УДК 004.91

Коваленко Т.В.,
Додонов В.О.,
Ланде Д.В., д.т.н.

ДОСЛІДЖЕННЯ ЗВ'ЯЗКІВ АГЕНТІВ В ІНФОРМАЦІЙНІЙ МЕРЕЖІ МІКРОБЛОГІВ

Інститут проблем реєстрації інформації НАН України

2005ste@ukr.net
ipri@ipri.kiev.ua

Розглянуто питання реалізації краудсорсингового підходу до моделювання інформаційно-аналітичної системи, який реалізується у вигляді відповідної інформаційної технології. Представлена технологія виявлення найбільш актуальних інформаційних повідомлень, які містять посилання на найважливіші інформаційні ресурси в мережі Інтернет. Результати моделювання підтверджені шляхом дослідження реальної мережі мікроблогів Twitter. Описано етапи створення корпоративної системи моніторингу мережевих інформаційних ресурсів, склад яких визначається посиланнями в мікроблогах. Наведено переваги такого підходу

Ключові слова: мережа зв'язків, активний агент, моделювання, мережа мультиблогів, база даних, моніторинг

Введення

Метою цієї роботи є обгрунтування можливості застосування краудсорсингового підходу при побудові інформаційного забезпечення інформаційно-аналітичних систем, що можуть застосовуються, зокрема, для формування сценаріїв інформаційного впливу, вирішення задач підтримки прийняття рішень. У роботі розглядаються розглядати застосування мультиагентного моделювання для дослідження параметрів масиву актуальних інформаційних ресурсів, отриманих шляхом моніторингу соціальних мереж і Інтернеті.

Обсяги інформаційних ресурсів в веб-просторі сьогодні ускладнюють оперативне отримання необхідної користувачам інформації навіть за допомогою найпотужніших мережевих пошукових систем (Google, Baidu, Яandex і ін.) [1] Одним із шляхів вирішення цієї проблеми є підключення оцінок великого числа людей, експертів в предметній області. Така можливість, яку можна назвати краудсорсинговою, відкривається за допомогою змістовного аналізу соціальних мереж, де користувачі «голосують» за ті чи інші інформаційні матеріали, встановлюючи на них гіперпосилання. Особливо це актуально в сегментних предметних областях, які відповідають потребам корпоративних користувачів. Незважаючи на те, що аналіз соціальних мереж сам по собі є складною науково-технічною задачею, існуючі пошукові можливості деяких з них, дають надію на рішення названої проблеми.

Разом з тим, процеси поширення інформації в мережах, що містить гіперпосилання на інформаційні ресурси, вимагали детального аналізу. Моделювання дозволяє досліджувати інформаційні процеси, виявляти закономірності, які можуть використовуватися як при вивченні механізмів передачі інформації в таких мережах, так і рівня її впливу на людей [3].

У цій роботі пропонується підхід до створення корпоративної системи моніторингу мережевих інформаційних ресурсів, склад яких визначається посиланнями в мережі мікроблогів Twitter [2].

Модель розповсюдження повідомлень в мережі мікроблогів

Для моделі розповсюдження інформації в мережі мікроблогів обирається мультиагентний підхід. Для створення відповідної моделі, перш за все, необхідно сформувати близький до

реальності віртуальний інформаційний простір, в якому функціонують агенти, з якими асоціюються окремі повідомлення в соціальній мережі, і які інкапсулюють в себе гіперпосилання на інформаційні ресурси мережі Інтернет [4-5]. Передбачається, що окремі агенти можуть [6]:

1) самозароджуватися;

2) породжувати нових агентів шляхом репостінгу (repost);

3) «вмирати» – зникати з простору агентів;

4) отримувати лайки (like) від інших агентів.

Агент має «енергію», що залежать від часу його існування, авторитетності (посилань на нього) і плодючості (кількості породжених їм агентів, репостів). Варіювання відповідними параметрами управління дозволили змоделювати профілі поведінки інформаційних сюжетів, що породжуються в цій моделі. В результаті проведених досліджень було досліджено еволюцію мультиагентної системи, знайдені аналогії з реальними тематичними інформаційними потоками. Були виявлені статистичні закономірності, що відносяться до життєвого циклу окремих повідомлень, розподіл яких, як було виявлено, відповідає розподілу Вейбулла. Дані моделювання були перевірені шляхом дослідження реальної мережі мікроблогів Twitter. Збіг результатів моделювання і параметрів розподілу реальної мережі дозволяють говорити щодо закономірностей, властивих реальним мережам, а також щодо адекватності моделі.

