УДК 004.738.52

Д. В. Ландэ

Структура новостного Web-пространства

Предлагается модель новостного Web-пространства, основанная на учете контекстных ссылок на отдельные источники информации. В отличие от существующих моделей Web-пространства представленный подход позволяет учитывать динамику новостных потоков, контекстные (а не только гипертекстовые) ссылки, эффект содержательного дублирования.

Эффективный анализ новостных информационных потоков в Интернет, построение систем синдикации новостей невозможны без некоторых сведений о структуре новостного Web-пространства [1]. Это пространство формируется динамичными потоками сообщений, публикуемых на Web-сайтах средств массовой информации, информационных агентств, отдельных организаций [2]. Чем может быть полезно знание структуры новостного Web-пространства на практике? Во-первых, это дает возможность выявления первоисточников информации [3], 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, "мусорных" источников. Кроме того, при оперативном поиске актуальной информации корректная модель новостного Web-пространства может способствовать обнаружению действительно полезных первоисточников и служб интеграции информации.

Если для обычного Web-пространства уже признана модель "галстука-бабочки", представленная в работах А. Бредера и его коллег [4], то публикации об архитектуре новостного Web-пространства автору неизвестны. Можно было бы применить вышеназванную модель и к новостной составляющей Web-пространства, однако такой подход нельзя считать корректным в силу нескольких причин.

1. Новостные потоки характеризуются динамикой [5], что сильно влияет на природу гиперссылок. Например, на наиболее актуальные сообщения в течение определенного времени ссылок может вообще не существовать.

2. Модель Бредера слабо учитывает особенности "скрытого" Web, т. е. тех информационных Web-ресурсов, на которые не существует прямых гиперссылок (исследователь рассматривал лишь ресурсы, охваченные поисковой системой AltaVista).

3. В новостных потоках необходимо учитывать не только гиперссылки, но и контекстные ссылки (причем это могут быть ссылки и на объекты из открытой части Web-пространства, и на ресурсы, доступные только по паролю, и даже офлайновые публикации изданий).

4. Модель Бредера не включает такого понятия, как содержательное дублирование информации.

5. При построении модели структуры новостного Web-пространства наибольшее внимание должно уделяться именно Web-сайтам, на которых публикуются новостные сообщения, а не отдельным Web-страницам или самим сообщениям.

В качестве экспериментальной базы для построения модели новостного Web-пространства автором использовался достаточно мощный информационный корпус — ретроспективная база данных системы контент-мониторинга InfoStream [2]. Система InfoStream применяется для решения задач автоматизированного сбора новостной информации с открытых Web-сайтов и обеспечения доступа к ней в поисковых режимах. Эта система, разработанная в Информационном центре "ЭЛВИСТИ", в настоящее время охватывает около 2000 источников, а ее ретроспективные базы данных представляют собой корпус объемом более 30 млн документов. Для построения модели использовалась база данных новостных сообщений за февраль 2006 г. объемом около 760 тыс. документов. Для каждого из источников был составлен запрос в следующем виде:

<код источника>#<шаблон для поиска>[#<шаблон для поиска>...#<шаблон для поиска>].

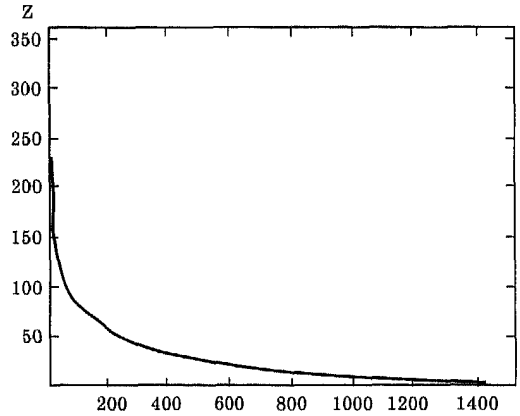
Совокупность запросов была объединена в конфигурационном файле, фрагмент которого представлен ниже:

```
srd00001#УКРОП#ukrop.com
srd00002#BBC#bbc.co.uk
srd00003#Champion.com.ua#champion.com.ua
srd00004#Crashes.ru#crashes.ru
srd00006#"Немецкая волна#Deutsche Welle#dwelle.de
srd00007#GazetaSNG.ru#gazetasng.ru
srd00008#idNews.com.ua#idnews.com.ua
srd00011#InoPressa#Инопресса#inopressa.ru
srd00012#Internet.ru#internet.ru
srd00013#K2Kapital#k2kapital.com
srd00014#KPNews.com#kpnews.com
srd00015#Lenta.Ru#Лента.ру#lenta.ru
srd00016#MIGnews.com#mignews.com
```

