Ранжирование источников информации в системе мониторинга новостей InfoStream

© Д.В. Ландэ, С.М. Брайчевский, А.Т. Дармохвал, А.Ю. Морозов

Информационный центр «ЭЛВИСТИ», Киев, Украина
dwl@visti.net smb@visti.net hval@visti.net alex@visti.net

Аннотация

Описано несколько подходов к ранжированию и отбору информационных источников при построении служб мониторинга новостей. Эти подходы базируются на принципе обеспечения максимальной полноты информации при минимальном количестве источников, выборе наиболее оригинальных, тематически стабильных, максимально цитируемых источников. Представлены результаты соответствующих статистических исследований. Рассматриваемые в работе принципы используются при отборе источников для корпоративных применений на базе системы мониторинга новостей InfoStream.

1 Введение

Эффективное использование возможностей современного информационного пространства невозможно без общих представлений о природе соответствующих информационных процессов, устойчивых закономерностях.

Системы интеграции и мониторинга новостей из открытых веб-сайтов сети Интернет сегодня все чаще становятся основными компонентами информационных служб различного уровня. Ни для кого уже не секрет, что даже самая закрытая новостная информация, передаваемая информационными агентствами, с минимальной временной задержкой становится доступной в Сети. Можно отметить разнообразный диапазон параметров информационных источников как по объемам публикуемой информации, так и по содержанию – от сообщений информационных агентств – до «живых журналов».

Вместе с тем мощные возможности Интернет порождает проблему оптимизации состава и количества источников, используемых корпоративной информационной системой с целью обеспечения приемлемого качества, удовлетворяющего потребностям пользователей.

В этой связи актуальными оказываются вопросы ранжирования и выбора источников новостной информации – веб-сайтов, к которым требуется обеспечить доступ через один интерфейс как в поисковом режиме, так и в режимах аналитического обобщения.

Принципам ранжирования как отдельных веб-документов, так и документальных массивов посвящено большое количество научных работ и практических разработок [1–5]. Ссылочное ранжирование веб-сайтов сегодня является отдельным направлением интернет-бизеса – SEO (search engine optimization) [6]. Вместе с тем, вопросам ранжирования и отбора информационных ресурсов с учетом их новостного контента, объемов и стабильности тематики публикаций уделяется значительно меньшее внимание [7, 8].

Безусловно, основным критерием при отборе источников для таких систем мониторинга новостей является их содержание. Как было показано в [8], распределение источников по контенту, соответствующему тематическим потребностям корпоративного пользователя удовлетворяет закону Бредфорда [9], соответственно, при отборе источников обязательно должно учитываться их ранжирование по степени соответствия тематике. Однако, реализация такого отбора приводит к известным сложностям. На практике такое ранжирование осуществляется экспертами [5] путем оценивая количества документов, релевантных некоторому отлаженному пакету тематических запросов, адресуемых к фрагменту базы данных, составленной из документов анализируемого источника. А это неизбежно приводит к элементу субъективизма со всеми вытекающими последствиями.

Поэтому нам представляется перспективным дополнить традиционный подход более объективными и строгими методами, позволяющими оптимизировать процесс формирования информационной базы систем интеграции контента.

Именно такой подход как основной, дополненный некоторыми представленными ниже, применяется в службе ИЦ «ЭЛВИСТИ» при отборе источников для корпоративных применений системы интеграции и мониторинга новостей InfoStream [10].

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

2 Распределение источников по количеству генерируемых документов

В качестве экспериментального корпуса в наших исследованиях использовался информационный массив, охватываемый системой интеграции и мониторинга новостей InfoStream. В частности, ниже приведены распределения, относящиеся к массиву документов за март 2008 года объемом свыше 1.2 млн. документов из более чем 2500 источников – открытых веб-сайтов.

