ВЫЯВЛЕНИЕ НОВЫХ СОБЫТИЙ ИЗ ПОТОКА НОВОСТЕЙ
DETECTION OF NEW EVENTS FROM NEWS FLOW

***Ландэ Д.В.**, (dwl@visti.net), **Брайчевский С.М.**, (smb@visti.net), **Григорьев А.Н.**, (gri@visti.net),*
***Дармохвал А.Т.**, (hval@visti.net), **Радецкий А.Б.**, (rad@visti.net)*
Информационный центр «ЭЛВИСТИ», Киев

Доклад посвящен популярной в настоящее время во всем мире тематике выявления, отслеживания и группировки новых событий из потоков новостей (New Event Detection, Tracking, Clustering). Приведен краткий обзор теоретических и практических разработок в этом направлении. Представлен оригинальный многокритериальный алгоритм выявления новых событий из потока новостей. В качестве основного метода настройки параметров алгоритма используется ретроспективный анализ и технология формирования сюжетных цепочек, созданная в рамках системы контент-мониторинга InfoStream®.

Характеристики информационных потоков зачастую определяются потоками событий реального мира [1]. Если событие новое и важное, то обязательно о нем будут много говорить и в дальнейшем, т.е. задача выявления новых событий из потока новостей является в каком-то смысле задачей предсказания дальнейшего появления множества подобных сообщений, задачей прогноза, актуальной как с научной, так и практической точки зрения. Кроме того, эта задача непосредственно связана с общей задачей нахождения исключений или аномалий, т.е. сообщений, которые в данный момент по каким-то критериям, например, лингвостатистическим, выделяются из общего потока, хотя в дальнейшем могут породить множество себе подобных [2-3]. Исследуемая проблема состоит в создании такого алгоритма выявления новых событий и/или параметров таких алгоритмов, которые на основе ретроспективного подхода к изучению потоков новостей позволят получить наилучшее соответствие между публикациями о новых событиях и сюжетами, которые будут выявлены в последующем.

Задачи выявления, отслеживания и группировки событий на основе анализа новостей активно обсуждаются специалистами во всем мире в течение продолжительного времени. Как выяснилось, они имеют большое практическое значение именно сегодня, когда режим доступа к системам интеграции новостей существенно облегчен. Данная проблема достаточно поднималась с 70-х годов XX-го века. В начале рассматривалась как "Topic Detection", а позднее как "New Event Detection" [4]. Первые работы связаны с Солтоном, векторно-пространственной моделью представления данных и традиционными методами кластеризации. Значительный вклад внес Р. Папка [5], который в своих работах и диссертации рассматривал записи о новых событиях как фрагментах документов, не удовлетворяющих запросам пользователей, построенных с учетом уже известных событий.

Подход Р. Папка кажется достаточно перспективным, однако ограничен массивом запросов пользователей. Основой же предлагаемого авторами подхода является автоматическое определение «близости» документов, их массивов и словарей. Релевантность, вычисляемая в подходе Папка, заменяется нами функциями подобия, ставшими уже традиционными в практике информационного поиска.

Общепринятая технологическая схема решения данной задачи, которой придерживались и авторы, может быть представлена следующим образом (рис. 1). Как правило, задача выявления новых событий из потока сообщений предполагает, что на вход соответствующего программно-технологического комплекса последовательно поступают новые документы (на рис. 1. - поток новых документов). В общем случае они могут поступать как политематический поток, но, как правило, как отобранные по тематическому запросу сообщения от информационного роутера (информационно-поисковой системы). Далее в соответствии с определенными алгоритмами, происходит непосредственное выявление новых событий. Новые события описываются в документах, для которых с помощью отдельных программных модулей во временной ретроспективе формируются цепочки подобных документов (сюжетные цепочки). Документы, отражающие различные новые события могут быть основой новых групп взаимосвязанных документов – кластеров (группировка событий). В свою очередь, каждый из этих кластеров со временем может стать основой формирования полноценной сюжетной цепочки.

Предлагаемый авторами подход базируется на таких предположениях, относящихся к документам, содержащим информацию о новых событиях:

а) минимальное время, прошедшее с момента публикации документа;

б) близость лексического состава документа к лексическому составу массива документов за небольшой промежуток времени (массив оперативных новостей);

в) существенное различие лексического состава документа от лексического состава массива документов за значительный период времени – окна наблюдения;

г) наличие в документе терминов, входящих в плюс-словарь (включающий важные для содержания новостей слова типа «теракт», «конфликт», «сенсация» и т.п.);

д) высокий ранг «авторитетности» источника, а также допустимости лексики заглавий новостей (определяемых экспертами).