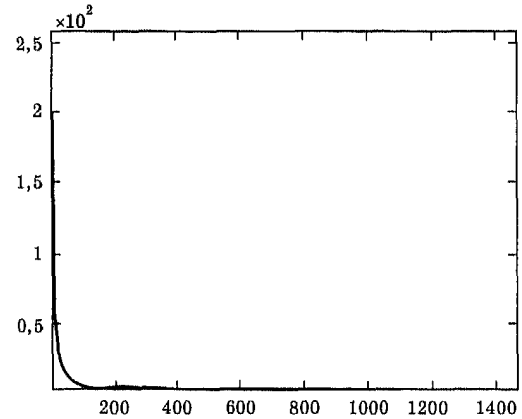
В результате специальной обработки такого пакета запросов для каждого сообщения, относящегося к определенному источнику (Web-сайту), были выявлены **исходящие ссылки** на другие источники (ссылки на собственный источник исключались). Было установлено, что исходящие контекстные ссылки присутствовали на 264 942 сообщениях с 1531 Web-сайта. Общее же количество Web-сайтов, участвующих в процессе взаимных ссылок, составило 1863. Было также выявлено 54 источника, не входящих в этот список, т. е. тех, на которые не вела ни одна из контекстных ссылок и сообщения из которых не ссылались ни на один из исследуемых Web-сайтов. Такие Web-сайты ("абсолютные острова") были вынесены за рамки модели.

Ниже приведен список Web-сайтов, обладающих **максимальным количеством исходящих ссылок:**

Web-сайт	Количество ссылок
RAMBLER	363
VLASTI.NET	271
RosInvest	270
"Обозреватель"	231
ИА "REGNUM"	217
"Деловая пресса"	202
"Россия-Он-Лайн"	193
"Оглядач"	191
RNews	183
Fin.org.ua	166
PRESIDENT.ORG.UA	164
"Промышленно-торговые новости"	160
"4 ВЛАДА"	159
"Украина промышленная"	156



Распределение по количеству исходящих ссылок



Распределение по количеству входящих ссылок

Рис 1 Распределения по количеству ссылок

Также было получено распределение новостных Web-сайтов по количеству **входящих ссылок**. Всего за февраль 2006 г. ссылки указывали на 1470 источников (без самоцитирования). Оказалось, что на 100 источников ведет свыше 80% ссылок. На рис. 1 представлены графики ранжированных распределений новостных Web-сайтов по количеству исходящих и входящих ссылок. Следует обратить внимание на то, что второй график значительно круче первого, — это говорит о большей равномерности распределения множества исходящих ссылок, чем входящих. Ниже приведен **начальный фрагмент ранжированного списка источников, на которые ведет максимальное количество ссылок:**

Web-сайт	Количество ссылок
РИА "Новости"	25 827
ИА "Интерфакс"	23 765
ИА "REGNUM"	22 354
УНИАН	17 847
ИТАР-ТАСС	14 157
"Reuters"	10 754
"Газета.Ru"	9354
РИА "РосБизнесКонсалтинг"	7653
УНІАН	5472
"Lenta.Ru"	5223
ИА "Интерфакс-Украина"	5073
ИА "Росбалт"	5031
"proUA"	4814
НТВ	4395

Кроме того, были выявлены источники, на которые не ссылаются, но которые обладают исходящими ссылками (393), и цитируемые источники, не ссылающиеся ни на кого (332).

Специальное место в исследовании занимало изучение **смыслового дублирования информации**. Следует отметить, что процент дублирующихся сообщений в системе InfoStream значительно меньше, чем во всем новостном Web-пространстве. Это объясняется подбором источников для сканирования (в число их не входят многие новостные интеграторы).

Выявление дублирующихся по содержанию новостных сообщений в системе InfoStream выполнялось на основе лингвостатистических методов, заключающихся в нахождении наиболее весомых слов в документах, которые выступают своеобразными ключами. Опыт показал, что в русско-и украиноязычных потоках новостей совпадение шести наиболее весомых ключевых слов с более чем 95%-ной вероятностью свидетельствует о содержательном дублировании документов*.

Исследование соотношения дублирующихся и оригинальных сообщений привело к неожиданному результату: количества оригинальных сообщений и их содержательных дублей, охватываемых системой InfoStream в 2005 г., почти совпали (рис. 2).

