УДК 519.711.3

О. О. Писарчук, Н. В. Стиров

МЕТОДИКА ФІЛЬТРАЦІЇ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ ІНТЕРНЕТ-ДЖЕРЕЛ ТА ЇХ КЛАСИФІКАЦІЯ

У статті сформовано методику фільтрації інформаційних повідомлень та їх класифікація. На конкретному прикладі показано можливості щодо її використання для реалізації завдань автоматизованої обробки інформації відкритих джерел інформаційно-телекомунікаційної мережі Інтернет.

Постановка проблеми. На теперішній час значного розвитку набули інформаційні технології. Основною складовою інформаційної сфери є віртуальний простір, прикладом його консолідуючого елемента може бути всесвітня мережа Інтернет, кількість користувачів якої зростає кожного дня, а обсяг доступної інформації становить сотні тисяч терабайт. Тому за сучасних умов інформаційна складова набуває дедалі більшої ваги і стає одним із найважливіших елементів забезпечення інформаційної безпеки. Виявлення інформаційних загроз потребує швидкої та якісної обробки значних обсягів інформації, що підвищує ефективність управлінських рішень стосовно реагування на загрози [1].

Реалізація зазначених заходів передбачає використання складних програмно-апаратних систем пошуку інформації, які містять об'ємні бази даних і знань. Тому забезпечення точності визначення спрямованості документів є достатньо складною проблемою для людини, що полягає у неоднозначності вибору відносно належності сегмента даних до однієї із категорій конкретного текстового документа. Це обумовлює необхідність обробки великої кількості текстових документів, суть яких може стосуватися всього спектра процесів життєдіяльності суспільства. Тому створення методики і відповідного програмного забезпечення, яке б надавало змогу фільтрувати та класифікувати необхідну інформацію мережі Інтернет, на сьогодні **актуальне** та необхідне науково-прикладне завдання.

Огляд останніх досліджень. Сучасний рівень розвитку програмно-апаратних засобів уможливив ведення баз даних оперативної інформації на різних рівнях управління. Для того, щоб вони сприяли прийняттю управлінських рішень, інформація повинна бути подана аналітику в потрібній формі, адже вона вміщує у собі великі потенційні можливості щодо виявлення корисної аналітичної інформації, на основі якої можна визначити приховані тенденції інформаційних загроз [2]. Для цього існує значна кількість систем глибинного аналізу текстів як вбудованих в інші більш комплексні системи, так і автономних. Вони розроблені на основі статистичного і лінгвістичного аналізу, а також штучного інтелекту, та призначені для проведення контентного аналізу, забезпечення навігації та пошуку в неструктурованих базах даних.

До існуючих програмних засобів цього напрямку можна віднести такі системи: Intelligent Miner for Text (IBM), TextAnalyst, Text Miner (SAS), SemioMap (Semio Corp.), Galaktika-ZOOM (корпорація «Галактика»), InfoStream (інформаційний центр «ЕЛВІСТІ»).

Такі системи мають розвинені графічні інтерфейси, надають доступ до різних джерел даних, функціонують в архітектурі клієнт-сервер та надають можливості зображення і візуалізації результату [8].

Основним їх недоліком є надлишковість знайденої інформації у відповіді на запит користувача, що є результатом дублювання інформації та її невідповідність запиту. Причиною цьому є те, що система не в змозі визначити, чи відповідає зміст даного документа інформаційним потребам конкретного користувача [3–4].

Метою статті є розробка методики фільтрації інформаційних повідомлень інтернет-джерел та їх класифікація для реалізації можливості виконання завдань автоматизованої обробки інформації відкритих джерел інформаційно-телекомунікаційної мережі Інтернет.

