

Обробка інформації в складних організаційних системах

УДК 004.9

А.Л. Єрохін, М.Д. Одарченко

Харківський національний університет радіоелектроніки, Харків

ПРОГРАМНА СИСТЕМА ДЛЯ ЗБОРУ І АНАЛІЗУ ІНФОРМАЦІЇ В СОЦІАЛЬНИХ МЕДІА

У статті пропонується вдосконалений підхід до збору та аналізу інформації в соціальних медіа, якими користуються великі підприємства з комерційною метою. Для проведення аналізу була розроблена програмна система з метою автоматизації збору інформації.

Ключові слова: соціальні медіа, моделі збору та аналізу інформації, бізнес-дослідження.

Вступ

В теперішній час невід'ємною частиною ведення бізнесу стала підтримка іміджу підприємства в мережі Інтернет. Безліч підприємств отримують прибутки, продаючи товари, або послуги тільки через Інтернет. Відгуки в соціальних мережах можуть знищити бізнес, або навпаки, - збільшити прибутки в декілька разів. Для успішності бізнесу необхідно максимально швидко реагувати на відгуки, що публікуються у вільному доступі, та приймати оперативні реакційні заходи, щодо збереження іміджу організації і якісного задоволення потреб своїх клієнтів.

Імідж організації — сукупне сприйняття конкретного підприємства споживачами, діловими партнерами, громадськими організаціями, контактними аудиторіями та персоналом [1]. Оскільки соціальним медіа характерна самоорганізація, то вони майже не контролювані і не передбачувані [2]. Крім того, у зв'язку з великою кількістю медіа джерел та людей, які приймають участь в створенні контенту, своєчасне відстеження згадувань організації є актуальною і складною проблемою для бізнесу.

Метою статті є дослідження способів автоматизації збору інформації в соціальних медіа та розробка програмної системи для автоматизації збору комерційної інформації в соціальних медіа.

1. Формалізація процедури аналізу інформації із соціальних мереж та постановка задач дослідження

Побудова математичної моделі оцінювання іміджу підприємства відноситься до задач, які важко формалізуються. Тому необхідно спочатку дослідити усі фактори, які впливають на імідж підприємства, а потім виділити найбільш значущі фактори

впливу і встановити функціональні залежності їхнього впливу. В соціальні медіа потрапляє інформація про справжнє ставлення користувачів, клієнтів до організації в цілому і до кожного її кроку: зміни в продукції, зміни в політиці взаємодії з клієнтами, зміни цін, нова рекламна компанія, тощо. Вчасно взята інформація із соціальних медіа допоможе скоректувати дії і плани компанії щоб краще задовольнити клієнтів та отримати більший прибуток.

Для організації можна виділити такі фактори впливу під час аналізу інформації з соціальних медіа:

- прогностична функція - f_1 ;
- оцінка досвіду конкурентів - f_2 ;
- дослідження реакції на зміни в роботі організації чи в її продукції - f_3 ;
- оцінка успішності рекламної кампанії - R ;
- показник ефективності протидії негативному інформаційному впливу конкурентів - G ;
- оцінка якості роботи організації, або окремих її частин - Q ;
- показник наявності джерел витоків секретної інформації - S ;
- показник знаходження нових клієнтів, або партнерів - N .

Прогностична функція f_1 має за мету прогнозування, як дослідження тенденцій. Оцінка досвіду конкурентів f_2 дозволяє уникати помилок, які зробили конкуренти і переймати позитивний досвід. Також можна відстежувати і помилки своєї організації, отримувати зворотній зв'язок з користувачами.

Задачами роботи є: дослідження методів збору комерційної інформації в соціальних медіа, виділення найбільш значущих факторів впливу, побудова формальної моделі для збору та аналізу інформації в соціальних медіа, дослідження можливостей автоматизації цих процесів та розробка програмної

системи для автоматизації збору та аналізу вказаної інформації.

