СТАБИЛЬНОСТЬ ИСТОЧНИКОВ КАК ОДИН ИЗ ПАРАМЕТРОВ ИНФОРМАЦИОННЫХ ПОТОКОВ

Д.В. Ландэ *(dwl@visti.net),*

А.Н. Григорьев *(gri@visti.net),*

С.М. Брайчевский *(smb@visti.net),*

ИЦ «ЭЛВИСТИ», Киев, Украина

Рассматривается понятие стабильности источников информации, ориентированное, прежде всего на новостные веб-сайты. Предлагается формула и алгоритм расчета уровня разброса информации от источника. Обосновывается практическая значимость этого параметра.

В настоящее время для решения широкого спектра информационно-аналитических задач все шире применяется информация, публикуемая на новостных веб-сайтах [1]. Один из возможных подходов к решению проблемы изучения сетевого информационного пространства основан на представлении его некоторым множеством источников, порождающих информационные потоки. Предполагается, что динамика этих потоков в определенном смысле более содержательна, чем динамика составляющих их сообщений.

При этом можно отметить очень высокий диапазон параметров этих источников, как по объемам публикуемой информации, так и по содержанию – от сообщений серьезных информационных агентств – до «живых журналов» школьников.

Источники информации, очевидно, характеризуются уровнем стабильности. Примером стабильных источников могут служить крупные информационные агентства, регулярно поставляющие потребителям примерно одинаковые объмы информации на протяжении длительного времени, а примером нестабильных – «живые журналы», многие из которых активно действуют в течение нескольких дней, а затем угасают.

Нестабильные источники по-своему интересны хотя бы тем, что, видимо, именно они ответственны за хаотичность динамической части сетевого информационного пространства. Однако они не связанны с его основными тенденциями и поэтому могут не приниматься в расчет при его систематических исследованиях. Напротив, ключевую роль здесь должны играть именно стабильные источники, отражающие (и в какой-то мере порождающие) реальные закономерности сетевой динамики.

На практике среди множества проблем подбора и анализа источников контента большое значение имеет учет параметров их стабильности, в частности, тематической. При этом тематическая стабильность и стабильность потока информации от источников зачастую играют решающую роль при проведении аналитических исследований. Например, такие важные свойства информационных источников, как их тематическая корреляция [2] и полнота, 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. Предполагается, 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 этом именно эта система, охватывающая более 30000 новостных сообщений в сутки, была выбрана в качестве экспериментальной платформы.

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 с некоторыми погрешностями в рубрикации при статистическом исследовании можно пренебречь.

Для подхода к изучению стабильности источников важно знать параметры их распределения по тематическим рубрикам, т.е. количество рубрик, соответствующих документам, входящим в эти источники. Результаты такого исследования, охватывающего 920 репрезентативных русскоязычных источников (опубликовавших за месяц более 100 сообщений), приведены на Рис. 1. Об относительно невысокой

тематической стабильности источников, порождающих общий информационный поток системы, свидетельствует тот факт, что около половины репрезентативных источников соответствуют более 20 рубрикам.

Для вычисления уровня разброса (нестабильности) источника информации использовалась формула, основанная на линейной метрике:

$$R = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M * \max(r_i)} \sum_{j=1}^{M} \left| r_{ij} - \frac{1}{M} \sum_{k=1}^{M} r_{ik} \right| ,$$

где N – количество рубрик, M – количество дней, $\max(r_i)$ – максимальное суточное количество вхождений рубрики i в документы источника за все время, r_{ij} –количество вхождения рубрики i за день j.

Из приведенной формулы следует, что значение R, на самом деле, учитывает не только тематический разброс, но и разброс по количеству вхождений рубрики, т.е. фактическое количество документов от источника, относящихся к данной рубрике. Даже если источник соответствует одной рубрике, но его наполнение не является стабильным, значение может существенно отличаться от нулевого.

Ранговая диаграмма распределения уровня разброса для источников – веб-сайтов, ежедневно публикующих новостные сообщения в ноябре 2005 года, по уровням тематической стабильности приведена на Рис. 2.