Виклад основного матеріалу. Автоматичні системи інформаційного пошуку використовують для зменшення «інформаційного перенавантаження». Пошукова система переглядає всі доступні інформаційні одиниці (документи) зі збірки і відбирає документи відповідно до інформаційного запиту. Оскільки реальні пошукові системи знаходять не всі відповідні документи, то можна говорити про точність пошукових систем. Результатом роботи пошукової системи є список відібраних документів, серед яких є відповідні до запиту. Для ідеальної пошукової системи список відібраних документів та відповідних до запиту повинні збігатися. У реальних пошукових системах у списках відібраних документів знаходяться і невідповідні до запиту [3]. Фільтрація інформаційних повідомлень здійснюється не тільки для визначення відповідності, але і для вирішення проблем, які пов'язані з неоднозначністю мови – один і той самий термін може позначати різні концепти, один і той же концепт може позначатись різними термінами.

Загалом зміст процедури фільтрації – це алгоритм, який, переглядаючи набір документів ($D_1, D_2, ..., D_n$), встановлює їх відповідність до пошукового запиту (ПЗ). Оскільки пошуковий термін зустрічається в документах різну кількість разів, можна говорити про різний ступінь відповідності до ПЗ. Цей алгоритм обчислює коефіцієнт відповідності (КВ) для кожного документа КВ (ПЗ, D_i), де $1 \le i \le n$. Технологізація інформаційних процесів актуалізує методи й методики пошуку та отримання інформації в новому інформаційному середовищі Інтернет, які відрізняються від традиційних і мають іншу класифікацію. Процедура отримання інформації в мережі Інтернет здійснюється поетапно (рис. 1) [3–5]:

I етап – визначення предмета пошуку (що необхідно знайти);

II етап – складання списку ключових слів (створення запиту);

III етап – вибір інформаційного простору (вибір місця для пошуку);

IV етап – визначення інструменту пошуку (вибір програмного забезпечення);

V етап – попередній пошук (результат першого запиту);

VI етап – аналіз одержаної інформації (вивчення одержаної інформації та корегування запиту);

VII етап – додатковий пошук інформації (здійснення подальшого пошуку, доки не знайдеться бажаний результат).

VIII етап – збереження знайденої інформації у спеціалізованій базі даних;

IX етап – обробка знайденої інформації (фільтрація);

X етап – відображення інформації в необхідній для користувача формі може здійснюватися у вигляді стовпчастої діаграми [2–4].

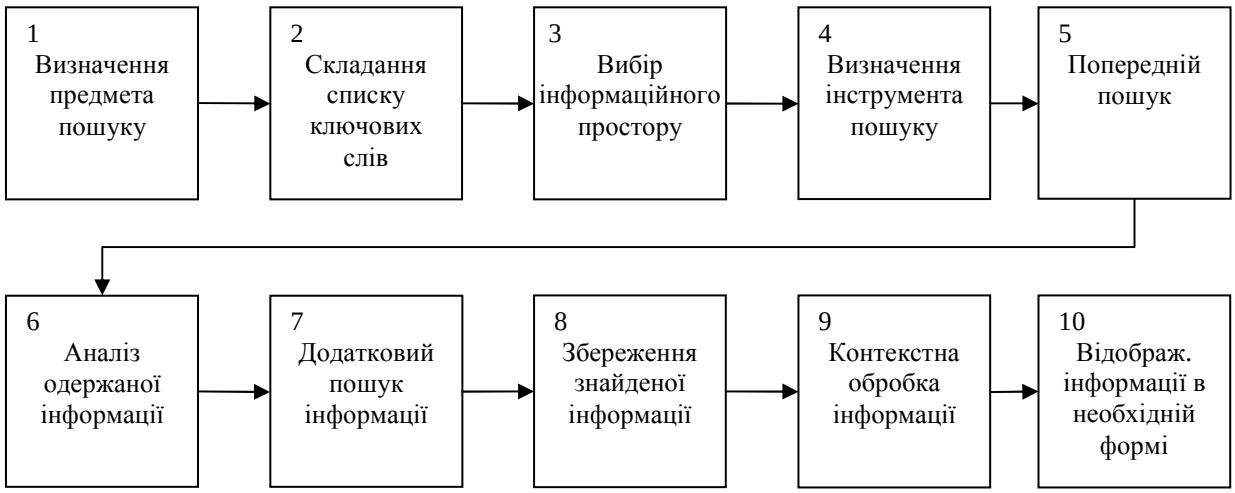

Рис. 1. Структурна схема алгоритму отримання інформації з мережі Інтернет

Бажано також визначити час, за який буде здійснюватися пошук інформації, оцінити альтернативні способи одержання інформації і ступінь її важливості у контексті визначеної проблеми.

