

**Оперативність і передрук в мережевих джерелах новин**

**Анотація.** Розглянуто явище дублювання інформації в мережевих джерелах новин. Досліджуються правові засади та статистичні дані на масивах новинних повідомлень, що узагальнюються системою контент-моніторингу. Результати можуть враховуватися розробниками і менеджерами пошукових систем і систем контент-моніторингу.

**Ключові слова:** дублювання інформації, інформаційні агентства, авторське право, передрук, контент-моніторинг, автоматичний аналіз.

**Аннотация.** Рассмотрено явление дублирования информации в сетевых источниках новостей. Исследуются правовые основы и статистические данные на массивах новостных сообщений, обобщаемых системой контент-мониторинга. Результаты могут учитываться разработчиками и менеджерами поисковых систем и систем контент-мониторинга.

**Ключевые слова:** дублирование информации, информационные агентства, авторское право, перепечатка, контент-мониторинг, автоматический анализ.

**Summary.** We study the phenomenon of information duplication in online news sources. Legal bases and statistical data are investigated. Approbation on news flows is carried out by means of system of content monitoring. The results can be considered by developers and managers of the search engines and content-monitoring systems.

**Keywords:** duplication of information, news agencies, copyright, reprint, content-monitoring, automatic analysis.

Традиційно джерелами новин вважаються інформаційні агентства (ІА). Дійсно, це їх основна діяльність, цим вони займаються професійно. Сьогодні робота ІА перемістилася в Інтернет, крім того утворюються джерела інформації, представлені виключно в інтернет-просторі – мережеві інформагентства [1]. Соціальні мережі вносять свій внесок у формування новинного інформаційного простору, причому, якщо про рівень достовірності інформації з соціальних мереж можна дискутувати, то оперативність при цьому завжди висока. Досвід авторів при вирішенні задач моніторингу свідчить про те, що новини все частіше потрапляють в соціальні мережі з веб-простору.

Цінність інформаційних повідомлень багато у чому визначається оперативністю, тому окремим завданням є оцінка запізнювання публікацій в Інтернет в порівнянні з часом розсилки відповідних повідомлень. Забігаючи наперед, скажемо, що в більшості розглянутих випадків час затримки виявилось негативним, тобто ІА копіювали повідомлення з веб-сайтів, та ще й зі значним запізненням.

Одним з ключових аспектів розвитку сучасних інформаційних технологій є специфіка взаємовідносин між інформаційними агентствами, що традиційно грають роль постачальників інформації, і ЗМІ, які були її основним споживачем у агентств. Ці взаємини значною мірою застаріли і потребують серйозних коректив як у технологічному плані, так і в плані організаційному, включаючи законодавче

регулювання. Головна причина такого стану справ полягає у швидкому розширенні впливу на інформаційні процеси мережевих технологій, у першу чергу Інтернету. Розвиток цих технологій призвів до якісних змін у структурі всього процесу інформування громадськості на всіх його ланках, в результаті чого ситуація вимагає вже кардинального перегляду основних механізмів, що лежать в основі функціонування медійних засобів.

Інформаційні агентства забезпечують своїх передплатників інформацією на умовах, які на сьогоднішній день виглядають щонайменше дивно. Зокрема, типовим умовою щодо використання матеріалів ІА є заборона на розмноження і поширення їх будь-якими способами.

Таким чином агентства намагаються захистити свою продукцію від копіювання, часто посилаючись на законодавство про авторські права. Разом з тим у статті 10 Закону України «Про авторське право і суміжні права» передбачено, що повідомлення про новини або поточні події не охороняються авторським правом. Аналогічно у статті 8 Закону РФ «Про авторське право і суміжні права» йдеться про те, що «повідомлення про події та факти, що мають інформаційний характер» не охороняються авторським правом. Таким чином, умови, декларовані більшістю ІА з посиланням на законодавство про авторські права, являються неправомірними, принаймні, по відношенню до їх основної продукції – інформаційних повідомлень.