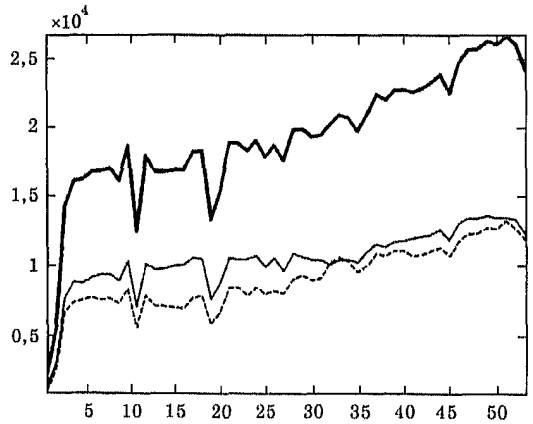


Рис. 2 Объемы информации, сканированной системой InfoStream в 2005 г , в разрезе недель. Условные обозначения сплошная жирная линия – общий объем сообщений, сплошная тонкая линия - оригинальные сообщения, пунктирная линия – информационные дубли

* Применение более "мягкого" критерия к множеству отобранных ключевых слов позволяет реализовать режим "поиска подобных документов"

Следует заметить, что устранение дублирующихся сообщений в информационных потоках требуется далеко не всегда. Существует ряд задач, в которых используется факт дублирования текстов сообщений в различных источниках (например, при определении важности сообщения или эффективности PR-кампаний). При построении модели новостного Web-пространства исследовался уровень дублирования для источников, ранжированных по количеству исходящих ссылок. Как выяснилось, до определенного значения (порядка 800) уровень дублирования значительно превышает средний, равный ~1/2. При небольшом количестве исходящих ссылок этот уровень понижается, однако при минимальном количестве ссылок снова возрастает. Можно считать, что значения рангов источников 1400 и выше соответствуют "зоне массового плагиата" (ссылок мало, а уровень дублирования высокий).

В результате проведенных исследований была принята **модель новостного Web-пространства**, которая представлена на рис. 3. Эта модель включает следующие зоны:

ядро, состоящее из трех областей: входной, выходной и коммуникационной зон (таких Web-сайтов оказалось 680, или 36,5%); зона ядра характеризуется средними и большими значениями уровней исходящих и входящих связей, однако допускает ранжирование по уровню этих коммуникаций;

входной полуостров: Web-сайты, которым соответствуют менее порогового значения входящих ссылок и любое, превышающее пороговое, количество исходящих ссылок (таких Web-сайтов оказалось 312, или 16,7%);

выходной полуостров: Web-сайты, которым соответствуют менее порогового значения исходящих ссылок и любое, превышающее пороговое, количество входящих ссылок (таких Web-сайтов оказалось 513, или 27,5%);

остров: Web-сайты, которым соответствует менее порогового значения исходящих и входящих ссылок (таких Web-сайтов оказалось 358, или 19,3%).

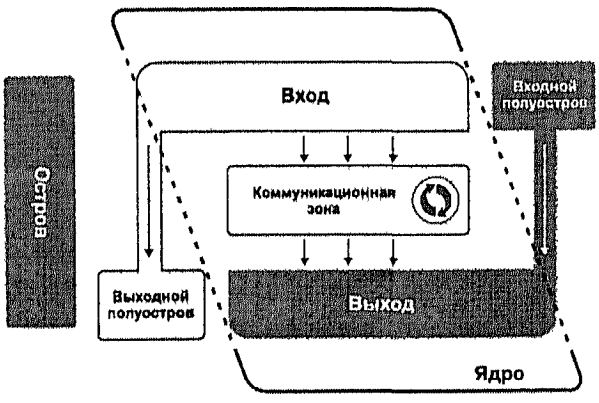


Рис. 3. Архитектура новостного Web-пространства

Основа модели строилась путем анализа полной картины распределения входных и выходных ссылок. При этом были созданы матрица инциденций и соответствующие графы связи, а также выявлены необходимые кластеры [6]. Вместе с тем оказалось, что само по себе отношение количества входящих и исходящих ссылок для каждого из источников достаточно точно характеризует его попадание в названные кластеры. Например, для разделения области ядра на входную, выходную и коммуникационную зоны можно рассмотреть ранжированный график логарифма отношения количества входных и исходящих ссылок для каждого из источников этой области (рис. 4). Центральная зона данного графика соответствует коммуникационной, левая — выходной, а правая — входной зоне.

