад, цей спосіб доцільно застосовувати для завдань, пов'язаних із визначенням рівня освіти у ВНЗ, вибору майбутньої спеціальності.

Формалізований запит для виявлення релевантних спільнот

Пошукові системи є надзвичайно продуктивним засобом для виявлення спільнот, у яких тематика дискусій належить до заданої предметної області. Однак базові запити до пошукових систем із використанням ключових слів ефективно здійснюють пошук за ключовими словами на веб-сторінці, але малоефективні для відбору веб-сторінок певних типів.

Крім цього, в результатах пошуку за ключовими словами виявляється велика кількість інформаційного шуму (нерелевантні дискусії, веб-сторінки, що не належать до спільнот тощо). Організація пошуку сторінок веб-форумів потребує, окрім базових, додаткових можливостей параметризування пошуку.

Для виявлення релевантних спільнот авторами розроблено формалізований параметризований запит до глобальних пошукових систем (рис. 1).

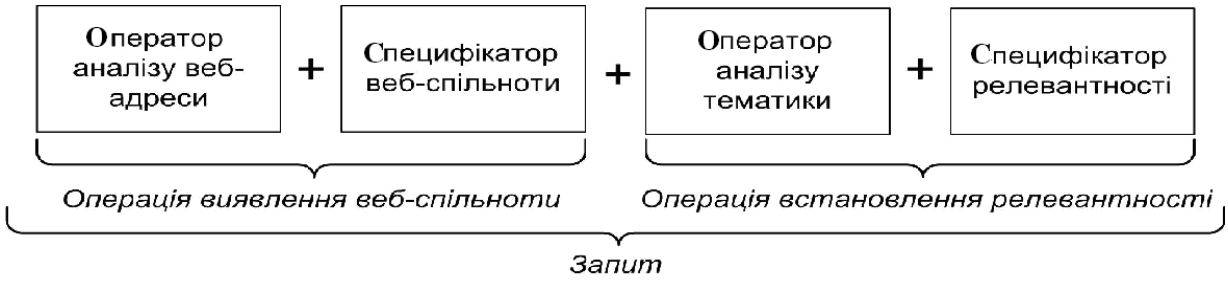

Рис. 1. Формалізований запит на виявлення релевантних веб-спільнот

Запит до глобальної пошукової системи на виявлення релевантних спільнот складається з таких операцій:

– операції виявлення веб-спільноти;

– операції встановлення релевантності.

Операція виявлення веб-спільноти дозволяє виявити і відібрати веб-сторінки, які належать до веб-спільнот. Функціонально ця операція складається з оператора аналізу веб-адреси і специфікатора веб-спільноти.

Оператор аналізу веб-адреси – це фільтр результатів пошуку, який вилучає ті веб-сайти та веб-сторінки, які не відповідають певній визначеній вимозі до структури веб-адреси. Така вимога встановлюється до URL-адреси веб-сторінки чи веб-сайту.

За формальними ознаками в URL-адресі можна ідентифікувати веб-сторінки чи веб-сайти спільнот. Тому оператор аналізу веб-адреси здійснює відбір сторінок за вимогою до URL-адреси, яка встановлюється специфікатором веб-спільноти.

Специфікатор веб-спільноти – це формальна ознака, яка вказує на те, що веб-сайт або веб-сторінка належить до веб-спільноти.

Специфікатор веб-спільноти це частина URL-адреси веб-сайту чи веб-сторінки, яка несе в собі ознаку, за якою можна ідентифікувати веб-спільноту.

Універсальним та однозначним специфікатором веб-спільноти є «forum». Наявність такого специфікатора в URL-адресах веб-сайтів та веб-сторінок свідчить про те, що вони належать до веб-спільнот.

Однак в URL-адресах багатьох спільнот специфікатора «forum» немає. Для таких спільнот специфікатор визначається з програмної платформи, на якій розміщена спільнота.

Кожна платформа має певний специфікатор веб-сторінки, яка належить до спільноти (див. таблицю).

Специфікатор веб-спільноти складається з основи та уточнення. Основа специфікатора знаходиться в URL-адресі сторінки і вказує на те, що ця сторінка є сторінкою відповідної спільноти.

