

Моделирование динамики новостных текстовых потоков

Ландэ Д.В.
ИЦ «ЭЛВИСТИ»
dwl@visti.net

Снарский А.А.
НТТУ «КПИ»
asnarskii@gmail.com

Брайчевский С.М.
ИЦ «ЭЛВИСТИ»
smb@visti.net

Дармохвал А.Т.
ИЦ «ЭЛВИСТИ»
hval@visti.net

Аннотация

В поведении текстовых информационных потоков, порождаемых в сети Интернет, наблюдаются две тенденции: постоянный рост объемов и усложнение динамической структуры. В связи с этим становится актуальной проблема моделирования динамики информационных потоков. Именно этому вопросу была посвящена данная работа.

Приводятся как теоретические выводы, так и результаты экспериментального анализа динамики информационных потоков, обрабатываемыми в рамках технологии контент-мониторинга InfoStream.

1. Введение

В последние годы в сфере информационных технологий понятие потоков стало одним из наиболее важных и актуальных [1, 2]. В связи с этим возрастает интерес к моделированию динамических процессов их генерации и распространения.

В литературе сегодня в основном обсуждается два класса моделей информационных потоков: линейные и экспоненциальные. Среди последних выделяется модель Бартона-Кеблера [3, 4]. Оба класса имеют существенную ограниченность – монотонный характер временной зависимости, поэтому они малопригодны для изучения реальной динамики сетевых информационных потоков.

Наблюдения временных зависимостей числа новостных сообщений определенного типа убедительно свидетельствуют о том, что механизмы их генерации и распространения связаны со сложными нелинейными процессами.

В настоящей работе предлагаются результаты моделирования характеристик тематических информационных потоков в рамках логистической модели. Эта модель в последние годы широко используется в различных областях. Несомненным ее достоинством является также сочетание в ней простоты исходных формулировок с гибкостью в постановке задач.

2. Документальные потоки

2.1. Общие замечания

При изучении и моделировании динамических свойств информационных потоков в рамках данной работы были приняты определенные допущения.

Предполагается, что существует система, сканирующая новостную информацию с веб-сайтов

сети Интернет (либо любой другой информационной среды) по мере публикации этой информации¹. Т.е. на входе такой системы – веб-сайты Интернет, а на выходе – последовательность сообщений, следующих одно за другим по мере публикации. В узком смысле в рамках данной работы под информационным потоком понимался дискретный числовой ряд, члены которого соответствуют количеству публикаций за единицу времени, выдаваемых такой идеальной системой.

При таком подходе фактически анализируется поток элементарных единиц контента. В качестве такой единицы будем рассматриваться документ. В узких рамках данной работы не различаются понятия «документ», «сообщение» или «публикация».

2.2. Тематические информационные потоки

Под тематическим информационным потоком (ТИП) будем понимать последовательность сообщений, соответствующих определенному тематическому запросу.

Итак, под ТИП в данной работе понимать число документов в единицу времени, относящихся к заданной теме, сканируемых из сети и фильтруемых по тематическому информационному запросу системой контент-мониторинга.

Рассмотрим общую картину динамики ТИП, ограничившись механизмами, типичными для новостного сегмента Интернет. Главное достоинство этого сегмента заключается в том, что он предоставляет в распоряжение действительно большие объемы постоянно изменяющихся данных, сравнительно легко допускающих тематическую фильтрацию.

Мы исходим из того, что организации-генераторы новостной информации в абсолютном большинстве работают в стационарном режиме, который может характеризоваться максимальной производительностью N . Это означает, что каждая организация-генератор производит поток информации, в среднем постоянный по количеству сообщений. Изменяются во времени лишь объемы сообщений, которые соответствуют той или иной

¹ Конечно, идеальной системы такого типа возможно не существует, но, например, в распоряжении авторов находилась система контент-мониторинга InfoStream® (торговая марка ИЦ «ЭЛВИСТИ», Украина), сканирующая около 50 тыс. новостных сообщений в сутки с открытых веб-сайтов RUNeta.

теме. Другими словами, рост количества публикаций по одной теме сопровождается уменьшением публикаций на другие темы [5], так что для каждого промежутка времени T имеем:

$$\int_0^T \sum_{i=1}^M n_i(t) dt = NT,$$

где $n_i(t)$ – количество публикаций в единицу времени, а M – общее количество всех возможных тем. Конечно, предполагается, что некоторая часть $n_i(t)$ равна нулю.

Основной интерес при таком подходе представляет изучение динамики отдельного тематического потока $n_i(t)$.

Конечно, приведенное идеальное уравнение искажается периодическими составляющими, связанными, например, с недельной цикличностью работы компаний – генераторов информации. О механизмах учета таких эффектов речь пойдет ниже.

2.3. Корреляционный анализ тематических информационных потоков

2.3.1. Проблематика

При моделировании тематических информационных потоков необходимо знание особенностей их реального поведения, без учета периодических составляющих, образуемых, например, праздничными датами при рассмотрении информационных потоков за несколько лет или различными объемами публикаций в различные дни недели.

