УДК 027.081:316.334.3:004

Артур Федорчук,
наук. співроб. відділу організації та використання документального фонду Фонду президентів України НБУВ

ІНФОРМАЦІЙНО-АНАЛІТИЧНІ РЕСУРСИ ІЗ СОЦІАЛЬНО-ПОЛІТИЧНОЇ ПРОБЛЕМАТИКИ: ТЕХНОЛОГІЧНИЙ АСПЕКТ

Публікація присвячена питанням технології контент-моніторингу засобів масової інформації та створенню на її основі інформаційних ресурсів соціально-політичної проблематики. Розглянуто історію та процес упровадження контент-моніторингу.

Ключові слова: засоби масової інформації, контент-аналіз, контент-моніторинг, бібліографічна база даних, інформаційні ресурси.

Одним з основних факторів забезпечення національних інтересів є оперативне використання накопичених людством знань у найважливіших сферах діяльності суспільства. Формою безпосередньої участі знань у суспільному житті при цьому завжди є інформаційний ресурс (ІР). Стан розвитку національних ІР, засобів їх створення та використання визначає здебільшого потенційні можливості успішного розвитку держави, забезпечення її національних інтересів.

Створення ІР передбачає в загальному випадку аналітико-синтетичне перероблення матеріалів, яке поєднує процедури аналізу та синтезу інформації. Тобто для створення нового ІР необхідно знайти та розчленувати потрібну інформацію з подальшим синтезом нового інформаційного продукту. Засоби аналізу й синтезу можуть бути різноманітні та залежать від предметної області, інформаційних потреб потенційних користувачів ІР та джерел інформації.

Друковані та інтернет-видання преси породжують інформаційний ресурс величезного обсягу. Проте це лише потенційний ресурс, бо для його повноцінного використання в аналітичних дослідженнях необхідно провести ще значну інформаційно-аналітичну роботу, спрямовану на пошук, систематизацію, узагальнення інформації та приведення її до вигляду, зручного для використання під час вирішення конкретного завдання.

Шляхи формування і використання ІР досить різноманітні. Проте можна окреслити коло інформаційних та інформаційно-аналітичних

завдань, які при цьому найчастіше вирішуються, а саме: пошук потрібної інформації, її класифікація, інформаційне згортання текстів джерел, формування інформаційних, інформаційно-аналітичних оглядів, довідок тощо. Відповідно, автоматизація вирішення цих завдань передбачає наявність розвинутих інформаційних технологій і систем [1].

У 1992 р. відділом організації та використання документального фонду Фонду президентів України НБУВ була розроблена технологія контент-моніторингу соціально-політичних процесів, яка передбачала створення нових інформаційно-аналітичних ресурсів на базі аналізу змісту публікацій преси.

Автоматизована технологія створення ІР мала декілька істотних особливостей:

1. Формування банку ключових фрагментів публікацій є поєднанням двох автоматизованих процесів – аналітико-синтетичного перероблення та модифікованої процедури контент-аналізу текстів публікацій.

2. Використання як одиниці формування текстового інформаційного масиву ключового фрагмента публікації.

3. Індексація ключових фрагментів публікації за допомогою фасетної класифікації.

Для виділення з тексту інформації був використаний один із напрямів інформаційного згортання – фрагментування. Термін «ключовий фрагмент публікації» був уведений для визначення виділеного аналітиком фрагмента тексту, що відповідає проблемним напрямам конкретної інформаційної системи та, зберігаючи авторську форму, найточніше характеризує зміст публікації стосовно проблемних напрямів та тематичних рубрик класифікатора системи.

Контент-аналіз публікацій складається з ряду послідовних етапів, які органічно вписуються й доповнюють аналітико-синтетичне перероблення інформаційних матеріалів. Унікальність запропонованої технології полягає в поєднанні змістовних (якісних) та кількісних методів аналізу текстів публікацій. Послідовність етапів змістовного аналізу проблеми, що досліджується конкретною інформаційною системою, умовно поділимо на стадії:

– змістовний (якісний) аналіз сукупності публікацій;

– формалізований (кількісний) аналіз інформаційних масивів – індексного, бібліографічного та масиву текстів ключових фрагментів публікацій.