2. Побудова формальної моделі збору та аналізу інформації в соціальних медіа

Про успішність рекламної компанії можна зробити висновки на основі кількості згадувань організації в соціальних медіа за певний період часу T та інтонаційного відтінку цих згадувань:

$$R = f(r_i \cdot k_i), \quad (1)$$

де r_i – характеристика згадування певного (i -го) інтонаційного відтінку, k_i – кількість згадувань певного (i -го) інтонаційного відтінку. r_i задається за допомогою експертних оцінок та може бути задане, наприклад, як дискретна множина оцінок $r_i = \{-10, -9, \dots, -1, 0, 1, 2, \dots, 10\}$. Тоді інтегральна оцінка успішності рекламної компанії R буде певним числом, що відображає ефективність рекламної роботи, і може бути як додатнім, так і від'ємним.

В окремих сферах бізнес-діяльності має місце негативний інформаційний вплив з боку конкурентів. В рамках даного дослідження будемо вважати, що ефективність протидії негативному впливу з боку конкурентів можна оцінити таким чином

$$G = 1 / \sum_j p_j, \quad (2)$$

де p_j – j -й негативний відгук в соціальних медіа.

На основі відгуків користувачів про якість продуктів та послуг, можна зробити висновки про ефективність роботи конкретних підрозділів організації та вжити заходів щодо підвищення якості. Оцінка якості Q також пов'язана з кількістю позитивних та негативних відгуків про роботу. Тому будемо вважати, що інтегральна характеристика R опосередковано враховує показники G та Q .

В окремих випадках інформація з соціальних медіа може допомогти знайти джерело витоків секретної інформації, або інформації, яка охороняється законом та вжити відповідних заходів.

Можна виділити наступні методи: метод пошуку; метод соціальної інтеграції; метод співбесіди.

Найочевиднішим способом збору інформації в соціальних медіа є пошук по ключовим словам, наприклад за назвою бренду чи організації. Цим методом можна досягти виконання всіх перелічених функцій, проте існують недоліки: велика кількість інформації, яку необхідно обробити, велика кількість джерел, між якими постійно треба перемикаєтись, щоб найбільше утримати в полі зору. Основні джерела це різноманітні соціальні мережі, в тому числі twitter.com, а також блогосфера і форуми.

При використанні методу пошуку не потребують детального аналізу наданих спеціалістами функцій $\{f_3, f_4, f_5, f_8\}$.

Найбільш поширеним є метод соціальної інтеграції, тобто взаємодія представників організації з

клієнтами і партнерами через соціальні мережі та блогосферу. Цей метод дає можливість спілкування зі споживачами, проведення опитів (голосувань), проте не виконує багатьох з можливостей, які можна досягти аналізом інформації в соціальних медіа. Не потребують детального аналізу функції $\{f_3, f_6, f_8\}$.

Менш поширеним є метод співбесіди, тобто проведення персональних інтерв'ю з рядом споживачів, які дадуть на це згоду. В соціальних медіа цей метод можна прирівняти до спаму, оскільки не кожен споживач захоче вступити в діалог. При використанні методу співбесіди не потребують детального аналізу функцій $\{f_2, f_3, f_4, f_6\}$, проте якісне виконання цих функцій вказаним методом потребує дуже великих затрат часу і майже не піддається автоматизації.

Отже, при використанні комбінації всіх методів можна виконати функції $\{f_2, f_3, f_4, f_5, f_6, f_8\}$ без необхідності детального аналізу великих обсягів даних спеціалістами-аналітиками, а при підключенні аналітиків можуть бути виконані і усі інші функції $\{f_1, f_7\}$. Найбільшу кількість задач розв'язує метод пошуку, але для практичного використання він потребує максимальної автоматизації.

Таким чином, за допомогою (1) та (2) можна отримати числові оцінки успішності проведеної рекламної компанії.