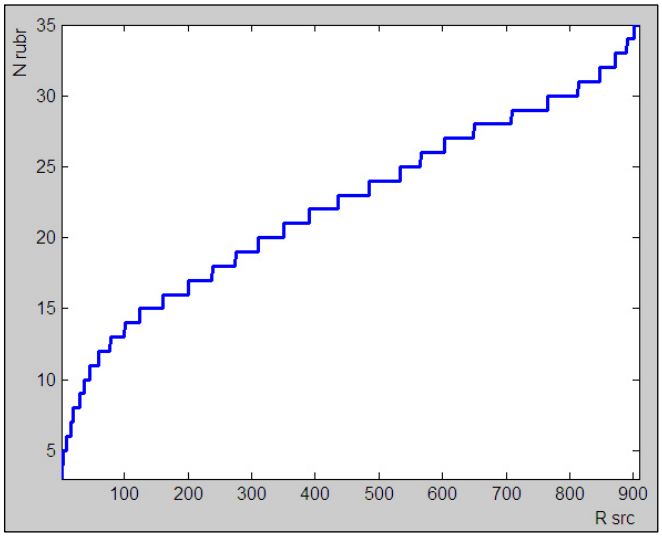

Рис. 1. Ранговая диаграмма “Источники – количество рубрик”

Определение стабильности документов выполнялось по такому алгоритму:

1. Проводился поиск документов в базе данных за определенный период.
2. Формировалась таблица, которая включала код источника информации, коды соответствующих ему тематических рубрик и их количество в разрезе дат.
3. Для каждого источника по приведенной выше формуле определялся уровень разброса R.
4. Информационные источники ранжировались по рассчитанным параметрам, и строилась соответствующая диаграмма.

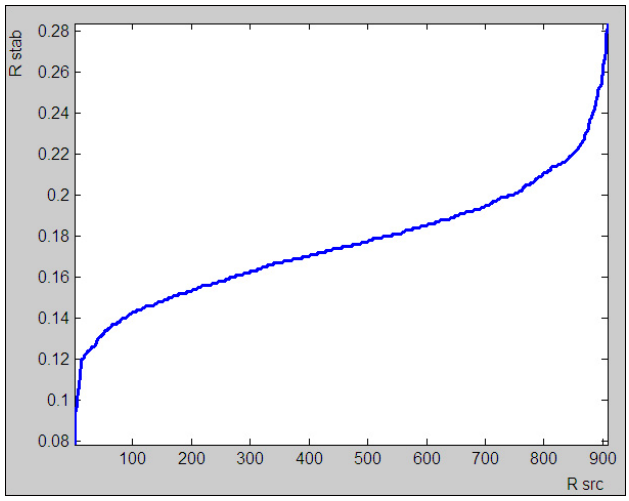

Рис. 2. Диаграмма “Ранг источника - коэффициент разброса”

Как оказалось, источники, содержащие до 5-6 рубрик обладают исключительной стабильностью, что, в общем, достаточно очевидно. Не совсем очевидным оказался факт резкого всплеска разброса для источников, включающих документы с 25 и более рубриками.

Результаты исследований стабильности источников 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 информационных источников» в интерфейсе системы контент-мониторинга информационных ресурсов.

Сегодня становится ясно, что разработка качественно новых средств работы с сетевыми ресурсами переходит в разряд приоритетных задач. Без приемлемых средств контроля за сетевыми информационными процессами невозможно обеспечить репрезентативность выборок. В любом случае, успешное продвижение в изучении современного информационного пространства невозможно без хотя бы общих предствлений о структуре и свойствах динамики сетевых информационных процессов, что в свою очередь требует выявления их устойчивых закономерностей.

Список литературы

1. Брайчевський С.М., Ландэ Д.В. *Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1, - 2005. - № 11. - С. 21-33*
2. Ландэ Д.В., Брайчевський С.М. *Определение тематической направленности запросов путем анализа набора рейтинговых источников / // Открытые информационные и компьютерные интегрированные технологи: Сб. научн. трудов. Вып. 29. – Харьков: Нац. аэрокосмический ун-т «Хаи», 2005. - С. 169-174*