ВЕБ-ПРОСТРАНСТВО И МАТЕРИАЛЫ ИНФОРМАЦИОННЫХ АГЕНТСТВ

WEB-SPACE AND MATERIALS OF NEWS AGENCIES

Ландэ Д.В. (dwl@visti.net), Брайчевский С.М. (smb@visti.net), Дармохвал А.Т. (hval@visti.net), Морозов А.Ю. (alex@visti.net)

(Информационный центр «ЭЛВИСТИ», Киев, Украина)

Исследуется в какой мере материалы, доступные подписчикам информационных агентств за плату, публикуются в открытом доступе на информационных веб-сайтах. Получено распределение сообщений информационных агентств по времени запаздывания, определено как удельное количество перепечаток материалов информационных агентств на веб-сайтах, так и сообщений из Интернет, включенных в состав лент информационных агентств.

Одним из ключевых аспектов развития современных информационных технологий является специфика взаимоотношений между информационными агентствами (ИА), традиционно играющими роль поставщиков информации, и СМИ, являющимися основным ее потребителем. На взгляд авторов предлагаемой статьи, эти взаимоотношения в значительной мере устарели и нуждаются в серьезных коррективах как в технологическом плане, так и в плане организационном, включая законодательное регулирование. Главная причина такого положения дел состоит в быстром расширении влияния на информационные процессы сетевых технологий и, разумеется, в первую очередь Интернет. Развитие этих технологий привело к качественным изменениям в структуре всего процесса информирования общественности на всех его звеньях, в результате чего ситуация требует кардинального пересмотра основных механизмов, лежащих в основе функционирования медийных средств.

Информационные агентства снабжают своих подписчиков информацией на условиях, которые на сегодняшний день выглядят по меньшей мере странно. В частности, типичным условием относительно использования материалов ИА является запрет на размножение и распространение их любыми способами. Таким образом агентства пытаются защитить свою продукцию от копирования, зачастую ссылаясь на законодательство об авторских правах. Вместе с тем в статье 8 Закона РФ «Об авторском праве и смежных правах» говорится о том, что «сообщения о событиях и фактах, имеющие информационный характер» не охраняются авторским правом. В аналогичном Законе Украины в статье 10 также предусмотрено, что сообщения о новостях или текущих событиях не охраняются авторским правом. Таким образом, условия, декларируемые большинством ИА со ссылкой на законодательство об авторских правах, являются неправомерными, по крайней мере, по отношению к их основной продукции – информационным сообщениям.

Не лучше обстоит дело и с содержательным аспектом проблемы. Никто, безусловно, не ставит под сомнение авторские права на те материалы, которые действительно имеют автора в обычном смысле слова (интервью, аналитические разработки, эксклюзивные репортажи и т. д.). Но говорить об авторских правах на сообщения об официальном визите главы государства или вступлении в силу нового закона явно лишено конкретного смысла. Мы уже не говорим о текстах законов, указов и т. п., для которых законодательно предусмотрен порядок обнародования.

Как всегда в подобных ситуациях, новые тенденции начинают прокладывать себе дорогу, не дожидаясь официальных решений, что неизбежно приводит к перераспределению не только ресурсов, но и функциональных ролей участников коммуникации. Поэтому для выработки обоснованных рекомендаций желательно было бы вначале разобраться в том, что и как происходит в действительности.

Целью данной работы было исследование того, в какой мере материалы, доступные платным подписчикам основных ИА, становятся доступными в открытом доступе на информационных веб-сайтах. Ценность информационных сообщений во многом определяется оперативностью, поэтому отдельной задачей была оценка запаздывания публикаций в Интернет по сравнению с временем рассылки соответствующих сообщений. Забегая вперед, скажем, что почти в третьей части рассматриваемых случаев время задержки оказалось отрицательным, т.е. ИА копировали сообщения с веб-сайтов, да еще и со значительным запаздыванием.

При проведении исследований авторы получили уникальную возможность доступа к подписным материалам ведущих ИА, представленных в украинском информационном пространстве. Кроме того, в распоряжении авторов находилась система контент-мониторинга InfoStream [1] – поисковая система, с помощью которой в реальном масштабе времени сканируется свыше 3000 информационных веб-сайтов, представленных в украинском и российском сегментах веб-пространства. Таким образом, в ходе исследования рассматривались два текстовых корпуса (точнее, набора «словесных сигнатур» текстов [2], представленных в этих корпусах) – сообщений ИА и текстов, сканированных из веб-пространства. Рассматривались сообщения ИА по общеполитической тематике, поступающие в течение 5-25 ноября 2007 года. Их объем оказался равным 8955

документов. Эти сообщения сравнивались с текстами, сканируемыми из Интернет в течение всего ноября 2007 года, количество которых составило свыше 1 млн. документов.

Технически задача нахождения дубликатов (в данном случае речь идет именно о дубликатах, а не о сообщениях по той же теме, но другими словами, т. е. учитывались перепечатки с незначительными искажениями) решалась методом, который описан в [2]. Этот метод относится к группе методов нахождения «подобных» документов [3-5], основанных на выделении некоторого множества опорных слов, имеющих наибольший TFIDF [2, 3]. В качестве некоторых «инвариантов» для отдельных сообщений использовались цепочки из 12 опорных слов, прошедших процедуру морфологической обработки (стемминга). Такое небольшое количество термов в цепочке, которая является своеобразной словесной сигнатурой, объясняется небольшой средней длиной новостных сообщений (2000-3000 символов).