Математичне підґрунтя алгоритмів обробки текстових повідомлень для блоку «Контекстна обробка інформації» включає такі методи: метод опорних векторів, метод найближчих сусідів, метод Роше.

У методі опорних векторів кожний документ відображається як вектор у багатовимірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексем зі словника аналізованих текстових масивів [3].

Для визначення належності до тематики скористаємось скалярним добутком векторів, звідки ми знаходимо кут між тематичним і текстовим векторами:

$$\cos\alpha = \frac{tf_1 tf_2 + idf_1 idf_2}{\sqrt{tf_1^{\,2} + idf_1^{\,2}}\sqrt{tf_2^{\,2} + idf_2^{\,2}}}, \qquad (1)$$

де tf_1 та idf_1 – координати опорного вектора тематики, а tf_2 та idf_2 – координати вектора текстового документа.

Документ для розгляду відбирається за мінімальним значенням величини $\cos\alpha$.

Алгоритм «найближчого сусіда» починається в довільній точці та поступово «відвідує» кожну найближчу точку, яка ще не була «відвідана». Пункти обходу плану послідовно включаються до маршруту. Причому кожен наступний пункт, що включається до маршруту, повинен бути найближчим до останнього вибраного пункту серед усіх інших, ще не включених до складу маршруту. Алгоритм завершується, коли «відвідано» всі точки [3–4]. Точками маршруту є контентні складові документа.

Вхідні дані для множини точок *V* розмірністю *N*. Вихідні дані: маршрут *T*, що складається з послідовності відвідування точок множини *V*. Послідовність роботи алгоритму «найближчого сусіда» включає:

1. Вибрати довільну точку V_1;
2. $T_1 = V_1$;
3. Для $i = 2$ до $i = N$ виконати;
4. Вибрати точку V_i, найближчу до точки T_{i-1};
5. $T_i = V_i$;

6. $T\,N + 1 = V_1$;

7. Кінець алгоритму та прийняття рішення про релевантне джерело [3]. Релевантне джерело обирається за мінімумом відстані V_i, T_{i-1}.

Метод Роше (класифікатор Роше) забезпечує рубрикацію документа, виходячи з його близькості до еталона рубрик. Еталоном для рубрики c є вектор $w = (w_1, w_2, \ldots w_i)$ у просторі ознак, обчислений за формулою:

$$w = \frac{a}{|POS(c)|} \sum_{d \in POS(c)} w_{di} - \frac{\beta}{|NEG(c)} \sum_{d \in NEG(c)} w_{di}, \qquad (2)$$

де $POS(c)$ та $NEG(c)$ – множина документів з навчальної вибірки, які належать і не належать рубриці c відповідно.

w_{di} – вага i-ї однини документа d.

Зазвичай, позитивні приклади набагато важливіші негативних, тому $a >> \beta$. Якщо $\beta = 0$, то еталоном рубрики буде просто центроїд усіх її документів. Документ визнається релевантним, якщо відстань до еталона мінімальна.

Кожний алгоритм має свої недоліки та переваги, тому на основі цих методів була сформована узагальнена схема їх застосування.

