**УДК 004.5**

# МОДЕЛИРОВАНИЕ ПОВЕДЕНИЯ ТЕМАТИЧЕСКИХ СЮЖЕТОВ НОВОСТЕЙ В ВЕБ-ПРОСТРАНСТВЕ

Д.В. Ландэ (*dwlande@gmail.com*)
Институт проблем регистрации информации НАН Украины, Киев

С.М. Брайчевский (*smb@visti.net*)
Информационный центр «ЭЛВИСТИ», Киев

Статья посвящена изучению поведения тематических сюжетов в новостных информационных потоках. Рассматривается модель, предполагающая существование трех независимых фаз в периоде существования сюжета, которая позволяет выделить ряд типичных профилей.

Веб-пространство представляет собой динамическую систему из связанных по смыслу элементов (документов), образующих в динамике своей эволюции информационные потоки [Kleinberg J., 2006] , [Del Corso G.M. et al, 2005], [Atkinson M. et al, 2009], [Lande D. et al, 2007], [Lande D. et al, 2005].

Основным объектом моделирования информационных потоков [Ландэ Д. и др., 2007] сегодня являются тематические сюжеты новостей, последовательности сообщений, соответствующих определенной тематике. Тематическим сюжетам новостей можно поставить в соответствие временные ряды, для решения задач анализа которых все чаще применяются дисперсионный, фрактальный, вейвлет-анализ [Астафьева Н.М., 2007], [Buckheit J. et al, 2007] , [Lande D. et al, 2009].

Многочисленные факты свидетельствуют о том, что динамика тематических информационных потоков определяется комплексом внутренних нелинейных механизмов, которые лишь частично коррелируют с реальностью. Тематический сюжет новостей, по-видимому, следует понимать как информационную категорию, не связанную непосредственно с объектами или явлениями реального мира. Например, сюжет может быть полностью вымышленным, с другой стороны, он может связан со многими событиями. Следует признать, что поведение

тематических сюжетов только опосредовано связано с динамикой событий.

В практическом плане часто оказывается удовлетворительным упрощенное понимание тематического сюжета как некоторой зависимой от времени величины $n(t)$, которая описывается некоторым нелинейным уравнением. Сегодня классическими считаются два класса моделей информационных потоков (соответственно, тематических сюжетов новостей): линейные и экспоненциальные. Оба класса имеют существенную ограниченность – монотонный характер. Поэтому они малопригодны для изучения реальной динамики в течение длительных интервалов времени. Сегодня при моделировании информационных потоков используются преимущественно нелинейные модели, применяются методы нелинейной динамики, теории клеточных автоматов, перколяции, самоорганизованной критичности [Ландэ Д. и др., 2009], [Додонов А. Г. и др., 2011].

В настоящее время существует несколько открытых информационных сервисов, в рамках которых можно наблюдать временную динамику объемов публикаций по тематикам, определяемым запросами. Так Google books Ngram Viewer (http://ngrams.googlelabs.com/), предоставляет визуализацию динамики количества книг, в которых упоминаются слова. На рис. 1 приведен пример динамики количества публикаций, в которых встречались слова «Хрущев» и «Брежнев» с 1940 по 1990 год.

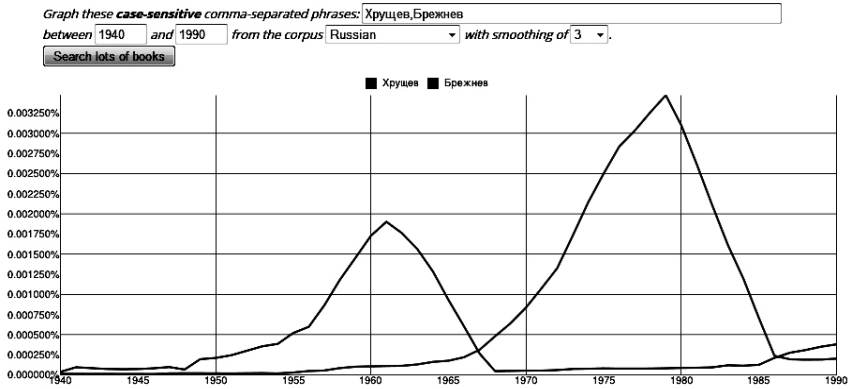


Рис.1. Динамика публикаций (Google books Ngram Viewer), содержащих заданные слова

Сервис «Яндекс пульс блогосферы» (http://blogs.yandex.ru/pulse/) также позволяет отображать динамику публикаций в блогах, содержащих заданные пользователем ключевые слова. На рис. 2 приведена динамика сообщений, соответствующих запросам «Олланд» и «Саркози» за период с ноября 2011 по май 2012 года.