Моделювання динаміки всього інформаційного потоку починається з одного агента. Поява нового агента можливо двома способами. Перший полягає в копіюванні існуючого агента за допомогою операції «репост». Також можливо самозародження агента, що відповідає публікації нового повідомлення. У кожний момент часу з певними ймовірностями, з кожним з агентів, може статися деяка подія. Агент з'являється з початковим значенням енергії E_0 і далі його енергія змінюється в залежності від подій, які з ним відбуваються. Будемо вважати, що можливі дві події: лайк і репост. За одиницю часу може статися одна з цих подій, обидві одночасно, або не відбутися жодної.

Позначимо e_t значення енергії агента на момент часу *t*. Тоді значення енергії в наступний момент часу можна записати в такий спосіб:

$$e_{t+1} = e_t + d_t,$$

де d_t є випадковою величиною із значенням у {-1, 0, 1, 2}. Згідно з правилами зміни енергії, введеними вище, збільшення енергії на 2 відповідає тому, що відбулися одночасно лайк і репост; збільшення на 1 – стався тільки репост; енергія не змінюється, якщо був лайк; і зменшується на 1, якщо не відбулося ні однієї з подій.

Так як значення енергії в наступний момент часу залежить тільки від значення енергії в попередній момент часу, то стохастична послідовність $(e_0, e_1, ..., e_t, ...)$ є марківським ланцюгом з перехідними ймовірностями p_{ij}. В результаті моделювання було визначено, що розподіл часу життя, а відповідно й кількість лайків і репостів в даній моделі відповідає щільності розподілу Вейбулла [8]:

$$f(x) = \begin{cases} \frac{k}{l}\left(\frac{x}{l}\right)^{k-1} e^{-\left(\frac{x}{l}\right)^k}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Параметри розподілу Вейбулла k і l були отримані методом максимальної правдоподібності. При зазначених початкових параметрах, отримані значення $k = 1.9, \quad l = 3.8$.

Отримані результати моделювання порівнювалися з проведеними результатами дослідження життєвого циклу новинних повідомлень в мережі мікроблогів Twitter, де, зокрема, аналізувалися показники зростання кількості спеціальних репостів (ретвіттів) обраних повідомлень.

База даних інформаційних матеріалів з мережі Інтернет

Авторами впродовж травня 2016 року здійснювався експеримент із збору повідомлень з мережі мікроблогів Twitter, для чого в пошуковому інтерфейсі цієї мережі на періодичній основі оброблявся пакет із 100 запитів із банківської і бізнес-тематики. В результаті були отримані наступні кількісні дані, пов'язані з кількістю інформаційних ресурсів мережі Інтернет, на які були вказані посилання з Twitter-повідомлень:

1. Відскановано близько 100 тисяч повідомлень по 100 елементарним запитам до Twitter за травень 2016 р.

2. 58% повідомлень містять гіперпосилання на веб-ресурси мережі Інтернет.

3. Кількість унікальних гіперпосилань 48%.

4. Функція розподілу кількості гіперпосилань на одні й ті-ж джерела - степенева.

5. Є проблеми ідентифікації зовнішніх посилань, пов'язана, перш за все, з використанням «коротких посилань» – переадресації з такими базовими адресами:

- http://migre.me/
- http://bit.ly/
- http://ow.ly/
- http://tinyurl.com/
- https://lnkd.in/
- https://goo.gl/
- http://wp.me/
- http://j.mp/
- http://dlvr.it/

6. Виявлено найбільш часто цитовані ресурси мережі Інтернет:

- https://youtu.be (https://www.youtube.com)
- http://fb.me/ (https://www.facebook.com/) – публічні сторінки
- https://vk.com/
- https://twitter.com/ - посилання на ту ж саму соціальну мережу
- https://plus.google.com/
- http://livejournal.com/

Таким чином, практика показала, що розподіл інкапсульованих в повідомлення соціальних мереж посилань на інформаційні ресурси мережі Інтернет відповідає степеневому розподілу (рис. 1).