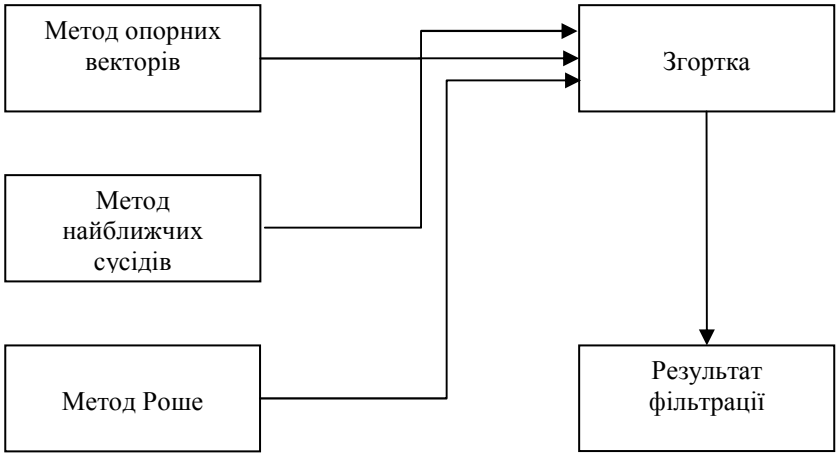

Рис. 2. Розширена структурна схема алгоритму фільтрації

Загалом задача фільтрації інтернет-повідомлень полягає у реалізації процесу сортування файлів (інформаційних) масивів за частотою появи в них заданого контенту та зворотньому аналізі заданого контенту (частота його появи), обмеженому в інформаційному полі. Кожен із наведених методів матиме свої результуючі оцінки щодо частоти появи контенту. Для підвищення точності результуючих оцінок пропонується отримати часткові оцінки за трьома методами та узагальнити їх у вигляді результуючої оцінки [7]. На підставі результату фільтрації визначається корисність інтернет-повідомлення відносно заданого контенту і може бути створена відповідна база даних джерел інформації для заданого контенту. Така інформація надається користувачу.

Сутність їх застосування полягає у цілісному зваженому врахуванні кожного із трьох рішень, отриманих за методами опорних векторів, «найближчого сусіда» та класифікатора Роше. Для цього застосовується багатокритерійний підхід, запропонований у роботі [7].

Згідно з методикою багатокритерійного порівняльного аналізу, перш за все, встановлюється образ (еталон), відносно якого здійснюється ідентифікація певного

об'єкта. Образом може виступати еталон результату фільтрації із сукупністю характеристик або сукупність критерійних вимог із визначення релевантного джерела. Об'єктом ідентифікації є довільний результат фільтрації. Надалі для образу й об'єкта ідентифікації встановлюється множина ознак, якими є значення частот появи повідомлення або критерійні вимоги до них. Тоді задачу ідентифікації можна розглядати як багатокритерійну з подальшим формуванням узагальнених ознак образу та об'єкта ідентифікації. Порівнюючи узагальнену ознаку об'єкта ідентифікації з еталоном можливо встановити міру відповідності між ними та визначити джерела інформаційних повідомлень за їх релевантністю [7].

Сукупність ознак еталона задається множиною T_i у кількості N:

$$S_E = \{T_i\},\ i = 1...N\,. \tag{3}$$

Сукупність джерел з ознаками характеризуються множинами

$$R_1\{T_j\}, R_2\{T_j\}, ..., R_m\{T_j\},\ j = 1...N_1, N_2, ..., N_k\,, \tag{4}$$

де $N_1, N_2, ..., N_k$ – кількість ознак m-го джерела.

При встановленні для ідентифікації ознак образу та об'єкта (джерела) у вигляді критерійних вимог множини (3), (4) трансформуються до вигляду

$$\begin{aligned} S_E &= \{T_1 \to extr, ..., T_i \to extr, ..., T_N \to extr\}, \\ R_1 &= \{T_{11} \to extr, ..., T_{1j} \to extr, ..., T_{1N_1} \to extr\}, \\ R_2 &= \{T_{21} \to extr, ..., T_{2j} \to extr, ..., T_{2N_2} \to extr\}, \\ R_m &= \{T_{m1} \to extr, ..., T_{mj} \to extr, ..., T_{mN_k} \to extr\}. \end{aligned} \tag{5}$$

Надалі здійснюється об'єднання сукупності ознак (4) до узагальненої ознаки згідно із згорткою проф. Вороніна А. М. для дискретних параметрів. Порівняно з іншими схемами оптимізації згортка за нелінійною схемою компромісів має такі переваги: оптимізаційні задачі розв'язуються за наявності обмежень, у межах яких гарантується унімодальність функції узагальненого критерію; відносно невелика обчислювальна складність алгоритму пошуку рішення.