На рис. 1 приведен график распределения (в полулогарифмическом масштабе) количества документов, опубликованных источниками, ранжированными по этому параметру. Центральная часть графика хорошо аппроксимируется прямой, что свидетельствует о близости представленной зависимости к гиперболической (т.е. о действии обобщенного закона Ципфа). На рис. 2 приведено общее количество документов, охватываемых системой мониторинга в зависимости от учитываемых в ней источников, также ранжированных по количеству опубликованных документов. Поскольку закон Ципфа предполагает аппроксимацию плотности распределения гиперболической зависимостью вида a/x, то функция распределения количества документов:

$$f(x) \sim \int \frac{a}{x} dx = a \ln x + C$$

в разумном приближении описывается логарифмическим законом.

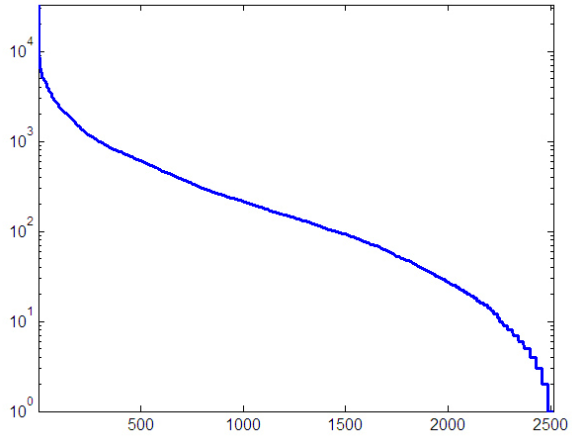

Рис. 1. Ранжированный список источников (ось 0X) по количеству публикаций (ось 0Y)

Приведенная зависимость позволила построить критерий отбора необходимой части источников для различных корпоративных применений из общего списка охватываемых системой мониторинга InfoStream, удовлетворительно решающих задачи пользователей.

Если предположить, что все источники давали бы одинаковый вклад по количеству опубликованных документов, то рассматриваемая зависимость была бы линейной и выражалась бы формулой:

$$f_{lin}(n) = n \frac{f_{\max}}{N},$$

где $f_{\max}$ – максимальный объем охватываемых документов, N – количество источников.

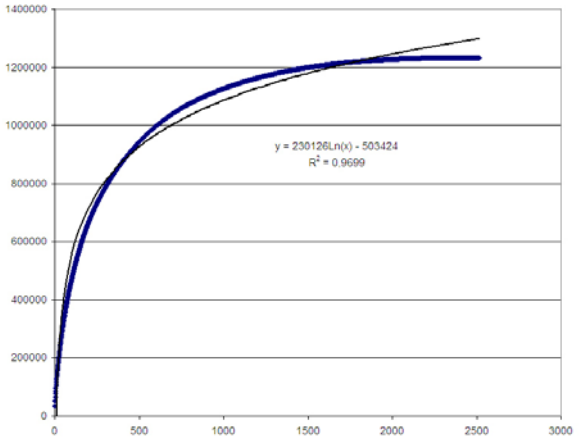

Рис. 2. Количество публикаций в системе мониторинга (ось 0Y) в зависимости от источников (ось 0X), ранжированных по количеству документов

Очевидно, что отклонение реальной зависимости от линейной сначала возрастает, а затем уменьшается до нуля. Будем называть количество источников пороговым n_p, когда значение реальной зависимости $f(n)$ максимально отклоняется от приведенной линейной:

$$n_p = \arg\max \left\{ f(n) - f_{lin}(n) \right\}.$$

На рис. 3 приведена иллюстрация значений n_p для различных значений N, т.е. когда выбирается N наиболее продуктивных источников. Что интересно (и вполне соответствует характеру функции $f(n)$), значения n_p практически линейно зависят от N (рис. 4): $n_p \sim 0.24N$, при этом количество охватываемых документов, соответствующих n_p при максимальном количестве источников (2514, рис. 5) достигает 80 процентов от $f_{\max}$.

При этом можно заметить, что построенная зависимость удовлетворяет принципу Парето: приблизительно 20% наиболее продуктивных источников публикуют 80% документов.