На рис. 1 приведено соотношение количества новостных сообщений, сканируемых системой контент-мониторинга InfoStream [6] в 2006 году, по дням недели (общее количество новостных сообщений превысило 10 млн.).

Известно, что для решения поставленной проблемы могут применяться разнообразные подходы – от простой «компенсации» приведенного процентного соотношения публикаций по дням недели, до методов, базирующихся на преобразованиях Фурье и вейвлет-анализе. Для обнаружения и обоснования последующего устранения периодических составляющих в данной работе были применены подходы, обоснованные в рамках корреляционной теории [7] с последующим применением метода «скользящей средней», которые представляются наиболее обоснованными и реализуемыми с инженерной точки зрения.

2.3.2. Основы корреляционного анализа периодических процессов

Если обозначить через X_t член ряда, образуемого тематическим информационным потоком, (количества документов, поступивших, например, в день t , $t = 1, N$), то функция автокорреляции для этого ряда X определяется как:

$$F(k) = \frac{1}{N-k} \sum_{t=1}^{N-k} (X_{k+t} - m)(X_t - m),$$

где m – среднее значение ряда X , которое в дальнейшем, не ограничивая общности, будем считать равным 0 (достигается присвоением значению X_t значения $X_t - m$). Предполагается, что ряд X может содержать скрытую периодическую составляющую.

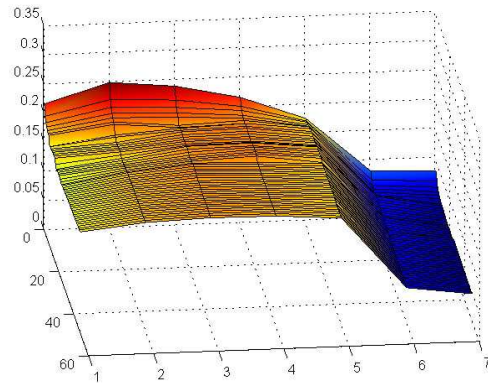


Рис. 1. Недельные колебания объемов информации (ось Z), сканируемых на протяжении 2006 г. (номер недели в году – ось Y) в процентах: Пн – 17.23; Вт – 18.44; Ср – 18.80; Чт – 18.73; Пт – 17.86; Сб – 5.57; Вс – 4.16 (день недели – ось X)

Известно, что функция автокорреляции обладает тем свойством, что если скрытая периодическая составляющая существует, то ее значение асимптотически приближается к квадрату среднего значения исходного ряда, т.е. нулю.

Когда рассматриваемый ряд периодический, т.е. может быть представлен как:

$$X_t = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\alpha t + \theta_n),$$

то его функция автокорреляции будет равна:

$$F(k) = \frac{a_0^2}{4} + \frac{1}{2} \sum_{n=1}^{\infty} a_n^2 \cos n\alpha k.$$

Этот результат показывает, что функция автокорреляции периодического ряда также является периодической, содержит основную частоту и гармоники, но без фазовых углов θ_n .

2.3.3. Корреляционный анализ «зашумленных» периодических процессов

Рассмотрим числовой ряд X , являющийся суммой некоторой содержательной составляющей N и синусоидальной сигнала S :

$$X_t = N_t + S_t.$$

Найдем функцию автокорреляции для этого ряда (значения сведены к нулевому среднему):

$$\begin{aligned} F(k) &= \frac{1}{N-k} \sum_{t=1}^{N-k} X_{k+t} X_t = \\ &= \frac{1}{N-k} \sum_{t=1}^{N-k} (N_{k+t} + S_{k+t})(N_t + S_t) = \\ &= \frac{1}{N-k} \sum_{t=1}^{N-k} N_{k+t} N_t + \frac{1}{N-k} \sum_{t=1}^{N-k} S_{k+t} S_t + \frac{1}{N-k} \sum_{t=1}^{N-k} N_{k+t} S_t + \frac{1}{N-k} \sum_{t=1}^{N-k} S_{k+t} N_t. \end{aligned}$$

Очевидно, первое слагаемое есть функция непериодическая, асимптотически стремящаяся к нулю. Так как N и S не когерентны, то взаимная корреляция между ними отсутствует, поэтому третье и четвертое слагаемое также стремятся к нулю. Таким образом, самый значительный ненулевой вклад составляет второе слагаемое – автокорреляция сигнала S . Т.е. функция автокорреляции ряда X остается периодической.

2.3.4. Корреляционные характеристики реального информационного потока

На рис. 2. представлен график информационного потока сообщений сетевых СМИ по теме «Коррупция в Украине», сформированный системой InfoStream, с помощью которой сканировалось 2500 новостных веб-сайтов. По запросу «(корупц|коррупц) & (Украин|Україн)» было отобрано свыше 83 тыс. публикаций за 456 дней.