Не краща ситуація і з змістовним аспектом проблеми. Ніхто, безумовно, не ставить під сумнів авторські права на ті матеріали, які дійсно мають автора в звичайному сенсі слова (інтерв'ю, аналітичні розробки, ексклюзивні репортажі і т. д.). Але говорити про авторські права на повідомлення про офіційний візит глави держави або набуття чинності нового закону явно не має сенсу, не кажучи вже про тексти законів, указів і т. п., для яких законодавчо передбачений порядок обов'язкового оприлюднення.

Як завжди в подібних ситуаціях, нові тенденції починають прокладати собі дорогу, не чекаючи офіційних рішень, що неминуче призводить до перерозподілу не тільки ресурсів, але і функціональних ролей учасників комунікації. Тому для вироблення обґрунтованих рекомендацій бажано було б спочатку розібратися в тому, що і як відбувається насправді.

Отже, інформ агентства генеруються новини на підставі: 1 – матеріалів власних кореспондентів з місць подій, явищ; 2 – інтерв'ю; 3 – реакції на прес-релізи компаній, кампаній; 4 – переказів, шляхом переказів інформації партнерів. Третє і четверте можна назвати генерацією новин вже умовно, – прес-релізи можуть публікуватися ще до надходження в інформ агентства, наприклад, на веб-сайтах їхніх генераторів, а переклади та перекази – це скоріше доведення відомих новин до своєї аудиторії. До діяльності, що не має прямого відношення до генерації новин, можна віднести також аналітичні узагальнення і передруки (в тому числі і коректні).

Ідеальна схема генерації новин в інформ агентствах як реакції на подію і подальшого передруку цих новин наведена на рис. 1.

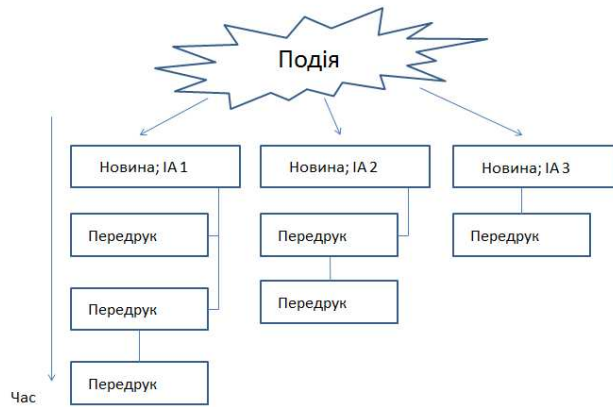


Рис. 1. Ідеальна схема генерації та передруку новин

На практиці перші повідомлення про подію в інтернет-просторі може публікувати зовсім не ІА, а, наприклад, користувач соціальної мережі або випадковий веб-сайт. У деяких випадках, інформантство, просто передруковує такі повідомлення (рис. 2).

Власне, хронологія життя новини в інтернет-середовищі, і може бути найкращим індикатором дублювання і передруків.

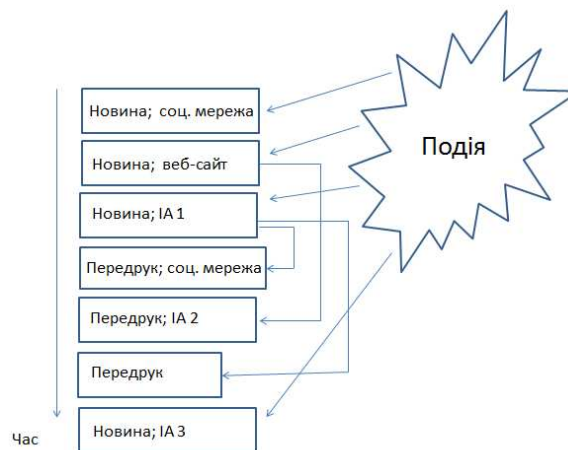


Рис. 2. Наближена до реальності схема хронології «життя» новини

Для вимірювання хронології появи новин в інтернет-просторі, їх генерації на тлі подібних за змістом повідомлень, автори використовували ресурс системи контент-моніторингу інтернет-новин InfoStream [2]. Ця система на даний час в оперативному режимі сканує близько 6000 веб-сайтів, представлених в українському і російському сегментах веб-простору, період сканування джерел програмою-ботом становить від 15 хвилин до 12 годин, в залежності від режиму поновлення відповідних джерел. Обсяг повідомлень, що обробляються, перевершує 100 тис. повідомлень на добу. В ході дослідження розглядалися два текстових корпуси – повні тексти повідомлень, сканованих з веб-простору, і масив «словесних сигнатур», що відповідають цим текстам [3, 4].