3. Дослідження можливості автоматизації процесі збору та аналізу інформації в соціальних медіа

Найпопулярніші соціальні мережі, серед яких [facebook](http://facebook.com), [twitter](http://twitter.com), [vk](http://vk.com), надають доступ до API (Application Programming Interface), в якому є можливість пошуку. Обмеження по кількості запитів – 1-2 запити за секунду, що дозволяє в реальному часі стежити за появою нових згадувань. Значна частина мережі Інтернет проіндексована пошуковою системою Google, що дозволяє виконувати пошук майже по всій блогосфері. До функціоналу пошукової системи входить можливість обирати давність згадування, щоб отримувати нові згадування, не повторюючи декілька разів обробку старих. Проте кількість пошукових запитів з одного IP дуже обмежена, тому для автоматизації збору інформації цим шляхом необхідно використовувати множину IP-адрес. В табл. 1 наведено порівняння основних джерел для автоматичного збору інформації.

Для розробки системи серед джерел були обрані 5 найефективніших з розглянутих, переважно з API: Google Search, Facebook, Twitter, "Вконтакте", Google+. Можливо використовувати інші системи в залежності від територіального знаходження спільноти, активність якої досліджується в рамках конкретної задачі. Наприклад, для Китаю більш підійшла б пошукова система Baidu, тому що там вона на першому місці за популярністю.

Порівняння основних джерел для автоматичного збору інформації

| Джерело | API | Обмеження і недоліки | Переваги |
|-----------------|--|--|---|
| Google Search | Ні | Обмежена глибина пошуку (перші 200 результатів запиту), обмежена кількість запитів з одного IP | Можливість фільтрувати результати за часом, щоб виділяти нові згадування. Великий обсяг |
| Facebook | Graph API | Допускається 600 звернень за 600 секунд з одним token та IP [3]. Багато з пошукових методів API отримали статус Depre-cated (“застарілі”) без заміни новими методами | Великий обсяг пошуку, можливість сортування результатів за часом додавання, контент різного типу, включаючи відеозаписи та заходи |
| Twitter | Twitter REST API | Допускається до 450 пошукових звернень кожні 900 секунд [4] | Короткі згадування, передаються в JSON форматі – відсутня необхідність виділяти їх з веб-сторінки |
| Вконтакте | API vk.com | API допускає до 3 звернень в секунду[5] | Контент різного типу. Пошук знаходить не тільки точне входження, але і написання на інших мовах і відмінках |
| liveinternet.ru | API з можливістю пошуку по блогах відсутнє | Невеликий обсяг інформації | Пошук одночасно за записами і коментарями. Можливість сортування за часом додавання |
| livejournal.com | API з можливістю пошуку по блогах відсутнє | Відсутність власної системи пошуку, використовується Яндекс. Обмежена глибина видачі результатів – 100 сторінок | Пошук розділено на записи і коментарі окремо. Великий обсяг інформації |
| Google+ | HTTP API | Безкоштовне використання має обмеження: 10000 звернень на добу, не частіше 5 за секунду. Відсутня можливість пошуку за коментарями | JSON API повертає результати у зручному вигляді, включаючи інформацію про медіа додатки, які прикріплені до запису |

4. Розробка моделі системи збору та аналізу інформації

Визначити емоційне забарвлення згадувань (позитивне, нейтральне, негативне) в більшості випадків може тільки оператор системи, оскільки це забарвлення суб’єктивне, залежить від контексту і мети аналізу. Наприклад, в записі блогу є банк, згадування якого досліджуються. Цей запис має гостро негативне забарвлення, проте негатив направлений не у відношенні до банку, а у відношенні до іншого клієнта банку, який не виконав умов договору з автором знайденого запису.

Таким чином, при автоматичному визначенні емоційного забарвлення система з великою вірогідністю буде помилятися, приписавши негативний відтінок, коли насправді по відношенню до мети аналізу правильний нейтральний, а людина-оператор зможе зрозуміти, до кого саме направлені емоції – до об’єкта дослідження, чи ні.

Бувають ще складніші для аналізу згадування, коли лексично забарвлення нейтральне, а зміст для різних людей має позитивне, або негативне значення, в залежності політики організації, персонального ставлення до події, політичних поглядів, світосприйняття, релігійних та інших переконань.