3 Наиболее цитируемые источники

Как уже было отмечено, цитируемость отдельных документов и веб-сайтов сегодня является одним из основных критериев оценки рангов документов в сетевых поисковых системах (PageRank, HITS, Salsa, TrustRank, h-индекс и др.) Идея оценки уровня цитируемости позволила построить одну из первых моделей веб-пространства [11]. Главное ее достоинство состоит в том, что она естественным образом сочетает в себе содержательный аспект с воз-

можностью использования количественных параметров, значения которых определяются вполне объективно.

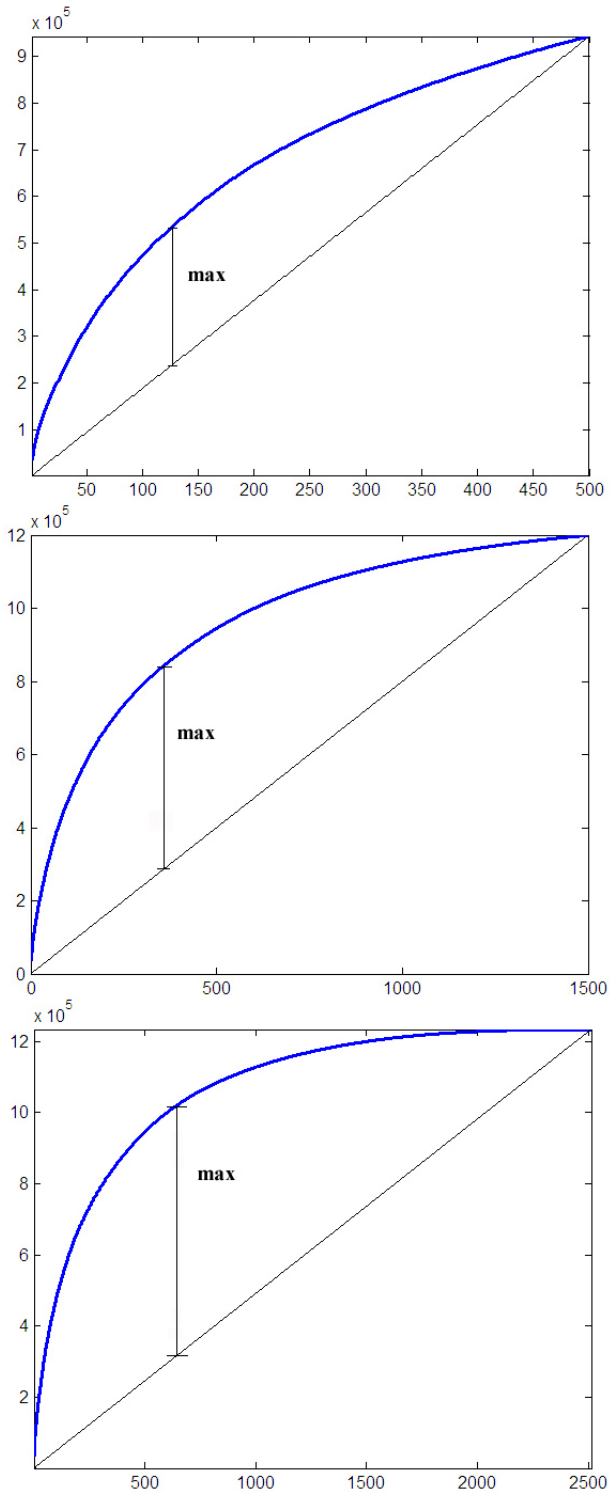

Рис. 3. Количество публикаций в системе мониторинга при подключении новых наиболее интенсивных источников (500, 1500, 2500)

Однако при изучении цитируемости новостных веб-ресурсов как информационных источников необходимо учитывать ряд условий, актуальных для таких источников.