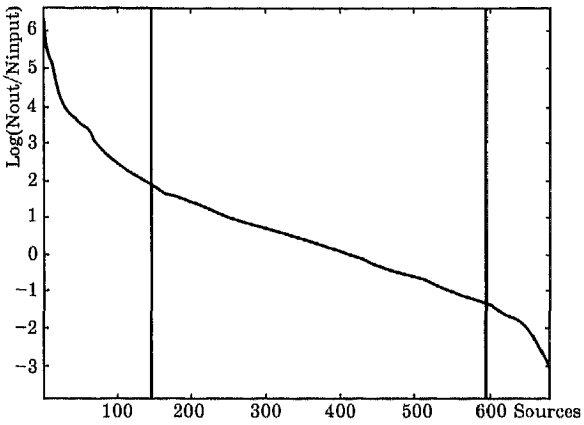


Рис. 4. Ранжированный график логарифма отношения количества ссылок

Интересным оказался график двумерного сечения значений $\log(N_{\text{out}} + 1)$, $\log(N_{\text{in}} + 1)$, где N_{out} — количество входящих ссылок, N_{in} — количество исходящих ссылок для каждого из источников (рис. 5). Этот график послужил основой идеальной схемы представления областей модели в зависимости от количества исходящих и входящих ссылок (рис. 6).

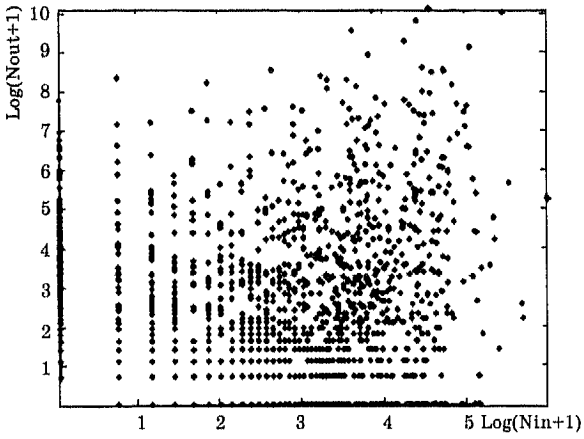


Рис. 5. График распределения зоны ядра в координатах "логарифм количества исходящих ссылок — логарифм количества входящих ссылок"

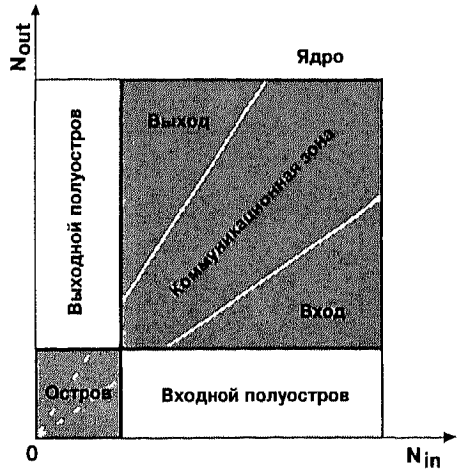


Рис. 6. Представление областей модели в зависимости от количества исходящих и входящих ссылок

В результате проведенных исследований была построена модель новостного Web-пространства (см. рис. 3, 6), основанная на контекстных ссылках, а также предложены подходы к выявлению ее основных зон и рассчитаны числовые соотношения этих зон. Вместе с тем данная модель предполагает дальнейшее развитие в следующих направлениях:

1) более точная идентификация контекстных ссылок;

2) совершенствование критерия определения зон на основе полного учета структуры ссылок и методов кластерного анализа;

3) совершенствование механизма определения содержательного дублирования информации (в том числе за счет механизмов настройки сканеров системы контент-мониторинга, учета авторитетности источников и возможных умышленных задержек публикации в Интернет).

СПИСОК ЛИТЕРАТУРЫ

1. Брайчевский С. М., Ландэ Д. В. Современные информационные потоки: актуальная проблематика // НТИ. Сер. 1.— 2005.— № 11.— С. 21-33.

2. Ландэ Д. В. Поиск знаний в Internet. Профессиональная работа.— М.: ИД "Вильямс", 2005.— 271 с.

3. Ландэ Д. В., Фурашев В. Н. Вопросы построения и использования многокритериальной модели выбора источников информации // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов.— Харьков, 2006. Вып. 30.— С. 76-85.

4. Broder A. Z., Glassman S. C., Manasse M. S., Zweig G. Syntactic Clustering of the Web / 6th international conference on World Wide Web // Computer Networks.— 1997.— Vol. 29, № 8-13. — P. 1157-1166.

5. Del Corso G. M., Univerisity A. G., Romani F. Ranking a stream of news // Proceedings of the 14th international conference on World Wide Web (Chiba, Japan). 2005.— P. 97-106.

6. Ландэ Д. В. Некоторые методы анализа новостных информационных потоков // Научные труды Донецкого национального технического университета. Сер. Информатика, кибернетика и вычислительная техника.— Донецк: ДонНТУ, 2005. Вып. 93.- С. 277-287.

Материал поступил в редакцию 25.05.06.