Как известно, коэффициенты корреляции для дискретного ряда измерений рассчитываются следующим образом:

$$R(k) = F(k) / \sigma^2,$$

где k – ширина «окна наблюдений»; $F(k)$ — функция автокорреляции; σ^2 — дисперсия X .

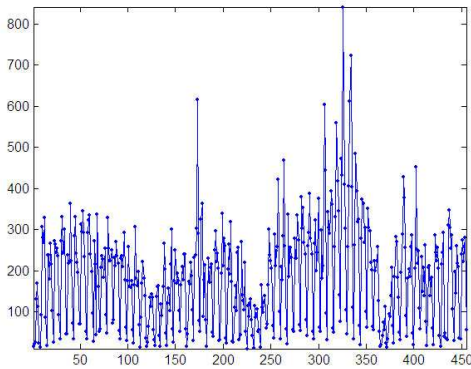


Рис. 2. Количество публикаций по заданной теме (ось Y) по дням (ось X)

Графическое представление коэффициента корреляции для исследуемого ряда наблюдений свидетельствует о разделении корреляционных свойств по дням недели (рис. 3).

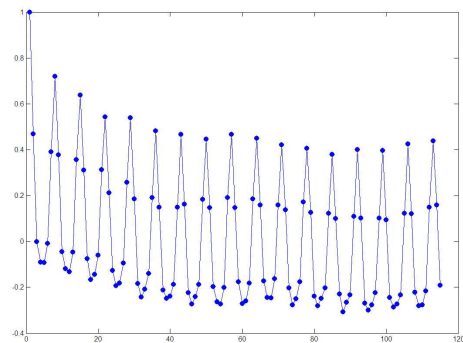


Рис. 3. Коэффициенты корреляции (ось Y) в зависимости от ширины окна наблюдений k (ось X)

Семь уровней значений, представленных на графике соответствуют 7 дням недели. Очевидно, что приведенные выше коэффициенты корреляции представляют собой выбранные значения непрерывной функции, имеющей явно выраженную гармоническую составляющую.

2.3.5. Сглаживание рядов, соответствующих информационным потокам

Для удаления периодической составляющей в исследуемом числовом ряду использовался метод взвешенной «скользящей средней».

В соответствии с результатами, представленными на рис. 5(период оказался равным 7), значения нового ряда «сглаженных» величин, определялись следующим образом:

$$S_i = \frac{1}{7} \sum_{i=t-3}^{t+3} X_i.$$

Очевидно, при изменении i от $t - 3$ до $t + 3$, происходит своеобразное «скольжение» по оси времени. На рис. 4 приведен график «сглаженного» ряда, соответствующего, рассматриваемому исходному.

Корреляционная функция «сглаженного» ряда (рис. 5) не содержит явно выраженных гармоник.

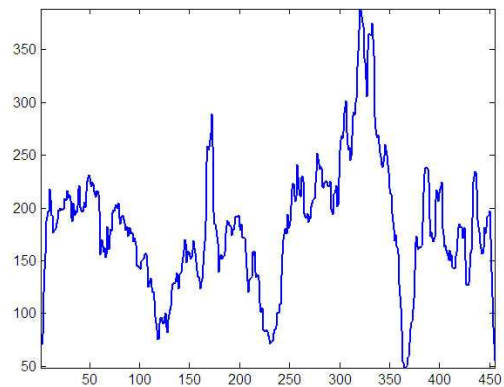


Рис. 4. «Сглаженный» ряд, соответствующий исходному

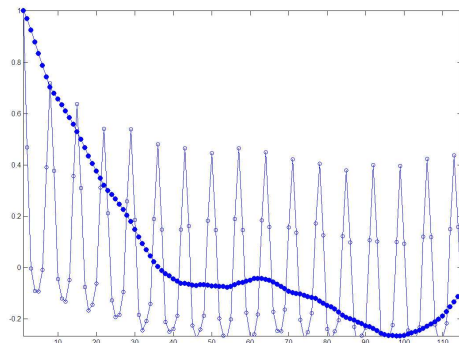


Рис. 5. Коэффициенты корреляции исходного (о) и сглаженного (.) ряда наблюдений

Предполагается, что «сглаживание» гармонич-

ческой составляющей позволит выявить особенности реального поведения тематического информационного потока, его динамику, соответствующую событиям реального мира, без учета периодики отдельных электронных СМИ. Таким образом при дальнейшем экспериментальном исследовании тематических информационных потоков в рамках данной работы исследовались «сглаженные» ряды с окном наблюдения в 7 дней, что позволяет избавляться от явно выявленных недельных циклов в объемах публикаций.

3. Логистическая модель

3.1. Динамика популяций

В отличие модели Бартон-Кеблера, в реальной динамике ТИП имеют место процессы и роста, и спада интенсивности. Поэтому для построения реалистической картины требуется использовать более гибкую модель.

В научной литературе понятие популяции часто используется в расширительном толковании, и поэтому вполне обосновано введение его и в обсуждение информационных потоков. Прежде всего, следует сказать, что документы в информационном потоке во многих отношениях напоминают популяции живых организмов. Они в определенном смысле «рождаются», «умирают» и дают «потомство».