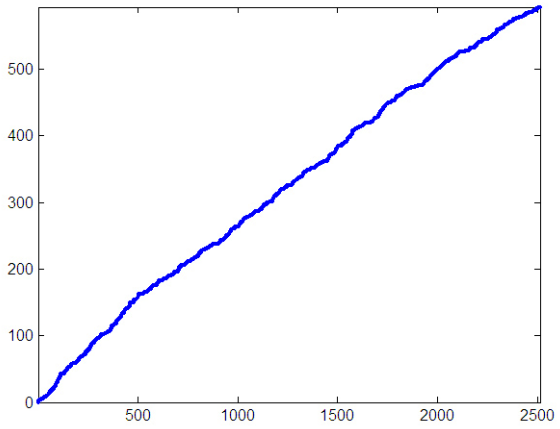

Рис. 4. Изменение порогового значения (ось 0Y) при изменении исходного количества источников (ось 0X)

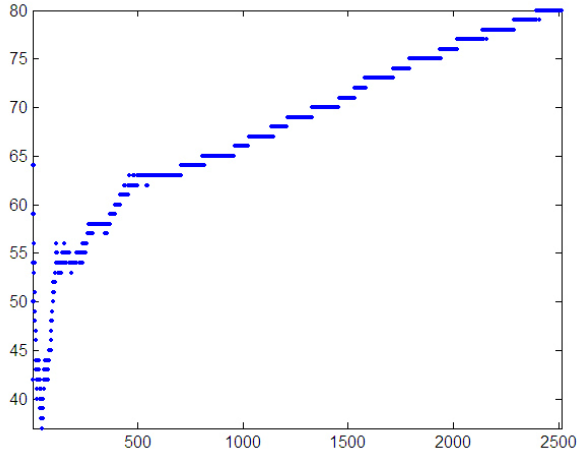

Рис. 5. Удельное количество документов, охватываемых системой (ось 0Y) при изменении исходного количества источников (ось 0X)

Можно было бы попросту применить модель А. Бредера [11] и критерий типа PageRank [1] к новостной составляющей Web-пространства, однако такой подход нельзя считать корректным по ряду причин:

- новостные потоки характеризуются повышенной динамикой [12], что сильно влияет на природу гиперссылок. Например, на наиболее актуальные сообщения в течение определенного времени ссылок может вообще не существовать;
- модель Бредера слабо учитывает особенности «скрытого» Web, т.е. тех информационных Web-ресурсов, на которые не существует прямых гиперссылок (в свое время им рассматривались лишь ресурсы, уже охваченные поисковой системой AltaVista);
- в новостных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части Web-пространства (это могут быть зачастую ссылки на ресурсы, доступные только по паро-

лю, или даже оффлайновые публикации изданий, возможно и присутствующих в Интернет);
- кроме того, модель Бредера не включает такого понятия, как содержательное дублирование информации.

Для построения модели для каждого из 2500 активных источников за март 2008 года был составлен запрос в следующем виде:
<код источника>#<шаблон для поиска> [#<шаблон для поиска>…#<шаблон для поиска>].

Совокупность подобных запросов была объединена в пакет, в результате специальной обработки которого для каждого сообщения, относящегося к определенному источнику – веб-сайту, были выявлены исходящие ссылки на другие источники (ссылки на собственный источник исключались). Было выявлено, что исходящие контекстные ссылки присутствовали на 484945 сообщениях с 2323 веб-сайтов.

Также было получено распределение новостных источников по количеству веб-сайтов, имеющих на них ссылки. Всего за месяц ссылки указывали на 1459 источников (без самоцитирования). Оказалось, что на первые 100 источников ведут ссылки с более 80% веб-сайтов. На рис. 6. представлен график ранжированного распределения новостных источников веб-сайтов по количеству сайтов, имеющих на них ссылки. Следует обратить внимание на то, что приведенный график позволяет достаточно четко выделить две зоны, имеющие различные статистические характеристики (углы наклона на графике): первая зона (левая на рис. 6), включающая информационные агентства и крупнейшие издания и вторая, которая соответствует сайтам госорганов, специальных изданий, компаний, публикующим пресс-релизы.