Крім того, згадування можуть бути представлені не тільки в текстовому вигляді, а і у мультиме-

дійному, без можливості конвертації в текст. В залежності від мети збору інформації, вона може потребувати додаткової перевірки, або поверхневого аналізу оператором системи. Наприклад, оператор може відсіяти контент, який очевидно сворено автоматичними системами, або який не відноситься до об’єкта дослідження, хоч і включає ключові слова, що використовуються для пошуку. Отже, точне визначення емоційного значення згадування в контексті задачі можливе тільки оператором системи, який знає мету даного етапу збору інформації. Проте існує можливість спростити роботу оператора завдяки покращенню ергономічності інтерфейсу системи та можливостям для групових операцій.

Представимо дії оператора системи у вигляді UML Use Case діаграми (рис. 1).

Оператор системи задає пошукові запити, за якими буде проводитись збір і аналіз інформації. Після збору він визначає емоційне забарвлення згадувань і система формує звіти по кожному запиту. Звіти включають візуалізації про зміни кількості згадувань та зміни відсоткового співвідношення емоційного забарвлення протягом періоду часу (тиждень, місяць, рік, весь час), що розглядається. Перегляд знайдених згадувань і визначення їх емоційних значень в контексті задачі – це основна робота оператора системи, яка може повторюватись будь-який період, допоки поставлена задача не буде ви-

рішена. Оператор може одразу виконувати реакційні дії стосовно знайдених згадувань (поза системою збору інформації).

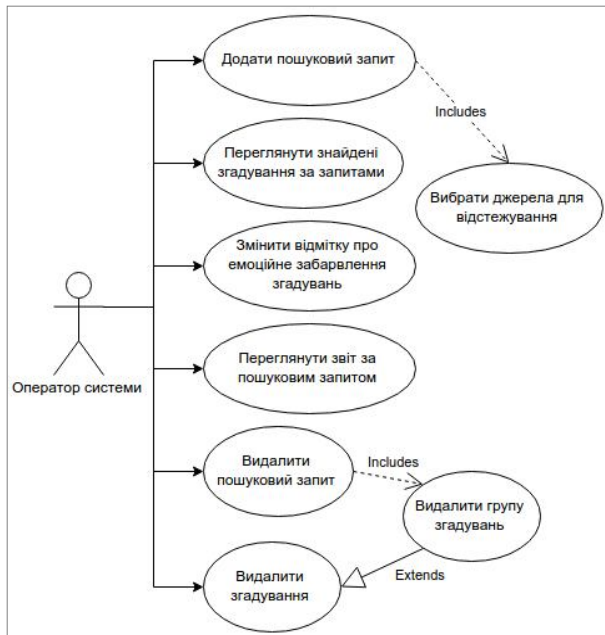


Рис. 1. UML Use Case діаграма системи збору і обробки інформації

На рис. 2 наведена діаграма потоків даних системи збору інформації. Для кожного джерела інформації створюється окрема підсистема обробки і окрема база даних. Кожне джерело має окреме представлення інформації, специфічні обмеження та потребує індивідуального підходу.

Для того щоб забезпечити максимальну частоту перевірки згадувань за пошуковим запитом, ігноруючи при цьому згадування, які вже були оброблені раніше, кожна підсистема потребує постійної роботи з базою даних. Джерел може бути велика кількість, це спричинить величезне навантаження на сервер баз даних.

Щоб цього уникнути, кожна підсистема користується окремою базою на своєму сервері, яка синхронізується з основною базою. Таким чином, в основній базі присутня вся інформація, проте навантаження розподілене між серверами, на яких працюють підсистеми.

Розроблено програмну систему для автоматизації збору інформації в реальному часі. Робота системи довела ефективність приведеного методу пошуку та можливість його автоматизації.

5. Аналіз роботи системи збору та аналізу інформації із соціальних медіа

Шляхом автоматизації пошукового методу збору інформації була зібрана інформація про декілька організацій.

