RCDL2008
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ, ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
Десятая Всероссийская научная конференция
Дубна, 7 - 11 октября 2008 года

Ранжирование источников информации в системе мониторинга новостей InfoStream

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

Информационный Центр «ЭЛВИСТИ»

ПРЕДМЕТНАЯ ОБЛАСТЬ

Системы интеграции и мониторинга новостей из открытых веб-сайтов сети Интернет сегодня все чаще становятся основными компонентами информационных служб различного уровня.

Можно отметить разнообразный диапазон параметров информационных источников как по объемам публикуемой информации, так и по содержанию – от сообщений информационных агентств – до «живых журналов».

Мощные возможности Интернет порождают 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.

Система контент-мониторинга InfoStream на основании анализа около 3000 источников информации в сети Интернет позволила построить зависимость суточных объемов тематических публикаций за 3 года по выбранной тематике (1096 суток, общее количество – свыше 320 тысяч).

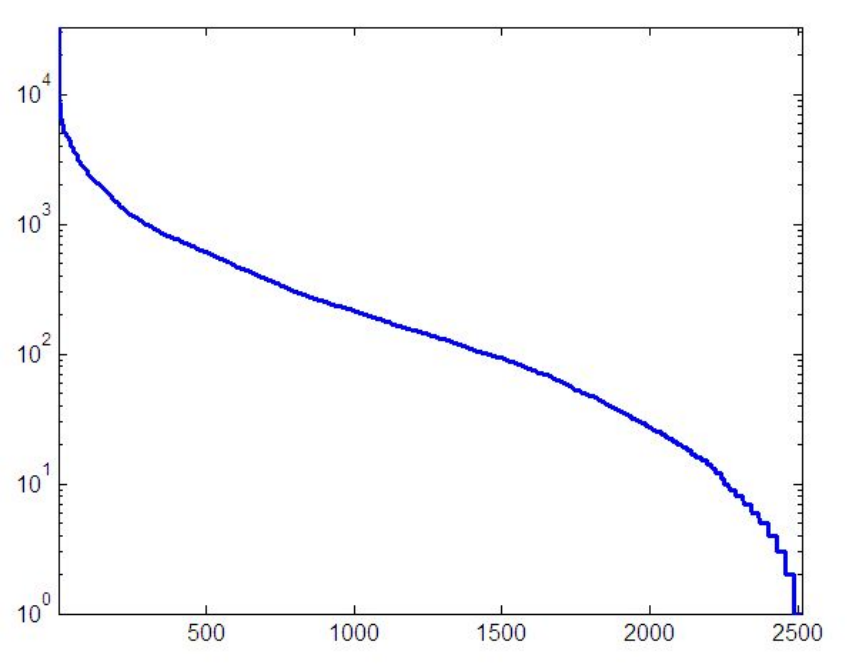
ТРАДИЦИОННЫЕ ПОДХОДЫ

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

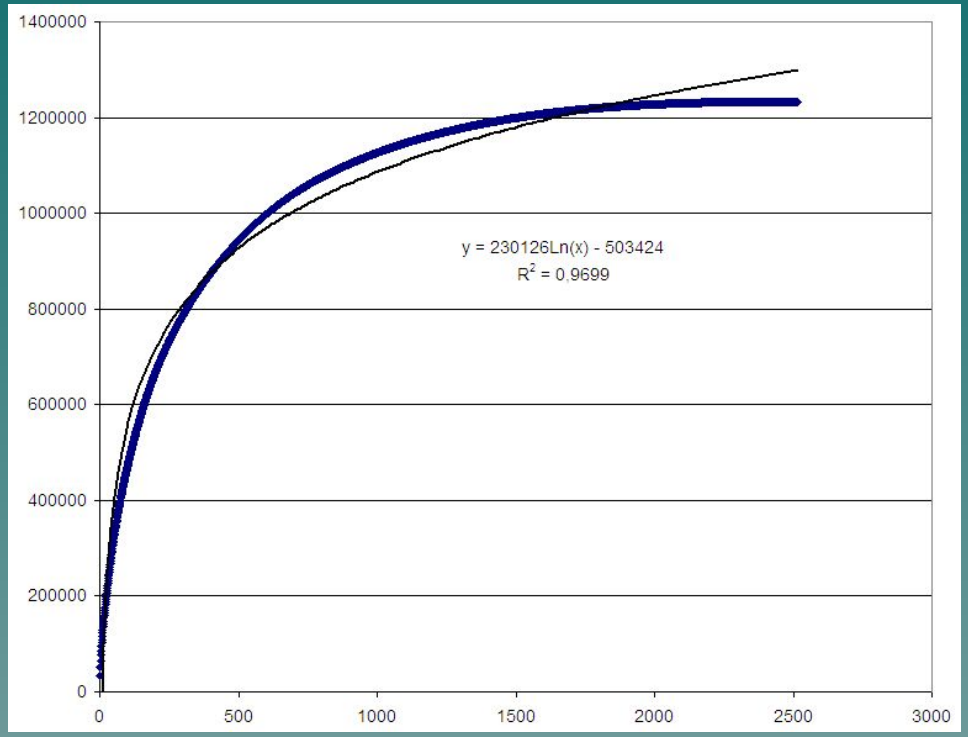
Основным критерием при отборе источников для таких систем мониторинга новостей является их содержание. Было показано, распределение источников по контенту, соответствующему тематическим потребностям корпоративного пользователя удовлетворяет закону Бредфорда, соответственно, при отборе источников обязательно должно учитываться их ранжирование по степени соответствия тематике. Однако, реализация такого отбора приводит к известным сложностям. На практике такое ранжирование осуществляется экспертами 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.