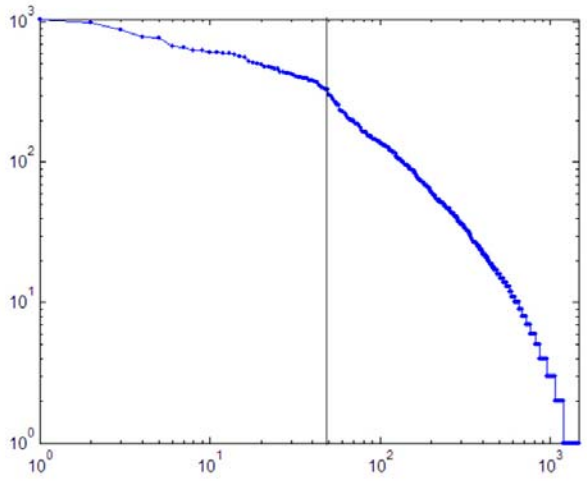

Рис. 6. Зависимость количества ссылающихся веб-сайтов (ось 0Y) от ранга новостного источника (ось 0X) в логарифмической шкале

Ниже приведен начальный фрагмент ранжированного списка новостных источников, на которые ведут ссылки с максимального количества веб-сайтов:

Web-сайт	Количество ссылающихся веб-сайтов
ИА «Интерфакс»	1051
«РосБизнесКонсалтинг»	983
"Reuters"	882
ИТАР-ТАСС	787
РИА «Новости»	773
УНИАН	675
Радио «Свобода»	662
НТВ	631
«Коммерсантъ»	623
BBC	598
«Комсомольская правда»	595

Следует отметить, что оценка уровня источника информации как «автора» преимущественно по количеству веб-сайтов, с которых на него ведут гиперссылки, вполне согласуется с предложенным Мораном и Лемпелем алгоритмом Salsa [13].

4 Выбор наиболее оригинальных источников

Специальное место в исследовании занимало изучение смыслового дублирования информации [14]. При этом следует отметить, что процент дублирующихся по смыслу документов в системе мониторинга InfoStream значительно меньше, чем во всем новостном Web-пространстве. Это объясняется подбором источников для сканирования, в число которых не входят многие новостные интеграторы.

Как уже отмечалось нами ранее, одной из главных особенностей новостной информации является наличие большого количества сообщений, дублирующих друг друга. Так, о событии мирового значения напишут все СМИ, причем, скорее всего, на одной из первых страниц. Потребитель же (за исключением некоторых специфических направлений аналитических исследований информационного пространства) желает получать по каждому событию одно сообщение. Поэтому исследование характера и свойств дублирования информации приобретает в современных технологиях исключительно важное значение. В том числе, крайне актуальной становится задача отбора наиболее оригинальных источников, позволяющих (по крайней мере статистически) исключить не только формальное, но и содержательное дублирование информации.

Технология InfoStream, позволяет с высокой достоверностью выявлять содержательные дубликаты сообщений [10]. Дублирование сообщений на веб-сайтах зависит от различных причин, поэтому проведенные измерения для ранжированного по количеству публикаций списка источников показывают разный уровень, при этом информация не носит наглядного характера. Вместе с тем, сглаживание с

помощью метода скользящей средней (с окном наблюдения, равным 20), позволил получить график (рис. 7), наглядно свидетельствующий об устойчивой тенденции: чем более продуктивен источник информации, тем больше он содержит заимствований из других источников.