З погляду технологій виявилось, що методи визначення нечітких дублікатів повідомлень, розвинені в останні роки як вітчизняними, так і зарубіжними дослідниками [5-7], виявилися дуже цікавими в розглянутому застосуванні.

В рамках технології системи InfoStream застосовується статистичний сигнатурний спосіб визначення дублікатів і схожих документів, який базується на кількості співпадаючих «опорних» слів у порівнюваних документах. В якості сигнатур для окремих повідомлень використовувалися ланцюжки з 12 опорних слів, що пройшли процедуру морфологічної нормалізації. Така невелика кількість термів в ланцюжку, який визначається своєрідною словесною сигнатурою, пояснюється невеликою середньою довжиною новинних повідомлень. При цьому для виявлення співпадаючих і схожих документів, наведених іншими мовами (чітких і нечітких дублікатів), використовуються також переклади опорних слів на українську та англійську мови [3, 8].

На основі можливостей системи InfoStream було створено сервіс, який дозволяє знаходити для кожного новинного повідомлення вибраного джерела всі дублікати, що охоплюються системою InfoStream. Цей сервіс цікавий для менеджерів джерел, що скануються, серед яких найбільшу зацікавленість проявляють представники інформаційних агентств. При цьому генерується два основних статистичних зрізу для джерел інформації: – це, по-перше, перелік повідомлень інформаційного агентства, доповнений переліком дублікатів (чітких або нечітких), публікованих на інших ресурсах (рис. 3), а, по-друге, перелік джерел, що публікують дублікати повідомлень вихідного джерела, який супроводжується переліком цих повідомлень. Крім того, для кожного аналізованого джерела визначається середній час випередження публікації його повідомлень у порівнянні з публікаціями дублікатів на інших веб-ресурсах.

Не всі джерела генерують новинні повідомлення, багато з них передрукують вже опубліковані в інтернет-просторі. Деякі інформаційні агентства спочатку на платній основі розсилають свої матеріали електронною поштою замовникам, і лише через декілька годин після цього публікують їх на власних веб-сайтах. У цьому випадку середній час випередження виявляється негативним, тобто відбувається запізнювання.

Для виявлення загальних тенденцій, пов'язаних з оперативністю і передруком в мережеских джерелах новин, зокрема, в мережеских ресурсах інформаційних агентств, авторами були відібрані 160 інформаційних агентств, щодня публікують новинні повідомлення російською та українською мовами. Для кожного з інформаційних джерел було розраховано час випередження публікації його матеріалів у порівнянні з публікацією на інших ресурсах інформаційних дублікатів, як чітких (критерій – збіг 6-и опорних слів з 12-и), так і нечітких (4-х з 12-и).

Виявилось, що матеріали лише близько 30 джерел в середньому випереджають час появи дублікатів (як чітких, так і нечітких) на інших веб-ресурсах (рис. 4). Діапазон від 16 годин випередження до 66 годин запізнювання. Середнє запізнювання повідомлення інфоагентства становить близько 12 годин.

На закінчення можна зробити декілька висновків.

Проведені дослідження показали, що, всупереч загальній думці, інформаційні агентства в мережі далеко не завжди є першоджерелами. Виявляється, багато солідних інформаційних агентств дуже часто самі передрукують матеріали з інших веб-сайтів, соціальних мереж. Сьогодні це говорить практично про кінець

епохи таких агентств, тобто об'єктивно повинні змінюватися форми їх роботи, підходи до формування та публікації матеріалів.