Рис.2. Динамика блогов, содержащих заданные слова

На сайте Национального корпуса русского языка (НКРЯ) в бета-режиме запущен сервис N-грамм (http://www.ruscorpora.ru/ngram.html), близкий по функциональности сервису Google books Ngram Viewer. На рис. 3 представлена динамика публикаций, соответствующих запросам «шовинизм» и «космополитизм» за период с 1820 по 2010 год.

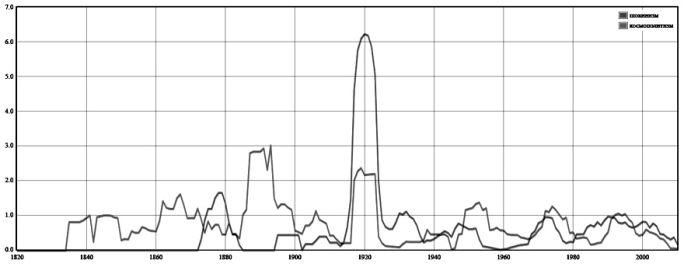


Рис.3. Динамика публикаций (Национальный корпус русского языка), содержащих заданные слова

Многие современные информационно-аналитические системы содержат в своем составе средства отображения статистики вхождения в базы данных понятий, соответствующих пользовательским запросам. В частности, авторами реализована подсистема статистики в рамках системы контент-мониторинга веб-пространства InfoStream [Григорьев А.Н. и др, 2012], реализующая данную функциональность.

В результате анализа многочисленных диаграмм поведения тематических сюжетов новостей, были выявлены наиболее типичные, базовые профили их поведения (рис. 4). Некоторые сюжеты развиваются

следующим образом: после быстрого информационного всплеска подготовки идет плавный спад (например, публикации о стихийных бедствиях, рис 4 *а*), некоторые, напротив предполагают длительную плавную информационную подготовку, после чего идет резкий спад (например, публикации об планируемых заранее мероприятиях). Существуют также тематические сюжеты, характеризующиеся симметричной кривой динамики, как узкие, кратковременные, так и растянутые во времени (рис 4 *в*). Ниже будут приведены примеры поведения реальных тематических сюжетов новостей из веб-пространства, полученные с помощью системы InfoStream.

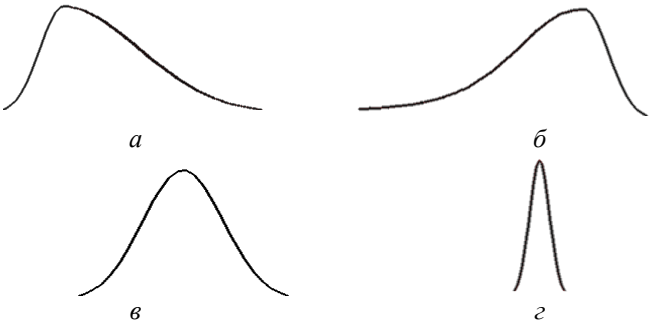


Рис.4. Базовые профили динамики тематических сюжетов новостей

В соответствии с многочисленными эмпирическими данными были определены установки предлагаемой ниже модели. В динамике сюжетов выделяется три фазы: предактуальная, актуальная, и постактуальная. Центральная идея заключается в том, что каждой фазе присущи собственные механизмы генерации сообщений, которые порождают разные динамические зависимости. Эти механизмы в настоящее время малоизучены, но, в любом случае, отношение читателей к свежей теме качественно отличается от их отношения к теме, известной уже давно. С другой стороны, поскольку речь идет об одном и том же процессе генерации сообщений, естественно ожидать, что эти механизмы должны быть структурно подобны. Поэтому в предлагаемой модели предлагается использовать одну и ту же форму зависимости, но с разными значениями параметров. Непрерывность полной зависимости обеспечивается «сшивкой» соответствующих выражений на границах областей.

Основной считается актуальная фаза, которая ограничивается начальным $t_{pre}$ и конечным $t_{post}$ моментами времени. Ее динамика описывается константой $n(t) = N_a$ (количество публикаций $N_a$ максимально и не зависит от времени).

Предактуальная фаза характеризуется ростом востребованности публикаций данного сюжета. Из наблюдений мы знаем, что такие

зависимости (в самых разных процессах природы и общества) обычно имеют точку перегиба, соответствующую максимальной скорости роста.

Такую динамику можно, например, описать выражением:

$$n_{pre} = N_a \exp[-c_{pre}(t - t_{pre})^2],$$

где $N_a$ – нормировочный множитель, зависящий от выбранного масштаба, а $c_{pre}$ – характеризует интенсивность обратных связей между генераторами сообщений и читателями.