В результате проведенных исследований удалось получить такие данные:

- из 8955 сообщений ИА на веб-сайтах было опубликовано 5567 сообщений (62 %);
- общее количество перепечаток на различных веб-сайтах составило 39901 (456 %). Соответствующее распределение, оказавшееся гиперболическим, приведено на рис. 1;
- количество перепечаток с положительным временем запаздывания (из материалов ИА - на веб-сайты) составило 28933 (73 %);
- количество перепечаток с отрицательным временем запаздывания (перепечаток из Интернет, помещаемых в ленты ИА) составило 10968 (27 %).

Ранжированный график распределения сообщений ИА по времени задержки публикаций приведен на рис. 2, на котором четко видны экстремальные отклонения в начальной и конечной области. Отклонение в начальной области характеризует большое время задержки включения в ленты ИА материалов, размещенных, как правило, на сайтах органов государственной власти (инертность ИА, отсутствие у них средств мониторинга веб-пространства). Отклонения в конечной области объясняются задержками перепечаток на веб-сайтах сообщений, получивших со временем некоторое новое продолжение. Вместе с тем центральная область графика (от 1000-го по 5000-е сообщение) имеет стабильный характер со средним значением около получаса.

Массовый характер перепечаток позволяет делать выводы о том, что все сообщения, интересные администраторам соответствующих веб-сайтов, были перепечатаны. По-видимому, примерно 38 % сообщений ИА оказались им недостаточно интересными.

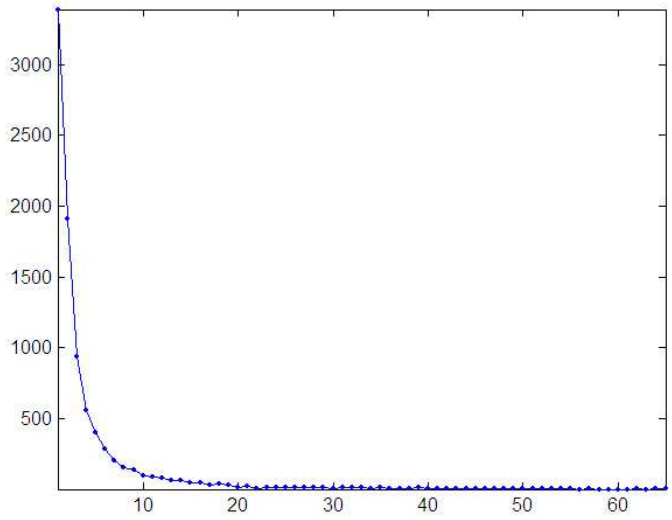

Рис. 1. Количество сообщений ИА (ось ординат), ранжированное по количеству перепечаток на веб-сайтах (ось абсцисс). Значению 0 соответствует количество неперепечатанных сообщений

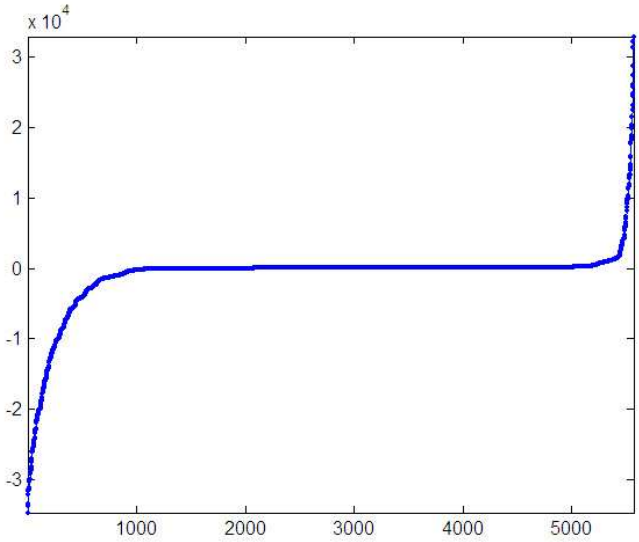

Рис. 2. Распределение сообщений ИА (ось абсцисс) по времени запаздывания в минутах (ось ординат)

В заключение можно сделать несколько выводов. С точки зрения технологий оказалось, что методы определения нечетких дубликатов сообщений, развитые в последние годы как отечественными, так и зарубежными исследователями, оказались очень интересными в рассматриваемом применении. Результаты заставляют задуматься, за что же платят подписчики информационным агентствам сегодня, когда большая часть информации с минимальной задержкой доступна в Интернет, а полноту могут обеспечить системы контент-мониторинга? По-видимому, за аналитический подбор этой информации, репрезентативность и достоверность. То есть, информационное агентство, если оно желает выжить в современных условиях, должно уделять повышенное внимание именно аналитической обработке информации, превращаясь в агентство информационно-аналитическое.

Естественно, полученные результаты могут учитываться также разработчиками информационно-поисковых систем и систем контент-мониторинга. Усиление аналитической составляющей таких систем уже сегодня позволяет им выступать на рынке рядом с крупнейшими информационными агентствами.

Список литературы

1. 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. (http://dwl.visti.net/art/booklet/booklet.pdf)
2. 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. (http://www.rcdl2006.uniyar.ac.ru/papers/paper_71_v2.pdf)
3. 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. (http://rcdl2007.pereslavl.ru/papers/paper_65_v1.pdf)
4. 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 (http://download.yandex.ru/IMAT2007/imat2007.pdf)
5. 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 (http://research.ihost.com/and2007/cd/ Proceedings_files/p139.pdf)