Найчастіше джерела, взагалі не генерують власної інформації, а лише займаються оперативним моніторингом та інтеграцією, мають позитивний час випередження публікації своїх матеріалів, що, взагалі є зручним для користувачів, які отримують інформацію оперативно і «з одного вікна», тобто вони в певному сенсі більш ефективні, ніж першоджерела.

УКРІНФОРМ: 2013.02.10			
<b>1</b>	<b>(23)</b>	<b>Негода знеструмила 30 населених пунктів на Київщині</b>	<b>2013.02.10 10:03</b> <b>УКРІНФОРМ</b>
1		Негода знеструмила 30 населених пунктів на Київщині	2013.02.10 10:02 Всесвітня служба УТР
2		Негода знеструмила 30 населених пунктів на Київщині	2013.02.10 10:03 <b>УКРІНФОРМ</b>
3		Негода знеструмила 30 населених пунктів на Київщині	2013.02.10 10:05 Перший Національний
4		Непогода обесточила 30 населенных пунктов в Киевской области	2013.02.10 10:15 Левый берег
5		У Київській обл. внаслідок негоди знеструмлено 30 населених пунктів	2013.02.10 10:24 РБК-Украина
6		В Киевской обл. в результате непогоды обесточены 30 населенных пунктов	2013.02.10 10:24 РБК-Украина
7		В Киевской обл. в результате непогоды обесточены 30 населенных пунктов	2013.02.10 10:29 Трибуна
8		Ненастье обесточило 30 населенных пунктов на Киевщине	2013.02.10 10:30 УКРИНФОРМ
9		В Киевской области из-за непогоды обесточены 30 населенных пунктов	2013.02.10 10:44 Коммерсантъ-Украина
10		Непогода оставила без света 30 населенных пунктов	2013.02.10 10:49 News24UA
11		Негода залишила Київщину без електрики	2013.02.10 11:01 УНІАН

...

<b>13</b>	<b>(5)</b>	<b>Українські тенісистки достроково програли іспанкам у матчі Кубка Федерації</b>	<b>2013.02.10 13:33</b> <b>УКРІНФОРМ</b>
1		Українські тенісистки достроково програли іспанкам у матчі Кубка Федерації	2013.02.10 13:32 Всесвітня служба УТР
2		Українські тенісистки достроково програли іспанкам у матчі Кубка Федерації	2013.02.10 13:33 <b>УКРІНФОРМ</b>
3		Українські тенісистки достроково програли іспанкам у матчі Кубка Федерації	2013.02.10 14:05 Перший Національний
4		Україна зазнала краху від іспанок у Кубку Федерації	2013.02.11 11:29 ТСН.ua
5		Україна зазнала краху від іспанок у Кубку Федерації	2013.02.11 13:04 ВолиньІнфо
<b>14</b>	<b>(5)</b>	<b>Американський шок-рокер Мерілін Менсон знепритомнів на сцені</b>	<b>2013.02.10 01:18</b> <b>УКРІНФОРМ</b>
1		Мерілін Менсон втратив свідомість на сцені	2013.02.08 13:32 Українські національні новини
2		У Канаді Мерліна Менсона знудило і "вирубало" на сцені	2013.02.08 17:42 ICTV Факти
3		Американський шок-рокер Мерілін Менсон знепритомнів на сцені	2013.02.10 01:18 <b>УКРІНФОРМ</b>
4		Американський шок-рокер Мерілін Менсон знепритомнів на сцені	2013.02.10 01:32 Всесвітня служба УТР
5		Американський шок-рокер Мерілін Менсон знепритомнів на сцені	2013.02.10 02:05 Перший Національний
<b>15</b>	<b>(5)</b>	<b>Українські хокеїсти завершили олімпійську кваліфікацію поразкою від словенців</b>	<b>2013.02.10 21:33</b> <b>УКРІНФОРМ</b>
1		Хокей. Україна почала кваліфікацію на Олімпіаду з поразки від датчан	2013.02.08 01:01 Корреспондент.net
2		Хокей. Україна почала кваліфікацію на Олімпіаду з поразки від датчан	2013.02.08 01:16 NewsBox.com.ua
3		Українські хокеїсти завершили олімпійську кваліфікацію поразкою від словенців	2013.02.10 21:33 <b>УКРІНФОРМ</b>
4		Українські хокеїсти завершили олімпійську кваліфікацію поразкою від словенців	2013.02.10 22:02 Всесвітня служба УТР
5		Українські хокеїсти завершили олімпійську кваліфікацію поразкою від словенців	2013.02.10 22:05 Перший Національний