У більшості випадків наявність основи специфікатора в URL-адресі сторінки є достатнім для того, щоб стверджувати, що виявлена сторінка належить до певної спільноти.

Однак деякі платформи мають ідентичні засоби для позначення сторінок як веб-форумів, так і сторінок блогів, довідкових сторінок тощо. Для таких платформ необхідно використати уточнення специфікатора.

Специфікатори веб-сторінок форумів для популярних платформ

Платформа веб-спільноти	Специфікатор веб-спільноти		
	Основа специфікатора	Необхідність уточнення	Уточнення специфікатора
Універсальний для всіх платформ	forum	Ні	–
phpBB	viewtopic	Ні	–
vBulletin	showthread	Ні	–
Simple Forum Machines (SMF)	index.php?topic=	Так	Powered by SMF
myBB	showthread	Ні	–
Invision Power Board (IPB)	thread.php	Так	Powered by myBB
	showtopic	Так	Powered by Invision Power Board
XMB (eXtreme Message Board), PHP-Fusion, Discuz тощо	viewthread	Ні	–
			Powered by ExBB
Word Press	“topic.php”	Так	Powered by BBpress

Уточнення специфікатора служить для перевірки того, чи основа специфікатора використовується для позначення сторінки веб-форуму.

Для уточнення є фраза, яка застерігає права розробників платформи веб-форуму і в більшості випадків складається з двох частин: «Powered by», яка вказує на програмне забезпечення, та «компанія-розробник», наприклад, «Invision Power Board».

Основа специфікатора та уточнення використовуються паралельно. Однак основа специфікатора розташована в заголовку сторінки, а уточнення – в тексті, тому пошук здійснюють у двох різних частинах сторінки.

Операція встановлення релевантності дозволяє виявити серед сторінок веб-спільнот ті, які можуть стосуватися об'єкта пошуку. Ця операція складається з оператора аналізу тематики і специфікатора релевантності.

Оператор аналізу тематики – це фільтр результатів пошуку, який вилучає з переліку ті сторінки веб-спільнот, які не пов'язані з заданою тематикою. Пов'язаність веб-сторінки з певною тематикою будемо визначати за її назвою. Назвою сторінки будемо вважати слово або словосполучення, яке знаходиться у заголовку сторінки, у тегу *<title>*.

Вважатимемо, що веб-сторінка пов'язана з заданою тематикою, якщо заголовок веб-сторінки містить відповідну ознаку, яка задається специфікатором релевантності. Відповідно оператор визначення тематики залишає в переліку лише ті сторінки веб-спільнот, у яких заголовок сторінки є релевантний заданій тематиці.

Специфікатор релевантності – це ознака, яка вказує на те, що сторінка веб-спільноти є релевантною об'єкту пошуку. Ознакою релевантності виступає слово або словосполучення в заголовку веб-сторінки. Оскільки заголовок веб-сторінки в більшості випадків збігається з темою дискусії, то присутність певного слова або словосполучення в заголовку свідчить про те, що дискусія може стосуватися об'єкта пошуку. Наприклад, якщо об'єктом пошуку є досвід щодо популярних книжок для дітей, то для нього специфікаторами релевантності є «книга», «книжка», «популярне видання» тощо.

Формалізований запит на виявлення релевантних дискусій

Формалізований запит на виявлення релевантної дискусії складається з трьох операцій (рис. 2):

– операції локалізації пошукового запиту;

– операції встановлення релевантності;

– операції виявлення об'єкта досвіду.

Операція локалізації пошукового запиту забезпечує пошук у межах заданої адреси веб-форуму. Операція має дві складові: оператор локалізації та специфікатор розташування веб-форуму.

Оператор локалізації – це обмежувач, накладений на зону дії пошукового запиту.

Обмежувач вказує рамки, в межах яких відбуватиметься пошук.

Специфікатор розташування веб-форуму – це вказівник на ресурс, на якому здійснюватиметься пошук. Ресурсом виступає веб-спільнота, яка була виявлена на етапі пошуку релевантних спільнот. Указівником на ресурс є URL-адреса, за якою можна доступитися до спільноти.