Во второй половине XX-го века были достигнуты значительные успехи в построении математических моделей динамики популяций [8, 9]. Наиболее перспективной, по-видимому, следует считать логистическую модель (ЛМ), предложенную П. Ферхюльстом для описания динамики народонаселения [10] и Р. Перлом для биологических сообществ [11]. В дальнейшем она оказалась крайне плодотворной во многих областях науки и техники. Логистическую модель можно рассматривать как обобщение модели Мальтуса [12], которая предусматривает пропорциональность скорости роста функции ее значению в каждый момент времени:

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

где k – некоторый коэффициент пропорциональности.

Идея ЛМ заключается в том, чтобы сделать коэффициент в уравнении Мальтуса функцией времени, причем так, чтобы решение не превышало заданного порогового значения. С этой целью используем параметр насыщения N . Тогда правую часть приведенного выше уравнения можно представить в виде: $k(N - n(t))$. Соответственно, модифицированное уравнение примет вид:

$$\frac{dn(t)}{dt} = kn(t)[N - n(t)] = k[Nn(t) - n^2(t)];$$

$$n(0) = n_0.$$

Иногда в этом уравнении удобнее перейти к относительной численности:

$$m(t) = \frac{1}{N}n(t);$$

$$p = kN.$$

Тогда:

$$\frac{dm(t)}{dt} = pm(t)[1 - m(t)].$$

Это уравнение имеет два класса решений (рис. 6, 7). Первый из них описывает рост $m(t)$, а другой – спад. Следовательно, система, описываемая уравнением, имеет два равновесных состояния: $m(t) = 0$ и $m(t) = 1$. Из них первое является неустойчивым, а второе – устойчивым. Действительно, при отклонении системы от равновесия $m(t) > 1$ включается механизм спада, а при отклонении $m(t) < 1$, соответственно, механизм роста.

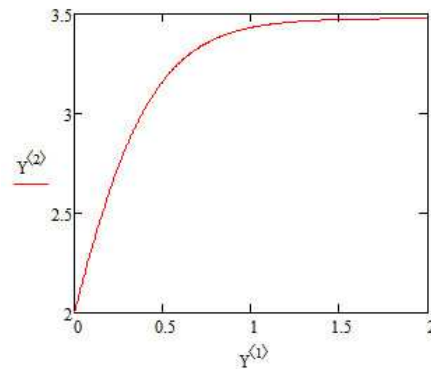


Рис. 6. Рост $m(t)$, ось Y , от времени (ось X)

Следовательно, данное уравнение описывает поведение системы, которое неизбежно приведет к установлению устойчивого равновесия $m(t) = 1$.

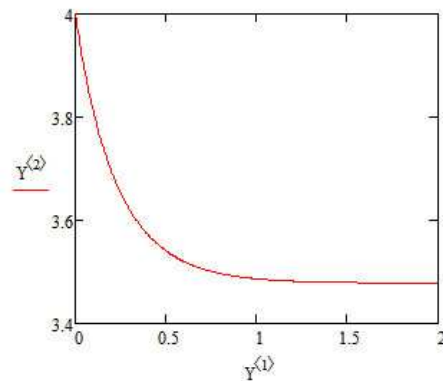


Рис. 7. Спад $m(t)$, ось Y , от времени (ось X)

Таким образом, стандартная ЛМ вполне пригодна для описания нелинейных процессов в сложных динамических системах, однако ограничивается лишь процессом установления устойчивого равновесия. Для более общих случаев это уравнение следует модифицировать.

3.2. Взаимодействие популяций

Приведенное выше логистическое уравнение описывает динамику единичной популяции, которая

взаимодействует исключительно с окружающей средой. В реальности подобные ситуации возникают редко, поскольку популяции активно взаимодействуют между собой. В теории популяционной динамики разработана классификация различных форм такого взаимодействия [13, 14], так что общая картина взаимоотношения популяций выглядит достаточно сложной и разнообразной. При этом следует учитывать, что взаимодействие популяций может быть не только прямым, но и опосредованным.

В рамках логистической модели описание n взаимодействующих популяций в общем случае осуществляется с помощью следующей системы уравнений:

$$\frac{dm_i(t)}{dt} = m_i(t) \left[p_i - \sum_{j=1}^n q_{ij} m_j(t) \right];$$

$$m_i(0) = m_{0i}.$$

Соответственно, тип описываемого процесса определяется величиной и знаком коэффициентов p_i и q_{ij} , причем следует иметь в виду, что в каждом уравнении диагональные члены $m_i m_i$ соответствуют внутривидовому, а перекрестные $m_i m_j$ – межвидовому взаимодействию. Данная система уравнений в принципе может описывать широкий спектр зависимостей, однако ее решения (относящиеся к реальным процессам), соответствуют одному из следующих режимов:

- стационарный;
- автоколебательный;
- квазистохастический.