Перша стадія містить етапи:

– виявлення публікацій, відповідно до проблемних напрямів інформаційної системи. Етап пов'язаний з якісним осмисленням інформації й визначенням можливості її використання в конкретному дослідженні;

– виділення з тексту публікації фрагментів, що релевантні проблемі;

– визначення в межах виділеного фрагмента окремих елементів проблеми та порівняння їх з конкретними значеннями класифікатора;

– індексація відібраних фрагментів.

Завершальним етапом першої (змістовної) стадії контент-аналізу тексту публікації є визначення та присвоєння кожному виділеному фрагменту сукупності індексів, які, залежно від фасетної позиції індексної формули, відповідають конкретному значенню класифікатора даної фасети. Технологією було передбачено, що в кожній фасетній позиції може бути один або декілька рівнозначних індексів, кожен з яких висвітлює відповідний елемент проблеми. Допускався також варіант, коли фасетна позиція не містила індексу.

Таким чином, за допомогою фасетної формули здійснюється формальний опис виділеного фрагмента тексту. Сукупність формул усіх виділених фрагментів публікації є формальним описом документа в контексті досліджуваної проблеми. За умови ж повного перегляду та контент-аналітичного оброблення визначеного кола джерел сукупність індексних формул усіх відібраних за певний проміжок часу публікацій формально описує досліджувану проблему. У свою чергу, сукупність усіх відібраних ключових фрагментів публікацій є масивом релевантної проблемним напрямам системи інформації. Таким чином, перша стадія контент-аналізу одночасно була аналітичною складовою процесу аналітико-синтетичного перероблення інформації [4].

Виявлені фрагменти публікацій вводились операторами у БД. Інструментом для роботи з БД у 1992 р. була обрана програма Absmarc, призначена для зберігання бібліографічної інформації у форматі MARC. Був розроблений формат запису, що включав 17 полів та забезпечував зберігання повного бібліографічного опису публікацій, заіндексованих текстів, їх ключових фрагментів, приміток, відомостей про авторів, посилань на персони, що представлені в публікації, і посилань на інші публікації. Фасетна формула розміщувалася безпосередньо перед початком чергового фрагмента. При цьому кожний запис БД описував окрему публікацію та містив будь-яку кількість заіндексованих фрагментів цієї публікації.

Подальші етапи технології формування банку ключових фрагментів публікацій містять стадії автоматизованої обробки інформаційних масивів.

Технологія контент-моніторингу передбачає, що кожен запис БД розбивається на три незалежні складові, кожна з яких функціонує самостійно. Їх сукупність формує три відповідні інформаційні масиви:

– змістовна інформація – сукупність ключових фрагментів публікації, яку описує даний запис БД;

– структурна характеристика публікації – сукупність індексів, що містяться у фасетних формулах цитат;

– бібліографічний опис документа.

Кожна з трьох складових запису пов'язана з двома іншими за допомогою перехресних посилань. Таким чином, з'являється можливість працювати з кожним інформаційним масивом окремо та подавати його як повну публікаційну матрицю, що може аналізуватися окремо за кожною з виділених складових і в їх сукупності. Отримана матриця є інформаційним формалізованим описом досліджуваної проблеми.

Крім того, інформаційний масив індексів ключових фрагментів публікації можна розглядати як вбудований у структуру БД реляційний масив, що дає змогу виконувати всі завдання керування даними, характерними для реляційних БД.