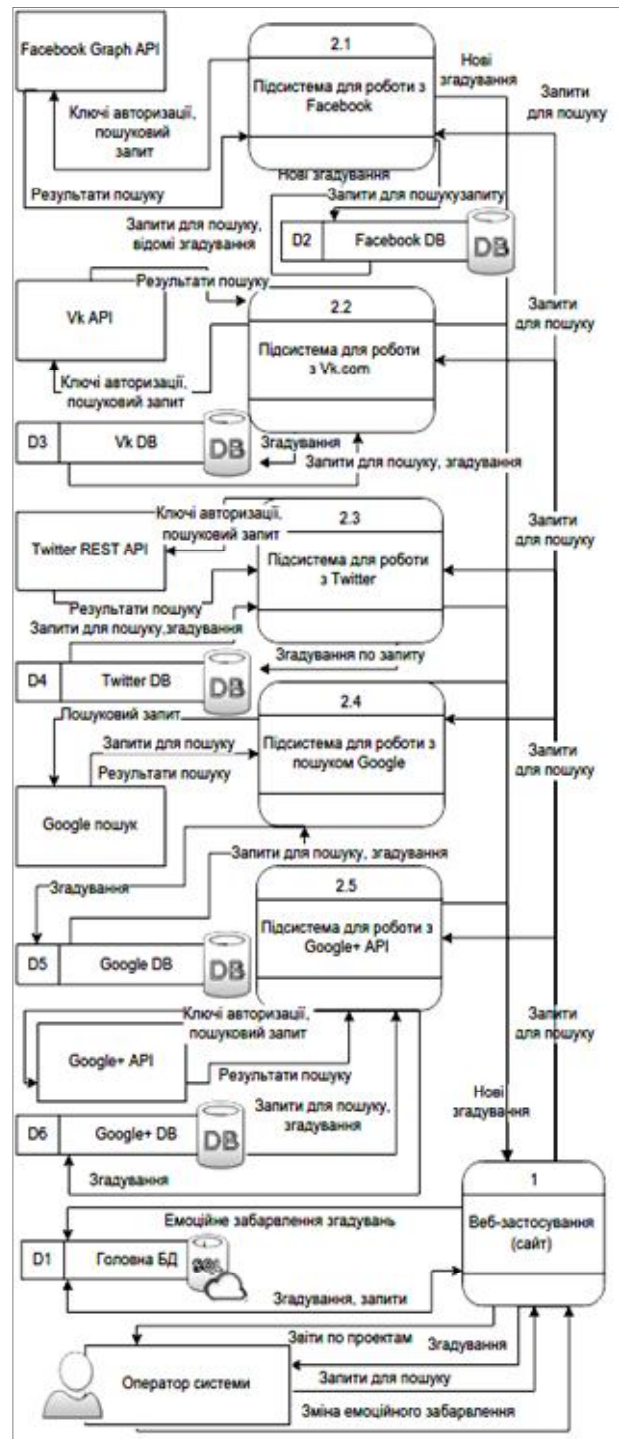


Рис. 2. DFD діаграма системи збору інформації

Інформація збиралась не з метою впливу на роботу цих організацій, а тільки з метою дослідження методу.

Однією з підслідних організацій був один досить відомий український банк, візьмомо його для прикладу. В середньому 1834 згадувань в соціальних медіа на день. 96.2% згадувань нейтральні, 3.5% - позитивні, решта - 0.3% негативні. Під негативними згадуваннями слід розуміти обурення в соціальних медіа. Відсоткове співвідношення наглядно показано на рис. 3.

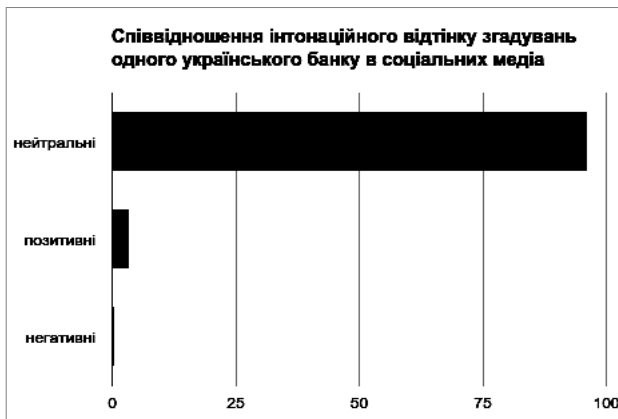


Рис. 3. Співвідношення інтонаційного відтінку згадувань одного українського банку в соціальних медіа

Слід зазначити, що аналіз зібраних даних проводився в ручному режимі, тобто дослідник вирішував, яке саме інтонаційне забарвлення має те чи інше згадування.