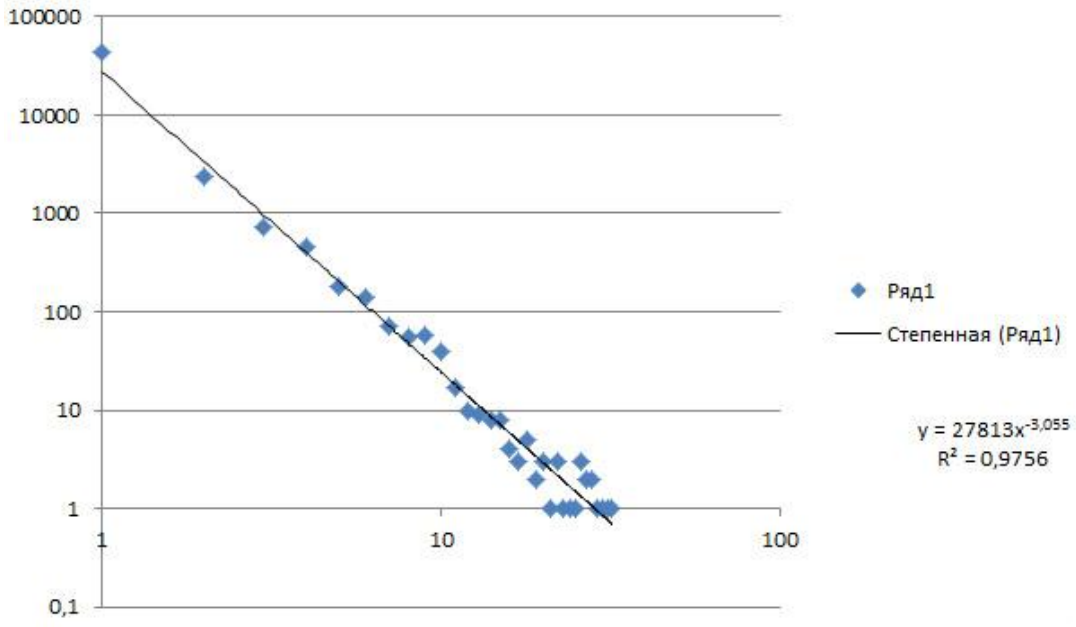

Рис. 1. Розподіл кількості повідомлень,
що містять посилання на зовнішні інформаційні ресурси

На основі отриманої інформації про розподіл повідомлень, на які реалізовані посилання з мережі мікроблогів, пропонується наступна «краудсорсингова» схема формування бази даних корпоративної системи моніторингу мережевих інформаційних ресурсів (рис. 2).

Рис. 2. Краудсорсингова схема формування бази даних корпоративної системи моніторингу мережевих інформаційних ресурсів

Наведемо основні етапи цього процедури:

1. В інтересах корпоративного користувача створюються запити до мережі мікроблогів, наприклад: Банки України; ДіамантБанк; Укрсиббанк; ПриватБанк; Банк Хрещатик; Платинум Банк; Кредит Дніпро

2. Сформований «широкий» пакет запитів передається програмі сканування мережі мікроблогів, в результаті чого на корпоративний сервер на регулярній надходять повідомлення, формально релевантні цим запитам.

3. Витяг з відсканованих повідомлень мережі мікроблогів гіперпосилань на зовнішні мережеві інформаційні ресурси.

4. Обробка гіперпосилань, розкриття «коротких адрес», сортування гіперпосилань, ранжирування окремих документів і зовнішніх джерел.

5. Сканування зовнішніх інформаційних ресурсів, відповідних виділеним гіперпосиланням, первинна обробка отриманих документів, приведення їх до вхідного формату використовуваної корпоративної інформаційно-аналітичної системи.

6. Завантаження сформованого інформаційного потоку в корпоративну інформаційно-аналітичну систему, надання в доступ корпоративним користувачам.

Мережа взаємозв'язку агентів

На основі сформованої бази даних, що містять повідомлення з мережі мікроблогів, які містять гіперпосилання на інформаційні ресурси веб-простору, формується графа взаємозв'язку аккаунтів авторів блогів. Для цього розглядається матриця $A = \left\| a_{ij} \right\|$, елементи якої відповідають наявності в повідомленні з аккаунта i гіперпосилання на ресурс j. До розгляду беруться лише такі аккаунти, тексти повідомлень з яких містять у себе гіперпосилання (у прикладі, що розглядається – гіперпосилання містили близько 58% всіх повідомлень, вони відповідали лише 10% аккаунтів).

Як матриця інцидентності графа взаємозв'язку аккаунтів авторів блогів розглядається матриця $C = A \cdot A^T$. Статистичні показники графа взаємозв'язку аккаунтів авторів блогів наведено в Табл. 1.

Таблиця 1. *Статистичні показники графа взаємозв'язку аккаунтів авторів блогів*

№ з/п	Параметр	Значення
1	Середній ступінь вузла	4,785
2	Діаметр графа	17
3	Щільність графа	0,004
4	Зв'язні компоненти	277
5	Середній коефіцієнт кластеризації	0,871
6	Середня довжина шляху	5,572

Відображення структури графа, вузлам якого відповідають аккаунти, а ознакою зв'язку – наявність посилань на ті ж самі інформаційні ресурси, що здійснювалося за допомогою програми Gephi (http://gephi.com/) наведено на рис. 3. Засоби Gephi дозволяють виявляти кластери найбільш зв'язаних вузлів, що й було здійснено рис. 4. На Рис. 5. наведено фрагмент графа з позначеними назвами аккаунтів вузлами. В табл. 2. Наведено перелік найбільш вагомих аккаунтів-вузлів графа взаємозв'язку аккаунтів авторів блогів.