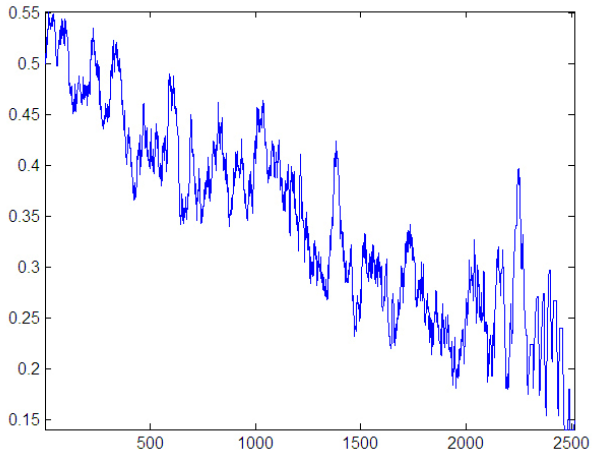

Рис. 7. Усредненное удельное количество дублирующихся документов (ось 0Y) по ранжированному по количеству публикаций списку источников (ось 0X)

5 Тематическая стабильность

Одной из важных характеристик информационных источников в новостном сегменте Сети является их стабильность, понимаемая как генерация постоянного числа документов в единицу времени (естественно, с учетом периодичности изданий) [15]. Примером стабильных источников могут служить крупные информационные агентства, регулярно поставляющие потребителям примерно одинаковые объемы информации на протяжении длительного времени, а примером нестабильных – блоги, многие из которых активно действуют в течение нескольких дней, а затем угасают.

Естественно, источник, регулярно выпускающий свою продукцию, с большей вероятностью отразит в своих публикациях важные события, чем источник, выходящий нерегулярно, от случая к случаю (он может попросту «проскочить мимо события»).

С другой стороны, крупные издания, обеспечивающие полноценное освещение нашей жизни, как правило на первое место выводят масштабные, значительные в общественном понимании события, о которых мы все равно так или иначе узнаем, если не из телевизора, так из разговоров в метро. События же меньшего, так сказать, общественного веса, но при этом, возможно, интересные и важные для отдельных групп потребителей, либо вообще отсутствуют, либо теряются «на последних страницах». Поэтому задача оптимального учета стабильности источников отнюдь не тривиальна и требует, на наш взгляд, серьезных исследований.

Ниже мы обратимся к одному из важных ее аспектов.

Как было показано, ежедневное общее количество документов, публикуемых на основных информационных веб-сайтах приблизительно постоянно, и колеблется в основном в зависимости от дня недели [8]. Вместе с тем тематика публикаций подвержена существенным колебаниям.

Один из возможных подходов к решению проблемы ранжирования источников информации основывается на подходе, заключающемся в изучении динамики порождаемых ими тематических информационных потоков.

На практике среди множества проблем подбора и анализа источников контента большое значение, в частности, имеет учет параметров их тематической стабильности. При этом тематическая стабильность и стабильность публикации информации источниками зачастую играют решающую роль при проведении аналитических исследований. Например, такие важные свойства информационных источников, как тематическую корреляцию и полноту, имеет смысл учитывать только для источников, публикующих документы относительно стабильной тематической направленности.

Тематическую стабильность источника можно определить как корреляцию наборов тематических рубрик, которым соответствуют документы из этого источника в различные периоды времени. Нам представляется, что конкретный набор рубрик мало влияет на предлагаемый ниже метод расчета стабильности источников (под тематической рубрикой в данном случае понимается тематика, семантика которой, в частности, находит свое отражение в виде запроса на информационно-поисковом языке). Предполагается, что документу присваивается та или иная рубрика, если он соответствует определенному запросу. Перечень рубрик и соответствующих им запросов был выбран авторами на основании опыта работы с политематическими новостными ресурсами сети Интернет. Эти рубрики и запросы установлены и апробированы в течение длительного времени в системе контент-мониторинга InfoStream. В настоящее время система включает 35 основных тематических рубрик.

При исследовании тематической направленности некоторых источников информации были обнаружены документы, отклоняющиеся от основной направленности этих источников. Такие документы, если их количество относительно невелико, не должны влиять на рассчитываемый ниже уровень стабильности источников. Конечно, автоматическая рубрикация во многом зависит от качества запросов, однако некоторыми погрешностями в рубрикации при статистическом исследовании можно пренебречь.