Постактуальная фаза характеризуется спадом востребованности публикаций данного сюжета. По причинам, аналогичным предыдущей, ее динамику опишем следующим выражением:

$$n_{post}(t) = N_a \exp[-c_{post}(t - t_{post})^2].$$

Таким образом, имеем:

$$n(t) = \begin{cases} n_{pre} = N_a \exp[-c_{pre}(t - t_{pre})^2], & t < t_{pre} \\ n(t) = N_a, & t_{pre} \le t \le t_{post} \\ n_{post}(t) = N_a \exp[-c_{post}(t - t_{post})^2], & t > t_{post} \end{cases}$$

В вырожденном случае $t_{pre} = t_{post}$ (актуальная фаза вырождается в точку или, в реальности, очень короткая) и $c_{pre} = c_{post}$ выражение для $n(t)$ соответствует распределению Гаусса – отклонения числа публикаций от максимума распределены по нормальному закону.

В зависимости от численных значений параметров $c_{pre}$, $c_{post}$, и $d = t_{post} - t_{pre}$ можно получать различные профили: с крутыми или пологими фронтами и различной длины «полочками».

Обратимся к некоторым теоретическим моментам, которые можно считать дискуссионными. Известное уравнение Мальтуса,

$$\frac{dn(t)}{dt} = an(t),$$

которое является основой классической экспоненциальной модели информационных потоков, обладает указанными выше недостатками.

Авторами ранее рассматривались для описания конкуренции тем логистические уравнения [Lande D. et al, 2005] типа:

$$\frac{dn(t)}{dt} = an(t)[c - n(t)]$$

Для данной задачи можем использовать следующее обобщение этих уравнений:

$$\frac{dn(t)}{dt} + a_i(t - b_i)n(t) = 0$$

для предактуального и постактуального интервалов времени. Смысл этого уравнения заключается в том, что скорости изменения количества публикаций в момент времени в предактуальной и постактуальной фазах зависят от величины отклонения от временных границ актуальной фазы с различными значениями параметров. Причем с ростом отклонения число сообщений уменьшается, как в одну, так и в другую сторону.

Решения приведенных уравнений имеют вид приведенных выше зависимостей ( $a_i = 2c_i, b_i = t_i$ ).

На рис. 5 изображена динамика тематического сюжета новостей, полученного по запросу «Фукусима». Очевидно, что диаграмма соответствует типу *а*, представленному на рис. 4 (резкий подъем, плавный спад).

Визуализации особенностей рядов измерений посвящены многочисленные исследования. В частности, для отображения неравномерностей во временном ряду использовался метод [Lande D., 2012], основанный на учете аномальных значений и популярной концепции одномерных клеточных автоматов. С помощью этого метода не детектируются абсолютные амплитудные всплески, однако он хорошо показал себя на «изрезанных» структурах данных, близких к фрактальным. К таким данным относятся, в частности, и временные ряды, связанные с объемами публикаций в веб-пространстве по определенным темам. Если область значений представляют собой выпуклое вверх множество, то визуальное представление клеточных автоматов принимает вид сплошной черной полосы, области изрезанности, нестабильности в ряде измерений могут вызывать появление «клетчатых» структур (рис. 7).

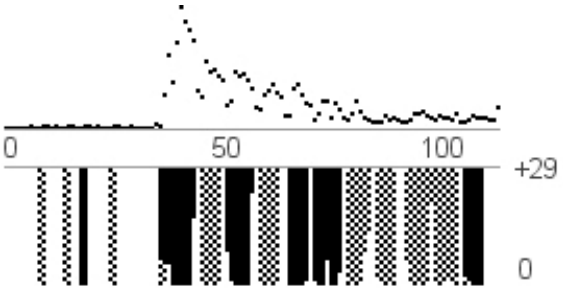


Рис. 5. Динамика сюжета новостей «Фукусима» типа *а* и клеточная диаграмма сюжета новостей «Фукусима»

На рис. 6 представлена динамика тематического сюжета новостей, полученного по запросу «Евровидение». В этом случае диаграмма соответствует типу *б*, (плавный подъем – информационная подготовка мероприятия, а затем резкий спад).