3.3. Обобщение логистической модели

Приведенное выше описание динамики популяций в рамках ЛМ может быть распространено ТИП. Для того, чтобы обеспечить возможность более адекватного описания динамики информационных потоков, необходимо несколько модифицировать приведенные выше уравнения, определяя в них коэффициенты как функции времени.

Прежде всего, введем зависящий от времени параметр, представляющий собой меру влияния событий внешней среды на число сообщений, относящихся к некоторой теме:

$$k = k(t) = k_0 + R(t).$$

где k_0 – константа, описывающая генерацию фоновых сообщений, а $R(t)$ – мера явного влияния событий.

Также потребуется учесть и то обстоятельство, что на самом деле процесс генерации сообщений не обязательно достигает насыщения при росте их числа до максимально допустимого значения N . Поэтому вводится в рассмотрение функцию $S(t)$, описывающая зависимость критического числа сообщений от времени.

С учетом сказанного, обобщенное логистическое уравнение примет следующий вид:

$$\frac{dn(t)}{dt} = [k_0 + R(t)]n(t)[N - R(t)n(t)];$$

$$n(0) = n_0.$$

Нормируя на N и, переобозначив коэффициенты, получаем более компактную запись:

$$\frac{dm(t)}{dt} = p(t)m(t) - q(t)m^2(t),$$

где

$$m(t) = \frac{1}{N} n(t);$$

$$p(t) = N[k_0 + R(t) = P_0 + P(t)];$$

$$q(t) = p(t)S(t).$$

В случае двух и более взаимодействующих тем ситуация выглядит существенно сложнее, поскольку коэффициенты в системе уравнений обладают менее очевидной интерпретацией. По аналогии естественно было бы ввести функции времени $p(t)$ и $q(t)$, стоящие соответственно перед линейным и диагональным членами, с тем же «физическим» смыслом:

$$\frac{dm_i(t)}{dt} = m_i(t) \left[p_i(t) - \sum_{j=1}^n c_{ij} m_j(t) - q_i(t) m_i(t) \right]$$

Процессы генерации сообщений на различные темы, как правило, не являются независимыми. В определенном смысле можно говорить о поглощении менее общих сообщений более общими, особенно в случаях генерализации событий. Не исключено, что могут иметь место и другие механизмы взаимовлияния, не столь очевидные.

В данной работе рассмотрены два общих случая, а именно, динамика одной невзаимодействующей темы и динамика двух взаимодействующих тем.

В каждом из них нам придется определять явный вид функций $p(t)$ и $q(t)$ для данной конкретной задачи, исходя из неких разумных допущений.

Из приведенных выше соотношений видно, что одной из особенностей ЛМ является требование $n_0 > 0$. Иными словами, логистическое уравнение не может описывать зарождение популяций. Оно пригодно исключительно для описания именно их динамики.

4. Монотематическая динамика

Наиболее простой случай временной зависимости числа сообщений, поступающих в информационный поток в связи с некоторым событием. В этом случае кривая динамики ТИП выглядит просто: вначале оно резко возрастает, достигает насыщения, а затем убывает, стремясь при этом к некоторому значению, близкому к нулю.

Аналогичная задача рассматривалась авторами ранее в работе [15], где в качестве $p(t)$ использовалась ступенчатая функция:

$$P(t) = \begin{cases} D, 0 \leq t \leq \lambda \\ 0, t > \lambda \end{cases}$$

Соответственно, для двух областей ($t \leq \lambda$ и $t > \lambda$) были получены уравнения:

$$\frac{du(t)}{dt} = P_0 u(t) - Qu^2(t) + Du(t);$$

$$\frac{dv(t-\lambda)}{dt} = P_0 v(t-\lambda) - Qv^2(t-\lambda),$$

где P_0 и Q – константы.

Полное решение этой системы определялось из условия «сшивки» на границе областей:

$$u(\lambda) = v(\lambda).$$

Полученные зависимости в целом совпали с опытными данными, относящимися к сообщениям определенной категории. Однако были выявлены и другие виды тематических информационных потоков, для которых данная модель оказалась неприемлемой.

Соответственно, вместо прямоугольной ступеньки используется гладкая функция, которая также будет иметь эффективную ширину, равную λ . Для реализации этого подхода было выбрано распределение Гаусса:

$$R(t) = ae^{-b(t-\tau)^2}.$$

Параметр τ фиксирует положение максимума временной зависимости реакции (в смысле роста числа публикаций) на происшедшее событие.

Вместе с тем, типичная качественная зависимость монотематического информационного потока от времени приведена на рис. 8.