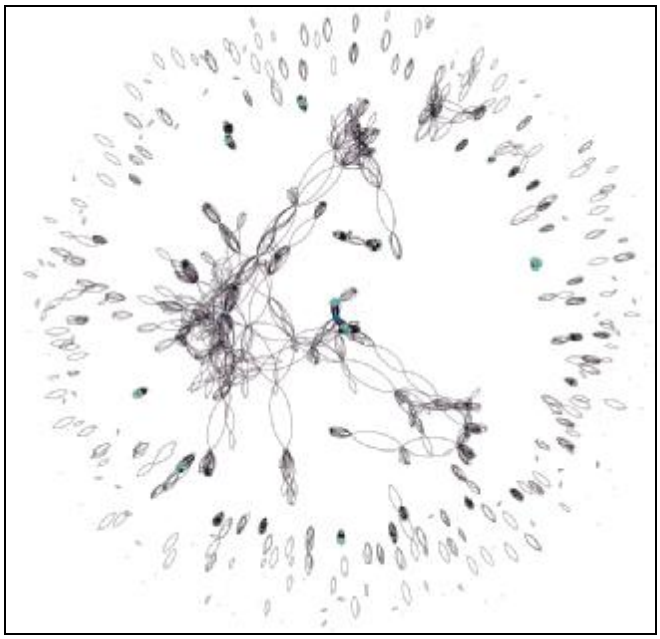

Рис. 3. Структура графа взаємозв'язку аккаунтів авторів блогів

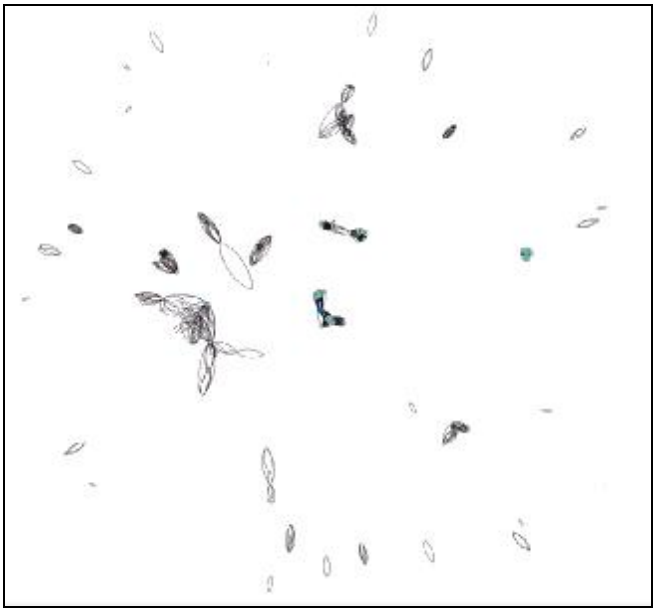

Рис. 4. Основні кластери графа взаємозв'язку аккаунтів авторів блогів

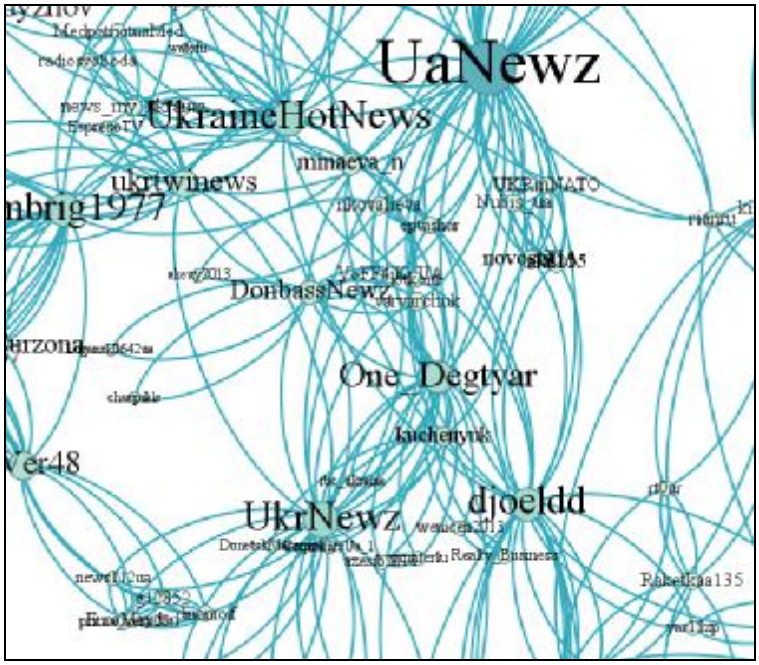

Рис. 5. Фрагмент графа взаємозв'язку аккаунтів авторів блогів

Таблиця 2. *Ступені найбільш вагомих вузлів графа взаємозв'язку аккаунтів авторів блогів*