Распределение источников по количеству генерируемых документов

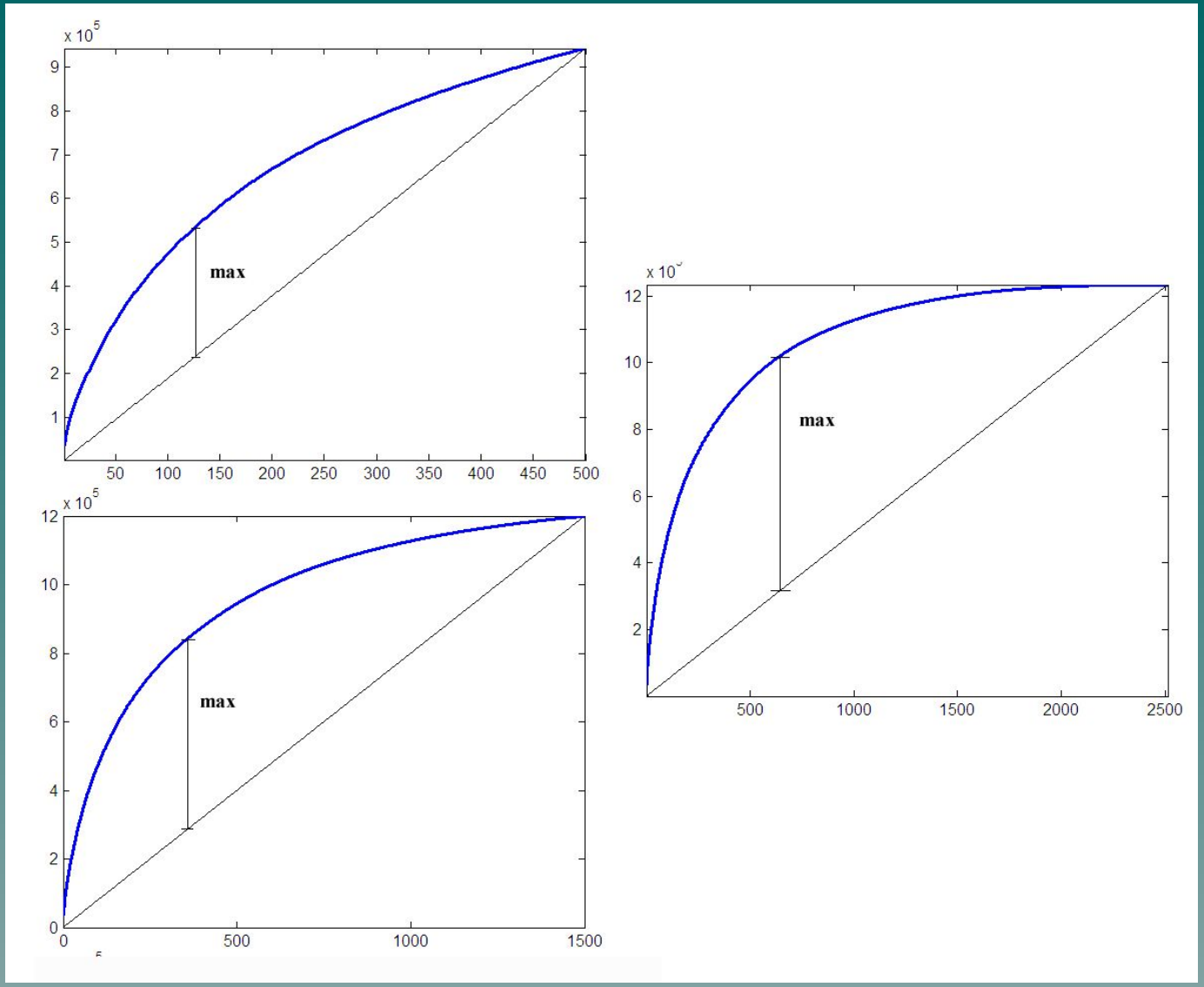
$$f(x) \sim \int \frac{a}{x} dx = a \ln x + C$$