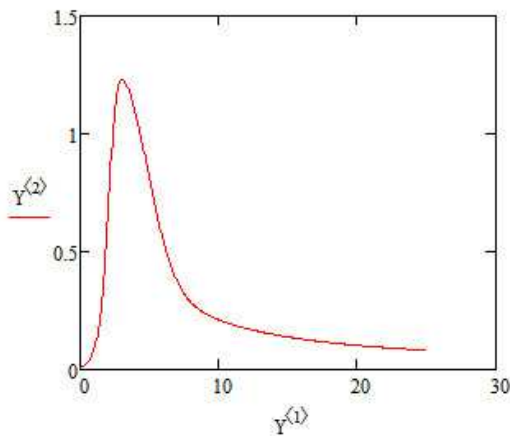


Рис. 8. Качественная зависимость монотематического информационного потока (ось Y) от времени (ось X)

Мы видим, что изображенная на нем кривая существенно отличается от стандартной гауссианы. Следовательно, вид функции $R(t)$ лишь частично определяет результирующую форму интересующей нас зависимости.

При выборе функции $q(t)$, которая описывает эффект изменения числа публикаций, вызванного изменением эффективного объема доступных

ресурсов, будем исходить из того, что он может проявляться в двух вариантах, отличающихся преобладанием того или иного типа обратных связей: самовозбуждения и самозатухания. В первом случае тема уже после завершения породивших ее событий становится все более и более актуальной, во втором же ее актуальность постоянно падает.

В общем виде выражение для $q(t)$ выберем следующим образом:

$$q(t) = (c_0 + ct)^h + d[1 + \sin(\omega t + \phi)],$$

$$h = \pm 1.$$

Первый член в этом выражении описывает в зависимости от знака h уменьшение ($h = 1$) или увеличение ($h = -1$) доступной области ресурсов, а второй – периодические ее изменения. Типичные зависимости приведены на рис. 9 и 10.

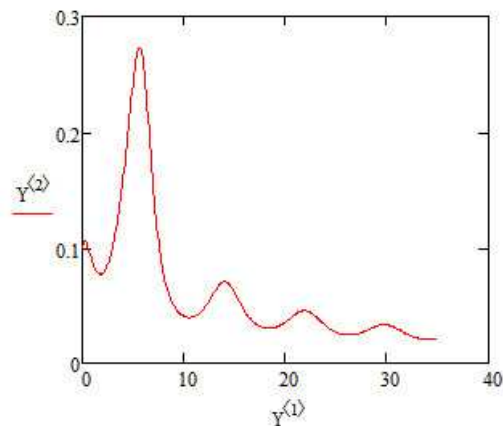


Рис. 9. Случай $h = 1$

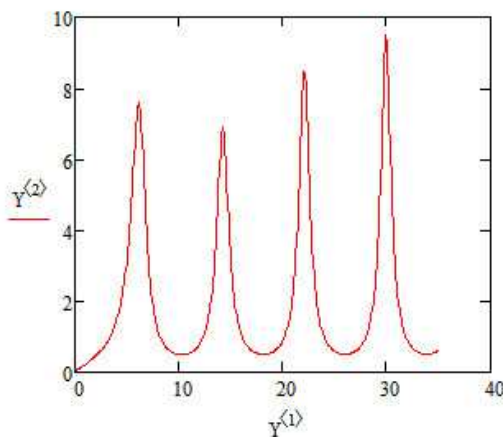


Рис. 10. Случай $h = -1$

Приведенные соотношения оказываются достаточно гибкими, чтобы описать общее поведение временных зависимостей числа публикаций.

Приведем в качестве примеров две зависимости, полученные при анализе реальных информационных потоков.

На рис. 11. приведены значения информационного потока по теме «отравление А. Литвиненко» за ноябрь-декабрь 2006 г. На рис.

12 приведены результаты моделирования данного информационного потока.

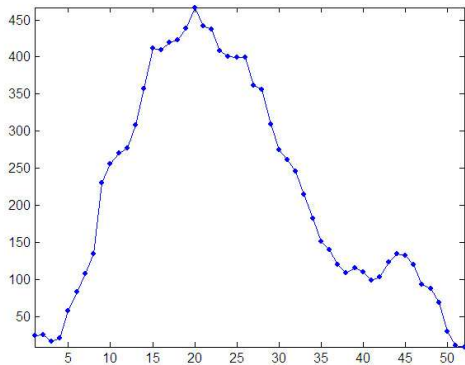


Рис. 11. Информационный поток по теме «отравление А. Литвиненко»

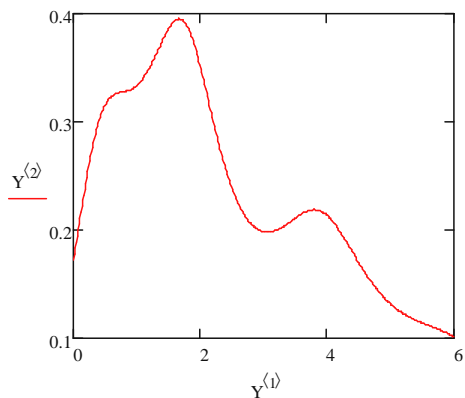


Рис. 12. Результат моделирования

На рис. 13. приведены значения информационного потока по теме «Хизбалла» за октябрь-декабрь 2006 г., а на рис. 14 - результаты моделирования данного информационного потока. В данном случае обращает на себя внимание явная цикличность рассматриваемого ряда. При этом цикличность, связанная с днями недели была заранее «сглажена» (см. п. 2.3.5).