№ з/п	Аккаунт	Ступень
1	Kaplja_O	43
2	UaNewz	27
3	Alekzzzzz	25
4	Surkova_Tatyana	23
5	dreamcatmaster	23
6	morozovbest	23
7	pochtas08010914	22
8	AlexSuherevskij	21
9	karabley	21
10	Porzia27	21
11	dostoverkin	21
12	Benderbr_rb	21
13	w3c_user	21
14	Archelmus	21
15	world_politika	21
16	alexd2w	21
17	raseyanin	21
18	You_Are_Future	21
19	uhbif18	21
20	UkrNewz	20

Переваги методу

Представлений підхід до формування баз даних на основі врахування посилань в мікроблогах на інформаційні ресурси забезпечує, поряд з істотним скороченням охоплення інформаційного простору, такі переваги:

1. Оперативність - інформаційне повідомлення потрапляє в базу даних інформаційно-аналітичної системи в режимі

реального часу в мірі того, як на нього зробив посилання перший користувач.

2. Охоплення головних інформаційних матеріалів за темою. Врахування думки аудиторії зацікавлених користувачів, контент повідомлень яких задовольняє широким корпоративним запитам. Можливість ранжирування інформаційних матеріалів виходячи з інтересів користувачів соціальних мереж.

3. Компактність баз даних, а, отже, зручність доступу кінцевих користувачів. Передбачуваність обсягів баз даних, динаміки їх наповнення.

4. Технологічна сумісність з існуючими інформаційно-аналітичними системами і системами контент-моніторингу.

5. Можливість виявлення інформаційних кампаній, операцій [7], вирішення задач підтримки прийняття рішень.

Висновки

В результаті описаних досліджень:

1. Побудовано мультиагентну модель поширення інформаційних повідомлень, які містять посилання на інформаційні ресурси в мережі Інтернет. Результати моделювання перевірені шляхом дослідження реальної мережі мікроблогів Twitter.

2. Знайдені закономірності можуть використовуватися при формуванні баз даних інформаційно-аналітичних систем, при вивченні аномалій в статистиці посилань на окремі інформаційні матеріали.

3. Досліджено закономірність розподілу кількості повідомлень мережі мікроблогів, що містять посилання на зовнішні інформаційні ресурси

4. Побудовано інформаційну технологію краудсорсингового формування бази даних за результатами моніторингу мережі мікроблогів.

5. Запропоновано метод побудови графа взаємозв'язку аккаунтів авторів блогів, що посилаються на ті ж самі інформаційні ресурси.

6. Запропоновано засоби візуалізації і кластеризації графа взаємозв'язку аккаунтів авторів блогів на бізі застосування програмного забезпечення Gephi.

Список літератури

1. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. - 264 с.

2. Li R., Lei K. H., Khadiwala R., Chang K. C. C.TEDAS: A Twitter-based Event Detection and Analysis System // Data Engineering (ICDE), 2012 IEEE 28th International Conference, 2012. – P. 1273 – 1276.

3. Додонов А.Г., Ландэ Д.В., Прищепа В.В., Путятин В.Г. Конкурентная разведка в компьютерных сетях. – К.: ИПРИ НАН Украины, 2013. – 248 с.

4. Woo J., Chen H. Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog // Springerplus, 2016. – № 22. – pp. 5-66.

5. Lerman K. Information Is Not a Virus, and Other Consequences of Human Cognitive Limits. Future Internet. 2016; 8(2):21.

6. Ландэ Д.В., Грайворонская А.Н., Березин Б.А. Мультиагентная модель распространения информации в социальной сети // Реєстрація, зберігання і обробка даних, 2016. – Т. 18. – № 1. – С. 70-77.

7. Додонов А.Г., Ландэ Д.В., Додонов В.А. Распознавание информационных операций: мультиагентный подход // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2016): материалы VI междунар. науч.-техн. конф. (Минск 18-20 февраля 2016 года) / – Минск: БГУИР, 2016. - С. 253-256.

Статтю подано до редакції 04.07.2016