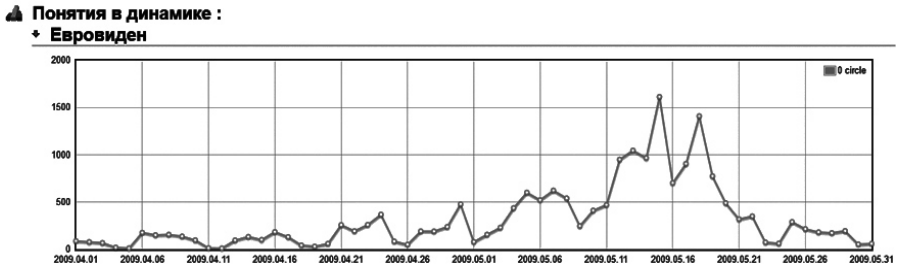

Рис. 6. Динамика сюжета новостей «Евровидение» типа б

В случае информационных потоков, которые ассоциируются с конкретными тематическими сюжетами новостей, необходимо описывать динамику каждого из таких потоков отдельно, принимая во внимание то, что рост одного из них автоматически приводит к уменьшению других и наоборот. Поэтому ограничение на количество сообщений по всем тематикам распространяется и на совокупность всех тематических сюжетов новостей. В случае изучения общего информационного потока наблюдается явление «перетекания» публикаций из одних, теряющих актуальность тематических сюжетов, в другие. Общая динамика должна описываться системой уравнений, каждое из которых относится к отдельному тематическому сюжету.

Выше была рассмотрена модель, в которой выделены три относительно независимые фазы. Предактуальная фаза характеризуется тем, что сюжет начинает привлекать интерес как генераторов сообщений, так и потребителей, в результате чего количество сообщений возрастает. Актуальная фаза наступает в тот момент, когда сюжет достигает максимальной востребованности у потребителей. В актуальной фазе число сообщений не меняется во времени. И, наконец, постактуальная фаза начинается в тот момент, когда интерес к сюжету начинает ослабевать, что сопровождается уменьшением числа сообщений.

Зависимость числа сообщений от времени в предактуальной и постактуальной фазах в рамках нашей модели соответствуют закону нормального распределения. Это означает, что модель может трактоваться как описывающая стохастические процессы усиления и ослабления интереса к тематическому сюжету и, как следствие, скорости увеличения и уменьшения числа сообщений по мере отдаления от актуальной фазы. Данная зависимость, вообще говоря, является ассиметричной, что вполне согласуется с реальными данными.

Отметим, что предложенная модель позволяет ограничивать сюжеты, поведение которых определяется естественными закономерностями медийного пространства, от сюжетов, освещение которых в медийных средствах испытывает влияние внешних факторов. Действительно, таким

индикатором может быть отклонение фронтов профиля от характерных форм распределения.

Авторы благодарны коллегам и соавторам по другим работам А.А. Снарскому, А.Т. Дармохвалу, А.Г. Додонову и В.Н. Фурашеву за внимание, поддержку и интерес, проявляемые при обсуждении рассматриваемых подходов.

## Список литературы

**[Atkinson M. et al, 2009]** Atkinson M. , Van der Goot E. Near real time information mining in mulitlingual news // in WWW '09: Proceedings of the 18th international conference on World Wide Web. ACM, 2009. – P. 1153–1154.

**[Buckheit J. et al, 2007]** Buckheit J., Donoho D. Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

**[Del Corso G.M. et al, 2005]** Del Corso G.M., Gulli A., Romani F. Ranking a Stream of News. In Processing of the 14th International World Wide Web Conference, 2005.

**[Kleinberg J., 2006]** Kleinberg J. Temporal dynamics of on-line information streams // Data Stream Management: Processing High-Speed Data Streams. – Springer, 2006.

**[Lande D., 2012]** Lande D.V. Visualization of features of a series of measurements with one-dimensional cellular structure // Preprint Arxiv: 1205.4234, 2012.

**[Lande D. et al, 2005]** Lande D.V., Braichevskii S.M. Dynamics of thematic information flows// arXiv:0805.4081, 2005.

**[Lande D. et al, 2007]** Lande D., Braichevski S, Busch D. Informationsfluesse im Internet // IWP - Information Wissenschaft & Praxis, 59(2007), Heft 5. – S. 277-284.

**[Lande D. et al, 2009]** Lande D.V., Snarskii A.A. Diagram of measurement series elements deviation from local linear approximations // Preprint Arxiv: 0903.3328, 2009.

**[Астафьева Н.М., 2007]** Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук, 1996. – 166. – № 11. – P. 1145-1170 (http://www.isuct.ru/~artcol/ articles/ Uspekhi_Fiz_Nauk/wavelet-analys.pdf).

**[Додонов А. Г. и др., 2011]** Додонов А. Г., Ландэ Д.В. Живучесть информационных систем. – К. : Наук. думка, 2011. – 256 с.

**[Григорьев А.Н. и др, 2012]** Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – Киев: ООО "Старт-98", 2007. – 40 с.

**[Ландэ Д. и др., 2007]** Ландэ Д.В., Снарский А.А., Брайчевский С.М., Дармохвал А.Т. Моделирование динамики новостных текстовых потоков // Интернет-математика 2007: Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. ун-та, 2007. – С. 98-107.

**[Ландэ Д. и др., 2009]** Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.