5. Динамика взаимодействующих тем

5.1. Стандартная логистическая модель

5.1.1. Простейшие формы взаимодействия тем

Изучение динамики в случае взаимодействия тем осложняется тем обстоятельством, что реальные ТИП содержат множество зависимостей, в отношении которых трудно сказать, какие из них взаимодействуют преимущественно между собой. Более того, трудно с полной определенностью отнести отдельные потоки к взаимодействующим либо невзаимодействующим.

В данной части работы основное внимание будет уделено в основном собственно анализу модели. Основные типы взаимодействий, описываемых системами логистических уравнений хорошо известны и включают несколько характерных

форм. В качестве примера опишем две наиболее интересные конкуренцию и симбиоз.

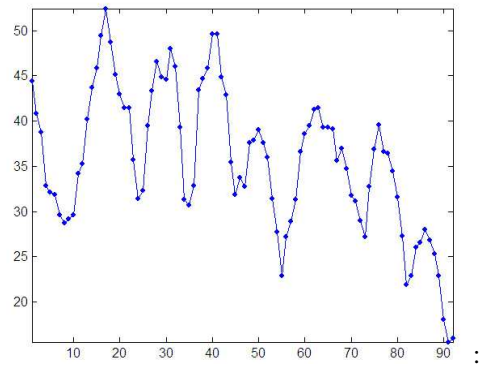


Рис. 13. Информационный поток по теме «Хизбалла»

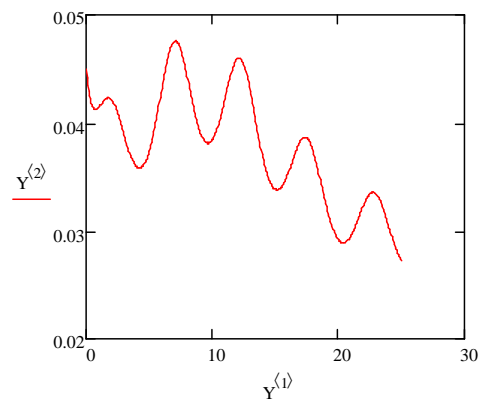


Рис. 14. Результат моделирования

5.1.2. Конкуренция

Случаю конкуренции соответствуют положительные значения обоих перекрестных коэффициентов q_{ij} . Это означает, что взаимодействие тем происходит таким образом, что рост числа публикаций по одной из них сопровождается сокращением числа публикаций по другой. На рис. 15 приведен два характерный случай конкуренции тем.

Интересным случаем конкуренции является автоколебательный режим, в котором числа публикаций совершают незатухающие колебания. Он может возникать при нулевых значениях диагональных коэффициентов q_{ii} .

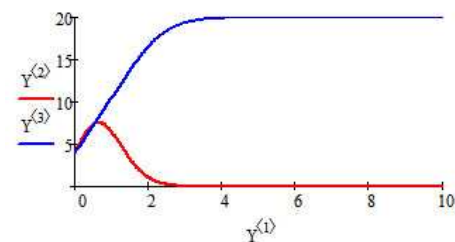


Рис. 15. Конкуренция

5.1.3. Симбиоз

Симбиоз возникает при отрицательных значениях

коэффициентов p_2 и q_{21} , то есть при условиях, когда тематические потоки не только потребляют определенные ресурсы, но и «подпитывают» друг друга. Пример приведен на рис. 16.

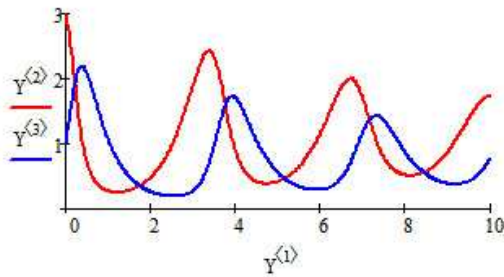


Рис. 16. Симбиоз

5.2. Обобщенная логистическая модель

В рамках обобщенной логистической модели, можно проследить за тем, как темы воздействуют друг на друга. Для этого вначале строится зависимость для двух отдельных тем, эволюционирующих каждая по закону, определяемому функциями $p(t)$ и $q(t)$ (рис. 17), а затем рассматривается их совместная динамика, при условии, что соответствующие законы остались неизменными (рис. 18).

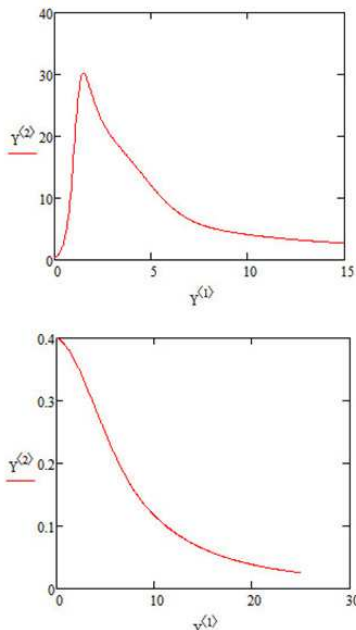


Рис. 17. Раздельная эволюция тем

На рис. 18 видно, что взаимное влияние тем носит не только количественный, но и качественный характер. Поведение каждой кривой теперь, действительно, определяется не только собственными функциями $p(t)$ и $q(t)$, но и динамикой другой темы.