Ранжированный список источников по количеству публикаций (ось 0Y)



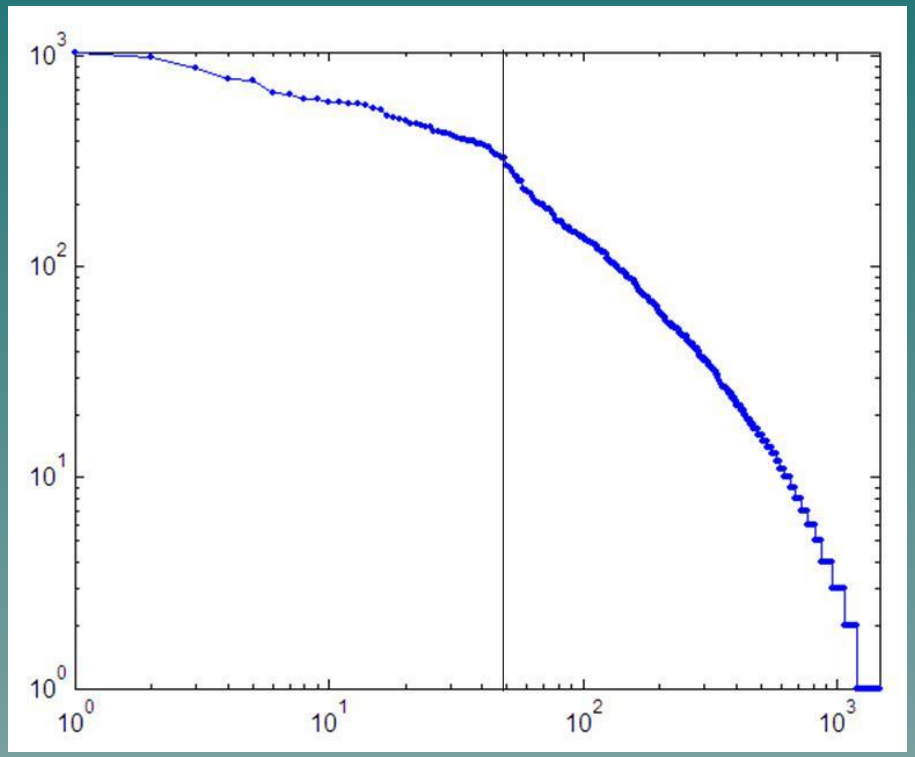
Количество публикаций в системе мониторинга в зависимости от источников, ранжированных по количеству документов



Количество публикаций в системе при подключении наиболее интенсивных источниковписок источников (500, 1000, 1500)