На наступному етапі спеціалізована програма вилучає інформацію з кожного запису баз даних, розбиває на фрагменти, що відповідають конкретній фасетній формулі, з посиланням на бібліографію першоджерела. Потім формується індексний файл, в якому розшифровуються класифікаційні індекси фрагментів, які складають фасетну формулу (з додатковою перевіркою їх правильності), присвоюється їм індекс сортування з посиланням на місце розміщення самого фрагмента в базі даних. Індекс сортування містить код значення кожного тематичного фасета і код хронологічного фасета, присвоєний на основі бібліографічної дати публікації. Аналітик отримує можливість роботи з індексним файлом, маючи широкі можливості з відбору інформації, що його цікавить, різноманітного сортування за тематичними й хронологічними фасетами, а також (за необхідності) автоматизованої переіндексації фрагментів, згідно з проблемою, що ним досліджується. У випадку необхідності він також може переглянути або роздрукувати повний текст обраних ним фрагментів, об'єднаних у тематичні рубрики, і вилучити або переіндексувати деякі фрагменти, а також внести корективи до сортування.

Ключові фрагменти публікацій, які лаконічно передають закладену в публікації ідею, стало можливим об'єднувати в будь-який спосіб у межах параметрів, представлених у фасетних формулах у вигляді індексів елементів проблеми. Технологією було передбачено, що домінантним стосовно інших може бути будь-який фасет. Залежно від цього акцент робився на тому чи іншому аспекті проблеми, яку можна було представити як у комплексі, тобто з урахуванням усіх її параметрів, так і частково, відокремлюючи окремі її аспекти. Отже, в інформаційній технології було закладено можливості багатоваріантного й широкоаспектного формування вихідної інформації.

Автоматизована технологія формування банку ключових фрагментів публікацій із різноманітних (попередньо визначених) джерел уможливлює отримання інформаційних продуктів з характерними особливостями:

1. Глибока структурованість огляду, яка досягається багаторівневим сортуванням інформації. Технологія дає змогу проводити незалежне сортування індексного та бібліографічного масиву, кількість рівнів сортування визначається кількістю фасетних позицій і полів БД, виділених для бібліографічного опису публікації. Теоретично обидві ці складові не обмежені, тобто глибину структурованості оглядів також теоретично не обмежено.

2. Об'єднання цитат у рубрики й підрубрики, назви яких збігаються зі значеннями індексів та/або елементів бібліографії публікації.

3. Посилання кожної цитати на першоджерело. Після кожного фрагмента публікації передбачено розміщення повного або скороченого бібліографічного опису.

Апробована на широкомасштабних дослідженнях матеріалів газетної періодики під час виборчих кампаній (як президентських, так і парламентських) інформаційна система, завдяки програмному інструментарію багатоаспектного використання інформаційних масивів, поступово переросла в більш глобальну систему контент-моніторингу соціально-політичних проблем та процесів, яка протягом багатьох років давала можливість відслідковувати імідж політичної еліти та політичних партій України, а також розвиток соціально-політичних процесів та конкретних подій в Україні [2, 3]. Для аналітико-синтетичного перероблення, автоматичного аналізу та формування інформаційно-аналітичних матеріалів використовувалися власні програмні засоби, які динамічно розвивалися та вдосконалювалися. Наприклад, був зроблений перехід від двофасетного до багатофасетного класифікатора, а трьохетапний процес аналітико-синтетичного перероблення матеріалів був втілений в одному програмному модулі.

Розроблена система опрацювання інформаційних масивів містила ряд інформаційних файлів, які, у свою чергу, містили відомості про фасети, значення їх індексів, варіанти порядку сортування інформаційних модулів і файл конфігуратора. Останній був визначальним щодо всіх інших у плані конкретного завдання синтезу інформаційних оглядів та/або формування кількісних результатів контент-аналітичного дослідження. Конфігуратор формалізував завдання, які ставить перед системою користувач-аналітик, і містив назви інформаційних файлів, які використовуються для індексування та сортування інформаційних модулів, назви БД, призначених для оброблення, назви вихідних файлів тощо [5].

Концептуальна модель багатофасетного класифікатора розробляється на основі аналізу проблемної галузі з урахуванням потреб замовника.