Згортка для дискретно заданих частинних критеріїв має вигляд

$$Y(y_0) = \sum_{f=1}^{b} \gamma_{0f} \left(1 - y_{0f}\right)^{-1} \to min\,, \tag{6}$$

де $f = 1...b$ – кількість включених у згортку частинних критеріїв;

γ_{0f} – нормований ваговий коефіцієнт;

y_{0f} – нормативний частинний критерій [7].

Згідно із згорткою (5) формуються узагальнені ознаки образу та об'єкта (об'єктів) ідентифікації з нормованими ознаками, що у прийнятих позначеннях має вигляд

$$P_l = \sum_{j=1}^{N,N_1,N_2,\ldots,N_k} \gamma_{0lj}[1-T_{0lj}]^{-1},\ T_{0lj} = T_{lj}\left[\sum_{l=1}^{m} T_{lj}\right]^{-1},\ l = 1\ldots m,$$
$$P_E = \sum_{i=1}^{N} \gamma_{0i}[1-T_{0i}]^{-1},\ T_{0i} = T_i\left[\sum_{i=1}^{N} T_i\right]^{-1}. \quad (7)$$

Результатом застосування згортки (6) є сукупність узагальнених ознак для образу та об'єктів ідентифікації $P_E, P_1, P_2, \ldots, P_m$. Числові значення міри відповідності об'єкта ідентифікації образу розраховуються як відношення узагальнених оцінок:

$$W_1 = \frac{P_1}{P_E}, W_2 = \frac{P_2}{P_E}, \ldots, W_m = \frac{P_m}{P_E}\ . \quad (8)$$

Таким чином, розроблена методика базується і відрізняється поданням та розв'язком задачі визначення релевантних джерел з використанням багатокритерійного підходу. При цьому сукупність ознак об'єкта ідентифікації та образу приводиться до узагальненої ознаки згідно із згорткою за нелінійною схемою компромісів. Методика дозволяє розв'язувати задачу пошуку релевантного джерела як за переліком ознак, так і за вектором критерійних вимог [6–7].

Приклад застосування методики фільтрації інформаційних повідомлень мережі Інтернет полягає у розв'язанні типової задачі такого запиту. Нехай за стрічкою інформаційних повідомлень, отриманих з веб-сайтів (www.comments.ua, www.gazeta.ua, www.feedsnews.ru), необхідно визначити активність інформаційної блогосфери щодо політичної обстановки та новітнього озброєння світу.

Використовуючи програмну реалізацію, почергово вирішуємо завдання. Після проведення аналізу інформації надається можливість завдяки програмній реалізації формування діаграми подібності з огляду на задачу (рис. 3).

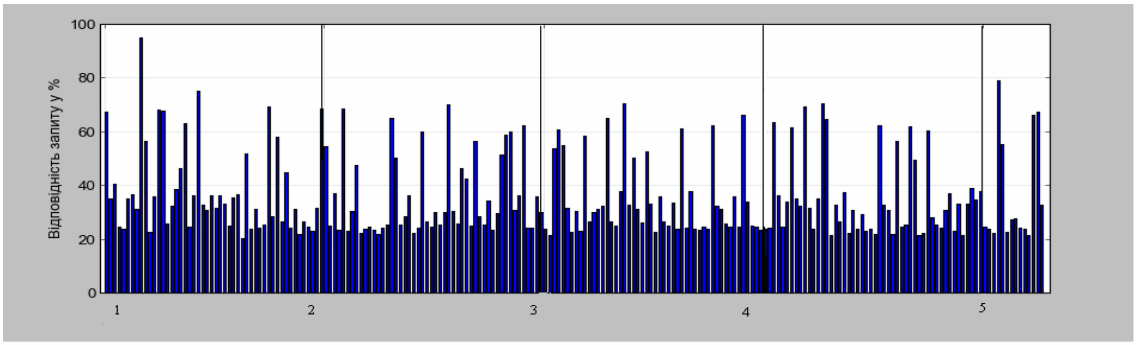

Рис. 8. Діаграма подібності

Цифрами позначено інформаційні контенти за пошуком інформаційних джерел Інтернету. Підсумком проведеного прикладу реалізації цільового завдання з використанням програми фільтрації інформаційних повідомлень було сформульовано типову задачу з використанням програмної реалізації.