е) отсутствие дублирования информации [3].

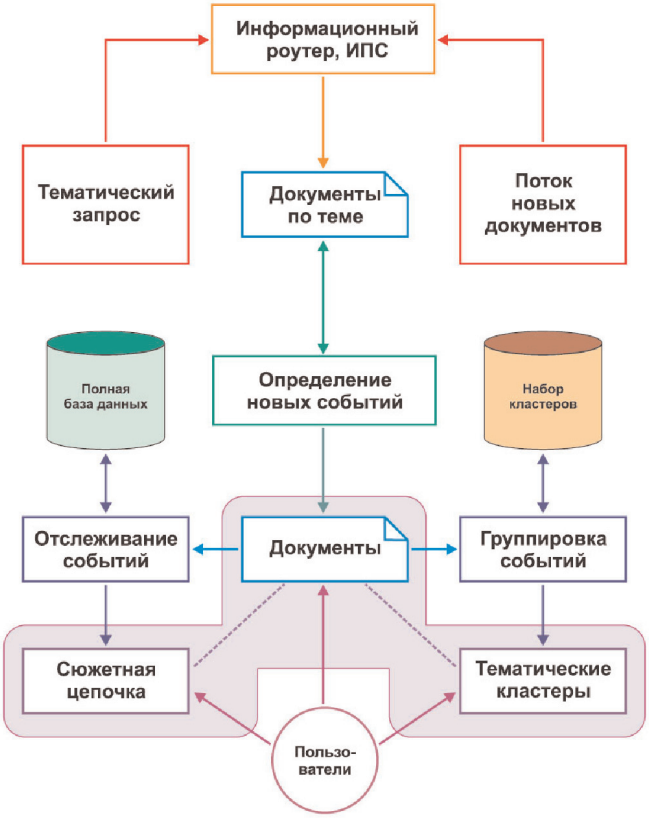

Рис. 1. Определение новых событий – элемент технологии контент-мониторинга

Введем обозначения: N – величина окна наблюдения потока новостей; n – величина массива оперативных новостей ($n < N$); D_i – i-й документ; $PlusDic$ – плюс-словарь; $sim(D_i, D_j)$ – мера близости документа i документу j; $sim(D_i, PlusDic)$ – мера близости документа i плюс-словарю; $Rang_i$ – ранг источника, соответствующего i-му документу.

В этих обозначениях мера близости лексического состава документа от лексического состава массива массива оперативных новостей вычисляется следующим образом:

$$\sum_{j=1}^{n} sim(D_i, D_j),$$

Соответственно мера близости лексического состава документа от лексического состава массива остальных документов из окна наблюдения вычисляется следующим образом:

$$\sum_{j=n}^{N} sim(D_i, D_j).$$

Было использовано приведенное в [6] определение меры близости документов, использующее аппарат условных вероятностей, а именно, как вероятность того, что некоторое слово w входит в документ D_i при условии, что оно входит в документ D_j:

$sim(D_i, D_j) = \text{Prob}(w \in D_i \mid w \in D_j)$.

Предлагаемая авторами формула для расчета ранга документа как «носителя» информации о новых событиях, учитывающая условия а) - е) может быть записана следующим образом:

$$\frac{Rang_i * sim(D_i, PlusDic) * \sum_{j=1}^{n} sim(D_i, D_j)}{\log(i+1) * \sum_{j=n}^{N} sim(D_i, D_j)},$$

с учетом введенных выше обозначений, а также того, что. Ввиду того, что нумерация документов в потоке проводится в обратном порядке, значение $\log(i+1)$ в знаменателе отражает вклад времени, прошедшего с момента публикации события.

На основе приведенной формулы может происходить ранжирование документов, поступающих в систему интеграции новостей. Построенный в рамках представленной технологии алгоритм используется в настоящее время в системе контент-мониторинга InfoStream, на вход которой ежедневно поступает свыше 40 тыс. документов. Данный алгоритм реализует прогнозно-аналитическую модель, основная методология оценки достоверности которой в настоящее время заключается в экспертном сравнении выявленных новых событий с основными сюжетами, полученными через некоторый интервал времени. Для настройки алгоритма экспертами использовались такие «рычаги», как параметры *N, n*, плюс-словарь, массив рангов источников информации, массив исключений для заглавий и адресов.