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

Наиболее цитируемые источники

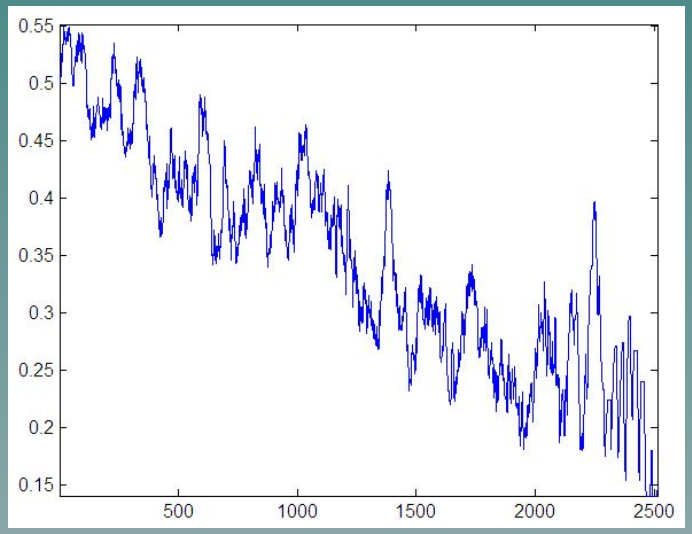


Web-сайт	Количество ссылающихся веб-сайтов
ИА «Интерфакс»	1051
«РосБизнесКонсалтинг»	983
"Reuters"	882
ИТАР-ТАСС	787
РИА «Новости»	773
УНИАН	675
Радио «Свобода»	662
НТВ	631
«Коммерсантъ»	623
BBC	598
«Комсомольская правда»	595

Зависимость количества ссылающихся веб-сайтов от ранга новостного источника в логарифмической шкале

Выбор наиболее оригинальных источников

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. 4), наглядно свидетельствующий об устойчивой тенденции: чем более продуктивен источник информации, тем больше он содержит заимствований из других источников.



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

Тематическая стабильность

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.

Авторами был предложен параметр тематической стабильности 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:

$$S_i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\{r_j^{(i)} - \frac{1}{M}\sum_{k=1}^{M} r_k^{(i)}\right\}^2},$$

где $r_j^{(i)}$ – количество вхождения рубрики i за день j,

M - количество значений ряда измерения (недель, например).

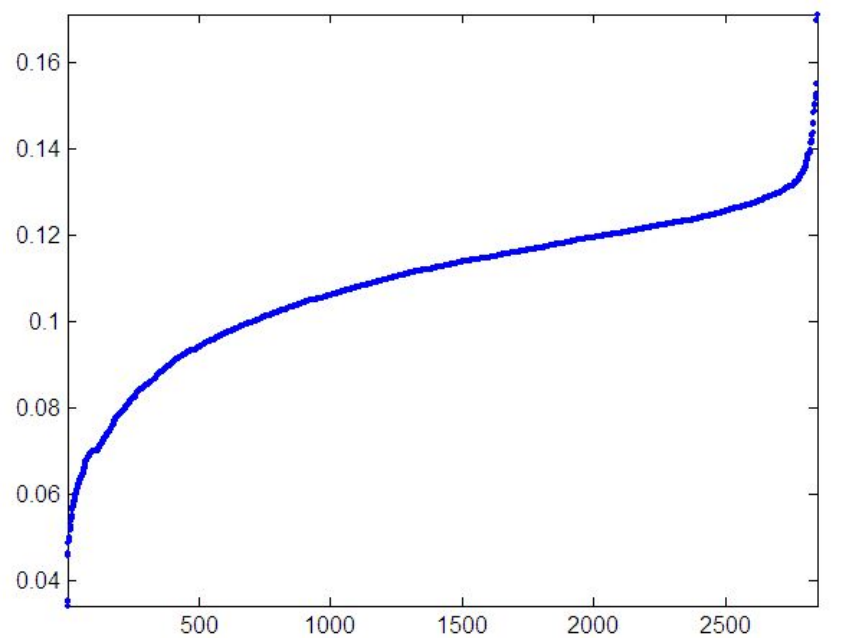
Значение R_i вычисляется следующим образом:

$$R_i = \max_{1<k<M} X_k^{(i)} - \min_{1<k<M} X_k^{(i)},$$

где $X_k^{(i)}$ – накопленное к моменту k отклонение по рубрике i, вычисляемое по формуле:

$$X_k^{(i)} = \sum_{j=1}^{k}\left(r_j^{(i)} - \frac{1}{M}\sum_{l=1}^{M} r_l^{(i)}\right).$$

Тематическая стабильность



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

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. Практически все ведущие информационные агентства, выпускающие политематическую информацию, тем не менее, вошли в состав наиболее тематически стабильных.

Некоторые выводы

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 приведенных критериев, дополняя его некоторыми неформальными соображениями.

Спасибо за внимание!

Д.В. Ландэ,
dwl@visti.net

Информационный Центр «ЭЛВИСТИ»,
Киев, Украина