Результатом виконання задачі є формування стовпчастої діаграми, що показує процентну відповідність публікації до заданого запиту та дає можливість побачити, яка з публікацій більш доцільна.

Висновок. Таким чином, у ході досліджень запропоновано методику фільтрації інформаційного повідомлення глобальної мережі, яка має актуальне спрямування, тому що існує велика кількість інформації, яка потребує структуризації. Відзнакою запропонованої методики є використання методу багатокритерійного порівняльного аналізу, який ґрунтується на зіставленні альтернатив. Запропонована методика доведена до практичного результату – розробленої програми фільтрації інформаційних повідомлень та їх класифікації. Практика її застосування довела дієвість запропонованих підходів.

СПИСОК ЛІТЕРАТУРИ

1. Основи інформаційних систем : навч. посібн. / В. Ф. Ситник, Т. А. Писаревська, Н. В. Єрьоміна,О. С. Краєва ; за ред. В. Ф. Ситника. – [2-ге вид перероб. і доп.]. – К. : КНЕУ, 2001. – 420 с.
2. Пермяков О. Ю. Інформаційні технології і сучасна збройна боротьба / О. Ю. Пермяков, А. І. Сбітнєв. – Луганськ : Знання, 2008. – 204 с.
3. Крысько В. Г. Секреты психологической войны (цели, задачи, методы, формы, опыт) / В. Г. Крысько, А. Е. Тараса. – Минск : Харвест, 1999. – 181 с.
4. Ларичев О. И. Теория и методы принятия решений, а также Хроника событий в Волшебных Странах : учебн. / О. И. Ларичев. – М. : Логос, 2000. – 296 с.: ил.
5. Батков Д. О. Информация: сбор, защита, анализ / Д. О. Батков, А. Г. Растомашкин. – М. : ООО Изд. Яуза, 2001. – 336 с.
6. Інтелектуальні системи підтримки прийняття рішень : навч. посібн. / Б. М. Герасимов, В. М. Локазюк, О. Г. Оксінюк, О. В. Поморова. – К. : Вид-во Європ. ун-ту, 2007. – 335 с.
7. Писарчук О. О. Методика багатокритеріальної ідентифікації технічних засобів та контрольованих ситуацій за сукупністю ознак / О. О. Писарчук // Збірник наукових праць ВІКНУ. – 2010. – № 26. – С. 90–96.
8. Ланде Д. О. Глубинный анализ текстов. Технология эффективного анализа текстовых данных [Электронный ресурс] / Д. О Ланде. – Режим доступа : http://visti.net/~dwl/art/dz/.

Подано 10.04.2014

А. А. Писарчук, Н. В. Стыров

МЕТОДИКА ФИЛЬТРАЦИИ ИНФОРМАЦИОННЫХ СООБЩЕНИЙ ИНТЕРНЕТ-ИСТОЧНИКОВ И ИХ КЛАССИФИКАЦИЯ

В статье сформирована методика фильтрации информационных сообщений и их классификация. На конкретном примере показаны возможности по ее использованию для реализации задач автоматизированной обработки информации открытых источников информационно-телекоммуникационной сети Интернет.

O. O. Pysarchuk, N. V. Styrov

METHOD OF FILTERING NEWS ALERTS INTERNET SOURCES AND CLASSIFICATION

The article generated method filtering broadcast messages and their classification. In the particular example shown possibilities of its use to achieve the objectives of automated data processing open-source information and telecommunications on the Internet.