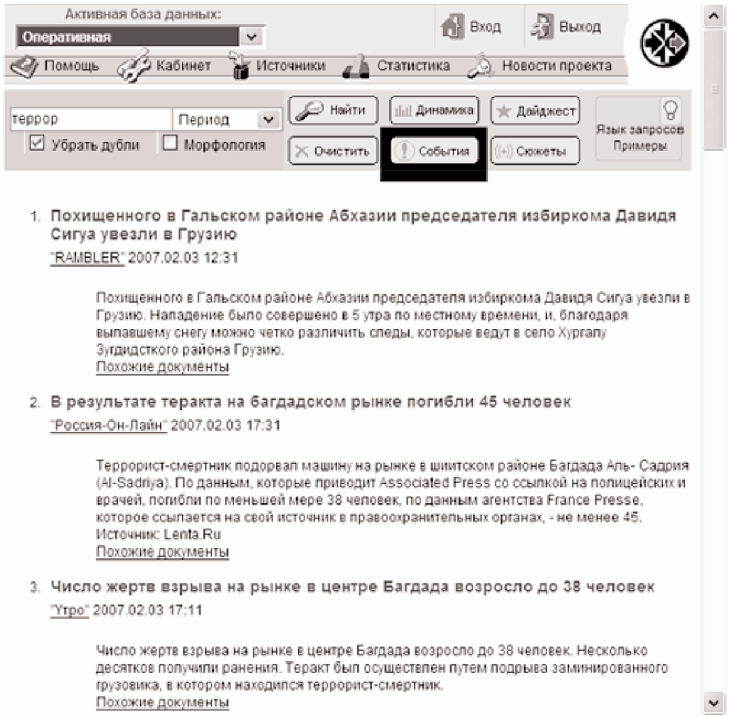

Рис. 2. Новая функция системы системы контент-мониторинга InfoStream

В настоящее время во многих популярных системах интеграции новостей задача выявления новых событий заменяется выявлением основных новостных сюжетных цепочек. Такой подход, конечно, частично решает названную задачу, однако, предоставляя пользователям ответ на вопрос «о чем больше всего пишут в последнее время», фактически отличается целевой функцией.

Авторами было проведено ретроспективное исследование с целью оценки, насколько сегодняшние события, определяемые в соответствии с предложенным подходом, станут основой сюжетов следующего дня. Оказалось, что таких событий не более 20%. Чаще всего большая часть сюжетов следующего дня повторяет сюжеты дня предыдущего. Приходится признавать, что не все новые события одинаковы по важности и порождают в дальнейшем значительные кластеры подобных документов.

Предложенный нами подход, конечно же, нельзя считать окончательным решением поставленной задачи. Например, не всегда изменение размеров окон наблюдения и объемов оперативных массивов может привес-

ти к адекватному выявлению новых событий, которые имеют свою предысторию. Рассматриваемый в статье плюс-словарь требует постоянного сопровождения, а в некоторых случаях «персонализации».

Однако, полученные практические результаты показали свою эффективность как существенное дополнение к поисковым режимам. При этом самое важное, пожалуй то, что пользователь привязывается не к новым сообщениям, а к новым событиям реального мира.

Список литературы

1. Ландэ Д.В. Основы интеграции информационных потоков - К.: Инжиниринг, 2006. - 240 с.
2. Ландэ Д. В., Фурашев В. Н. Выявление новых событий в рамках системы контент-мониторинга. / Научно-техническая информация. Сер. 2. –М., 2006. - №12, - С. 12-16.
3. Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках. Труды Восьмой Всероссийской научной конференции (RDCL'2006). - С. 115-119.
4. Kurt, H. On-line New Event Detection and Tracking in A Multi-Resource Environment, MS. Thesis, Bilkent University, 2001.
5. Papka, R. On-line News Event Detection, Clustering, and Tracking. Ph. D. Thesis, University of Massachusetts at Amherst, September 1999.
6. Kumaran, G., Allan, J. and McCallum, A., Classification Models for New Event Detection, CIIR Technical Report. – 2004.