Для вычисления уровня стабильности источника информации использовалась формула, основанная на так называемом R/S-анализе [16]. Следует отметить, что этот подход имеет непосредственное отношение к фрактальному анализу, применение которого, однако, выходит за рамки настоящей ра-

боты. R/S -анализ позволяет исследовать «изрезанность» кривой, образуемой временным рядом на основе отношения разброса значений к среднеквадратичному отклонению.

Авторами был предложен параметр тематической стабильности K временного ряда интенсивности публикаций на веб-сайтах (источниках), который выглядит следующим образом:

$$K = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i}{R_i},$$

где N – количество тем (рубрик) источника;

S_i – среднеквадратичное отклонение по рубрике i;

R_i – размах значений по рубрике i.

Значение S_i вычисляется по формуле:

$$S_i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\{r_{ij} - \frac{1}{M}\sum_{k=1}^{M} r_{ik}\right\}^2},$$

где r_{ij} – количество вхождения рубрики i за день j, M – количество значений ряда измерения (недель, например).

Значение R_i вычисляется следующим образом:

$$R_i = \max_{1<k<M} X_{ik} - \min_{1<k<M} X_{ik},$$

где X_{ik} – накопленное к моменту k отклонение по рубрике i, вычисляемое по формуле:

$$X_{ik} = \sum_{j=1}^{k}(r_{ij} - \frac{1}{M}\sum_{l=1}^{M} r_{il}).$$

На рис. 8. представлена кривая значений коэффициентов стабильности для источников (было измерено поведение свыше 2500 источников за 2007 год), ранжированных по этим значениям.

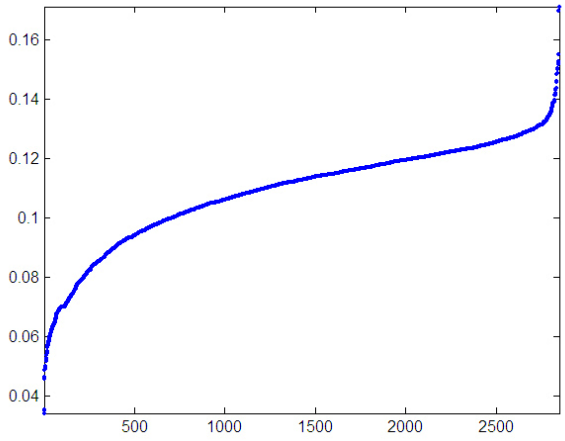

Рис. 8. Ранжированный список источников (ось 0X) по параметру тематической стабильности (ось 0Y)

В частности, самыми тематически стабильными документами (значения правой верхней части диаграммы), оказались периодические профессиональные издания, такие как «Континент Сибирь», «Зеркало недели», «Русский Вестник», «Политический журнал», «Власть денег» и т.п., с определенной периодичностью печатающие постоянное количество сообщений по тематикам, распределенным в приблизительно в одинаковых пропорциях. Подтвердилась гипотеза о том, что именно профессионализм информационного источника коррелирует с тематической стабильностью. В частности, практически все ведущие информационные агентства, выпускающие политематическую информацию, тем не менее, вошли в состав наиболее тематически стабильных.

Кроме приведенной тематической, исследовалась более простая, диаграмма внетематического распределения источников, ранжированная по коэффициентам стабильности. Полученные данные еще раз подтвердили тот факт, что электронные издания более склонны изменять тематику публикаций, чем свои объемы, выраженные общим количеством публикаций.

6 Заключение

Сегодня становится ясно, что разработка качественно новых средств работы с сетевыми ресурсами переходит в разряд приоритетных задач. В частности, без развитых средств наблюдения за сетевыми информационными источниками невозможно обеспечить соответствующих репрезентативность выборок, а эта задача сегодня является одной из самых актуальных при отборе источников для корпоративных применений систем контент-мониторинга.