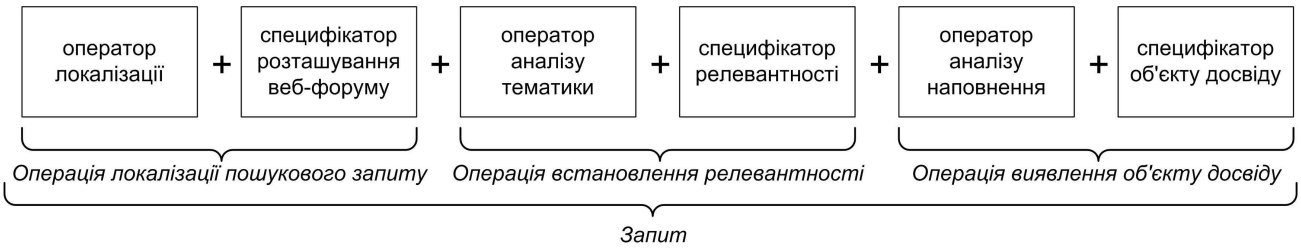

Рис. 2. Формалізований запит на виявлення релевантних дискусій на веб-форумі

Операція встановлення релевантності складається з оператора аналізу тематики та специфікатора релевантності.

Операція виявлення об'єкта досвіду слугує для пошуку в тілі сторінки для виявлення об'єкта досвіду. Ця операція дозволяє виявити згадки про об'єкт досвіду в дискусії і тим самим установити доцільність здійснення пошуку досвіду щодо об'єкта в цій дискусії. Операція виявлення об'єкта досвіду складається з оператора аналізу наповнення та специфікатора об'єкта досвіду.

Оператор аналізу наповнення виконує пошук певної заданої лінгвістичної конструкції в наповненні сторінки. Цей оператор відфільтровує з результатів пошуку ті сторінки, наповнення яких не містить заданої лінгвістичної конструкції, і залишає серед результатів ті сторінки, в яких задана лінгвістична конструкція зустрічається хоча б один раз. Лінгвістичною конструкцією тут виступає специфікатор об'єкта досвіду.

Специфікатор об'єкта досвіду – це ознака, за якою можна перевірити наявність згадувань про об'єкт досвіду в дискусії. Таким специфікатором виступають лінгвістичні конструкції, які прямо або опосередковано вказують на об'єкт досвіду. Наприклад, якщо об'єктом шуканого досвіду є видавництво «Нова Книга», то специфікаторами цього об'єкта будуть такі конструкції: "Нова Книга", "Nova Knyha", "novaknyha" тощо.

Висновки

Досліджено проблему виявлення вебсторінок, які є релевантні заданому об'єкту пошуку. Для розв'язання цієї проблеми розроблено формалізовані запити до глобальних пошукових.

Література

1. *Esuli, A.; Sebastiani, F.* 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. – Proceedings of the International Conference on Language Resources and Evaluation. Genoa: 417–422.

2. *Wiebe, J.; Bruce, R.; Martin, M.* 2004. Learning Subjective Language. – Computational Linguistics. 30(3): 277–308.

3. *Kim, S.; Hovy, E.* 2004. Determining the Sentiment of Opinions. – Proceedings of Conference on Computational Linguistics (COLING-04).Geneva: 1367–1373.

4. *Pang, B.; Lee, L.* 2008. Opinion Mining and Sentiment Analysis. – Foundations and Trends in Information Retrieval. 2(1-2): 1–135.

5. *Mullen, T.; Malouf, R.* 2006. A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. – Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs. Berlin: Springer-Verlag: 159–162.

6. *Strapparava, C.; Valitutti, A.* 2004. WordNet-Affect: An Affective Extension of WordNet. – Proceedings of the 4th International Conference on Language Resources and Evaluation: 1083 – 1086.

7. *Dave, K.; Lawrence, S.; Pennock, D.* 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification in Product Reviews. – Proceedings of ACM WWW2003. Budapest: 519 – 528.

8. *Turney, P.; Littman, M.* 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. – ACM Transactions on Information Systems. 4(21): 315–346.

Стаття надійшла до редакції 25.10.2011.