Таким чином, можемо зробити висновок про успішність проаналізованої організації та її позитивний імідж в соціальних медіа. Окрім оцінки успішності організації, в результаті дослідження були знайдені докази негативного інформаційного впливу з боку конкурентів: з десятків різних облікових записів в twitter розсилалися записи з абсолютно ідентичним текстом, негативним по відношенню до даної організації.

В той же час було помічено і протилежне: дана організація розповсюджує про себе позитивну інформацію, використовуючи однаковий текст і безліч різних облікових записів та джерел.

Висновки

У статті розглядаються підходи до автоматизації збору інформації в соціальних медіа та методи її застосування для підвищення прибутку організації.

В результаті дослідження запропонована формальна модель збору та аналізу інформації із соціальних медіа.

ПРОГРАММНАЯ СИСТЕМА ДЛЯ СБОРА И АНАЛИЗА ИНФОРМАЦИИ В СОЦИАЛЬНЫХ МЕДИА

А.Л. Ерохин, Н.Д. Одарченко

В статье предлагается усовершенствованный подход к сбору и анализу информации в социальных медиа, которыми пользуются большие предприятия с коммерческой целью. Для проведения анализа разработана программная система с целью автоматизации сбора информации.

Ключевые слова: социальные медиа, модели сбора и анализа информации, бизнес-исследования.

SOFTWARE SYSTEM FOR COLLECTING AND ANALYZING INFORMATION ON SOCIAL MEDIA

A.L. Yerokhin, M.D. Odarchenko

The article provides an improved approach to the collection and analysis of information in social media, used by large enterprises for commercial purposes. For analysis software system designed to automate the collection of information.

Keywords: social media, models gather and analyze information, business research.

Спроекована і розроблена система автоматизації збору і обробки інформації.

Розроблена програмна система відкриває широкі можливості для збору і комерційного застосування інформації з соціальних медіа. При використанні такої системи пошуковий метод значно ефективніший ніж методи соціальної інтеграції та опитування, оскільки не потребує ініціативи від осіб, про яких, або від яких необхідно отримати інформацію та охоплює різні джерела інформації.

Були розглянуті можливості для аналізу, що дає зібрана із соціальних медіа інформація.

Список літератури

1. Падафет Ю.Г. Механізм формування іміджу державної установи: автореф. дис. на здобуття наук. ступеня канд. наук з держ. упр.: 25.00.02 [Електронний ресурс] / Ю.Г. Падафет; Нац. акад. держ. упр. при Президентові України, Харк. регіон. ін-т держ. упр. – Х., 2005. – 16 с.
2. “Соціальні медіа” у Вікіпедії [Електронний ресурс]. – Режим доступу до ресурсу: http://uk.wikipedia.org/wiki/Соціальні_медіа.
3. Google група “Koala & The Facebook Graph API”, тема “Calls to stream have exceeded the rate of 600 calls per 600 seconds” [Електронний ресурс]. – Режим доступу до ресурсу: <https://groups.google.com/forum/#!topic/koala-users/3cb9TEoH8mA>.
4. Twitter API Rate Limits [Електронний ресурс]. – Режим доступу до ресурсу: <https://dev.twitter.com/rest/public/rate-limits>.
5. Вконтакте – “Ограничения и рекомендации” [Електронний ресурс]. – Режим доступу до ресурсу: https://vk.com/dev/api_requests.
6. Теория моделирование потоков данных (процесов) [Електронний ресурс]. – Режим доступу до ресурсу: http://www.business-process.ru/designing/methodology/dfd/dfd_theory_dfd.html.

Надійшла до редколегії 2.06.2015

Рецензент: д-р техн. наук, проф. І.П. Захаров, Харківський національний університет радіоелектроніки, Харків.