Для інформаційних систем соціально-політичного спрямування на базі періодичних видань використовується класифікатор, який містить п'ять фасетних позицій:

1. Об'єктний фасет (так званий проблемний).
2. Суб'єктний фасет (чия думка висвітлювалась у фрагменті публікації).
3. Тематичний фасет (питання, що висвітлювались у публікації).
4. Фасет модальності публікації.
5. Фасет першоджерел (містить назви центральних видань і регіон для місцевої преси).

Поліпшення комп'ютерного та програмного забезпечення, створення локальної мережі та поява нових завдань зі створення електронних інформаційних ресурсів, а також забезпечення доступу до них читачів зумовили необхідність адаптації технології контент-моніторингу до нових загальнобібліотечних програмних засобів та систем автоматизації бібліотеки. Нова версія системи автоматизації бібліотек – «ІРБІС64» – почала впроваджуватись у НБУВ з 2009 р., що створило перспективи автоматизації практично всіх галузей бібліотечно-інформаційної діяльності. На основі стандартних рішень, що пропонує «ІРБІС64», було розроблено спеціалізовану базу даних «Інформаційно-аналітичний огляд преси», адаптовану до потреб відділу, та забезпечено її експлуатацію в мережевому режимі.

Для кожного користувача налагоджено персональний профіль, в якому розділено доступ до інформаційних ресурсів, які він має опрацьовувати, та зареєстровано прізвище співробітника, за яким автоматично ведеться статистика його роботи із записами бази даних.

Для опрацьованих матеріалів налагоджено вихідну форму у форматі RTF, яка виводить аналітичний матеріал в упорядкованому ієрархічному вигляді відповідно до введених тематичних та географічних рубрик і може бути представлена замовнику за будь-який період часу.

Було вирішене також питання конвертації БД, що свого часу були створені у форматі MARC за допомогою програми Absmarc та використовувались як інформаційні масиви в технології контент-моніторингу відділу.

Поєднання можливостей інтегрованого середовища САБ «ІРБІС64» та технології контент-моніторингу дало змогу зробити процес досліджень більш технологічним і зручним для аналітиків та отримати унікальний інформаційно-аналітичний продукт широкого спектра використання та представлення. Сьогодні користувачеві (читачеві) можуть бути надані як структуровані відповідно до запиту тематичні інформаційно-аналітичні

матеріали за будь-який проміжок часу, аналітичні та статистичні результати досліджень, так і доступ до БД у середовищі САБ «ІРБІС64» з можливостями різноманітного пошуку та відбору потрібної інформації.

Список використаної літератури

1. *Горовий В.* Бібліотеки як сучасні центри української інформатизації / В. Горовий // Наук. пр. Нац. б-ки України ім. В. І. Вернадського. – К. : НБУВ, 2009. – Вип. 25. – С. 23–35.

2. *Танатар Н. В.* Сучасні інформаційні технології прес-моніторингу передвиборних кампаній / Н. В. Танатар, А. Г. Федорчук // Українська періодика: історія і сучасність : VI Всеукр. наук.-теорет. конф., 11–13 трав. 2000 р. : доп. та повідомл. – Л., 2000. – С. 342–344.

3. *Танатар Н. В.* Усе – про політичну еліту / Н. В. Танатар, А. Г. Федорчук // Вісн. НАН України. – 1997. – № 1–2. – С. 91–92.

4. *Федорчук А. Г.* Використання інтернет-ресурсів для контент-моніторингу передвиборних кампаній / А. Г. Федорчук, Н. В. Танатар // Наук. пр. Нац. б-ки України ім. В. І. Вернадського. – К. : НБУВ, 2008. – Вип. 21.– С. 227–236.

5. *Федорчук А. Г.* Теоретико-методичні засади аналізу інформаційного потоку соціально-політичного спрямування / А. Г. Федорчук, Н. В. Танатар // Бібліотекознавство, документознавство, інформологія. – 2004. – № 2. – С. 33–38.