Отметим лишь несколько практических применений ранжирования информационных источников. Во-первых, это даст возможность выявления первоисточников информации, например, для размещения в них рекламных материалов, материалов информационного влияния и т.п. Во-вторых, можно сократить затраты времени и средств путем игнорирования, исключения из поиска и анализа заведомо слабых, «шумовых» источников. Кроме того, для оперативного нахождения актуальной информации корректное ранжирование может способствовать нахождению действительно полезных первоисточников и служб интеграции информации.

Результаты данных исследований источников информации могут использоваться при ранжировании выдачи информационно-поисковых систем, подсчете медиа-рейтингов, позволяют рекомендовать пользователям наиболее тематически стабильные и оригинальные источники информации, например, для включения их в список «персональных» в интерфейсах систем контент-мониторинга информационных ресурсов.

Следует отметить, что, несмотря на то, что в данной работе приведено четыре критерия ранжирования источников информации, окончательный «универсальный» критерий не приводится. Теоретически его можно было бы записать, например, как линейную комбинацию приведенных критериев с некоторыми экспертно определяемыми коэффициентами. Однако практика, диктуемая информационными потребностями корпоративных пользователей, показывает, что при выборе источников информации останавливаются на одном из приведен-

ных критериев, дополняя его некоторыми неформальными соображениями (вкусовые предпочтения, учет региональных факторов и т.п.).

В заключение выражаем благодарность А. Григорьеву за организационную поддержку данного исследования, С. Бороденкову и Р. Мазуркевичу и А. Стукаленко за участие в обсуждении затронутых проблем и кропотливую работу с контентом информационных источников.

Литература

[1] Brin S. and Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7, - 1998.
[2] Kleinberg J.M. Authoritative sources in a hyperlink environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.
[3] Kleinberg J.M. Authoritative sources in a hyperlinked environment. Journal of the ACM,
[4] 46(5):604–632, 1999.
[5] Luo G., Tang Ch., Yu P.S. Resource-adaptive real-time new event detection // In Processing of SIGMOD'07, June 11-14, 2007, Beijing, China.
[6] Kent, P. Search engine optimization for dummies . Hoboken, Wiley, 2004, 354 p.
[7] Gianna M. Del Corso, Antonio Gullн, Francesco Romani. Ranking a stream of news. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. – pp. 97–106.
[8] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1, 2005. – № 11. – С. 21–33.
[9] Bradford S.C. Sources of Information on Specific Subjects // Engineering: An Illustrated Weekly Journal (London), 137, 1934 (26 January), pp. 85–86.
[10] http://infostream.ua
[11] Broder A., Kumar R., Maghoul F., Prabhakar R., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the web. URL: http://www.almaden.ibm.com/cs/k53/www9.final/
[12] Ландэ Д.В. Структура новостного Web-пространства // Научно-техническая информация. Сер. 2. – М., 2006. – № 8. – С. 17–20.
[13] Lempel R. and Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect // In Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, 2000. – pp. 387–401.
[14] Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 8-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 115–119.
[15] Ландэ Д.В., Григорьев А.Н., Брайчевский С.М. Стабильность источников информации как один из параметров информационных потоков // Компьютерная лингвистика и интеллектуальные технологии : труды международной конференции Диалог'2006 – М. : Наука, 2006. – С. 332–334.
[16] Федер Э. Фракталы. – М. : Мир, 1991. – 254 с.

Ranking of information sources in the news monitoring system InfoStream

D.V. Lande, S.M. Brajchevskiy, A.T. Darmokhval, A.Y. Morozov

A few going is described near ranking and selection of informative sources at the construction of services of news monitoring. These approaches are based on principle of providing of maximal plenitude of information at the least of sources, choice of the most original, thematically stable, maximally quoted sources.

The results of the proper statistical researches are presented.

The principles examined in-process are used for the selection of sources for corporate applications of the news monitoring system InfoStream.