Рис. 3. Перелік повідомлень джерела інформації, кожне з яких доповнено переліком дублікатів

Результати заставляють задуматися, за чим передплатники звертаються до інформаційних агентств сьогодні, коли більша частина інформації з мінімальною затримкою (і навіть випередженням) доступна на інших веб-сайтах, а повноту можуть забезпечити системи контент-моніторингу? Можливо, за аналітичною добіркою цієї інформації, репрезентативною і достовірною. Тобто, інформаційне агентство, якщо воно бажає вижити у сучасних умовах, має приділяти підвищену увагу саме аналітичній обробці інформації, перетворюючись на агентство інформаційно-аналітичне.

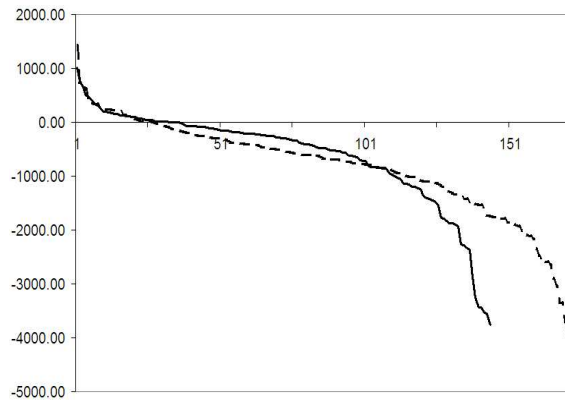


Рис. 4. Середній час випередження/відставання публікації матеріалів (у хвилинах) на сайтах інформаційних агентств

Результати, отримані за допомогою запропонованих методів, допомагають менеджерам інформаційних агентств при організації функціонування своїх веб-ресурсів. Крім того, такі вимірювання можуть надавати додаткові дані для рейтингування джерел інформації. Отримані дані можуть також враховуватися розробниками пошукових систем і систем контент-моніторингу, що вже сьогодні займають провідні місця на ринку новин поряд з великими інформаційними агентствами.

### Використана література

1. Ландэ Д.В., Брайчевский С.М., Дармохвал А.Т., Морозов А.Ю. Веб-пространство и материалы информационных агентств // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог" (Бекасово, 4-8 июня 2008 г.). Вып. 7 (14). – М.: Изд-во РГГУ, 2008. – С. 303-305.
2. Григорьев А.Н., Ландэ Д.В. и др. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: ООО «Старт-98», 2007. – 40 с.
3. Ландэ Д.В., Жигало В.В. Технология полнотекстового поиска в мультязычных сетевых ресурсах // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2012. - Казань: Изд-во "Фэн" Академии наук РТ, 2012. – С. 101-105
4. Д.В. Ландэ, А.Т. Дармохвал, А.Ю. Морозов. Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006. – С. 115-119.
5. Ю.Г. Зеленков, И.В. Сегалович Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль, Россия, 2007. – Том 1, С. 166-174.
6. Никконен А.Ю. Устранение избыточности и дублирования сюжетов новостных сообщений // Интернет-Математика. Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. Ун-та, 2007. – С. 157-167.
7. J. Bourdaillet. Alignment of Noisy Unstructured Text Data // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data. Hyderabad, India - January 8, 2007. P. 139-146.
8. Ландэ Д.В., Жигало В.В. Підхід до рішення проблеми пошуку різномовного плагіату // Сб.наукових праць "Проблеми інформатизації та управління". – Київ: НАУ, 2008. – Вип. 2(24). – С. 125-129.