Приведем в качестве примеров две взаимные зависимости реальных ТИП.

На рис. 19. приведены совместные значения информационных потоков по темам Nokia и Motorola за ноябрь-декабрь 2006 г, а на рис. 20 – результат моделирования этого процесса. На рис. 21

приведены совместные ТИП, определяемые двумя личностями – Ахметовым и Богатыревой в контексте украинской Партии Регионов. Соответственно, на рис. 22 приведены результаты моделирования.

Как мы видим, оба приведенных случая отражают состояние конкуренции информационных потоков.

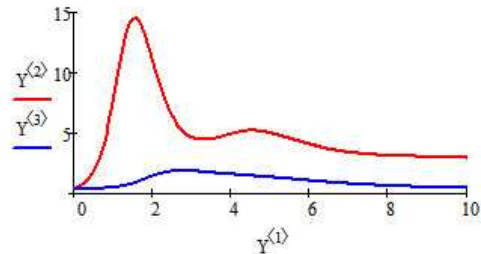


Рис. 18. Совместная динамика тем

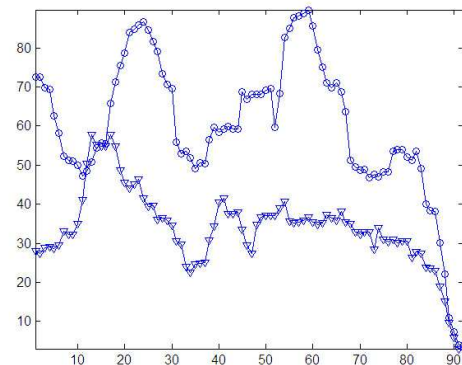


Рис. 19. Потоки по темам Nokia (∇) и Motorola (\circ)

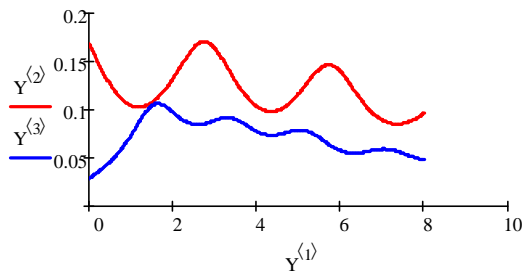


Рис. 20. Результаты моделирования

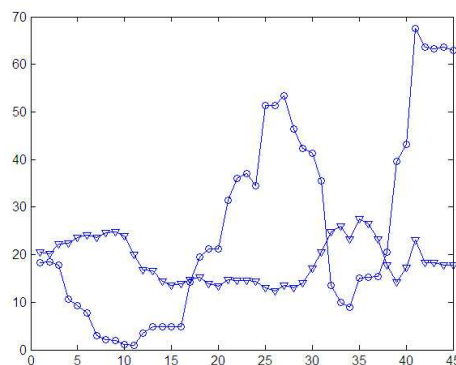


Рис. 21. Потоки по персонам – Ахметов (∇) и Богатырева (\circ)

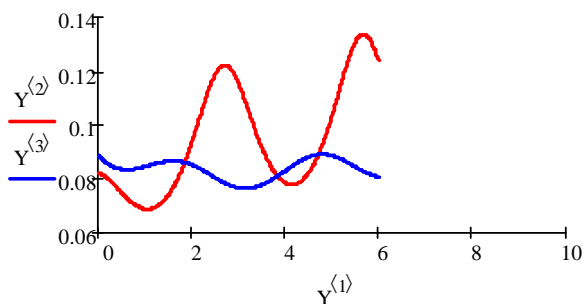


Рис. 22. Результаты моделирования

6. Объектная визуализация тематических информационных потоков, Wordlet-диаграмма

Предполагается, что часть проблем, связанных с ориентацией пользователей в информационных ресурсах Интернет решится за счет средств визуализации динамики ТИП. Вопросы визуализации результатов поиска посвящено большое количество научных работ [16-18]. Вместе с тем, визуализации тенденций и объектного распределения ТИП больших объемов не уделяется существенного внимания.

Авторами предлагается форма визуального отображения информационного потока в разрезе объектов и дат, представляющая собой прямоугольную таблицу, будем называть ее Wordlet-диаграммой, ячейки которой заполнены значениями количества сообщений ТИП за определенную дату, соответствующих определенному объекту. Столбцам этой таблицы соответствуют даты, а строкам – объекты, являющиеся своеобразными содержательными фильтрами информационного потока. Объектам могут соответствовать информационные запросы. Визуально Wordlet-диаграмма представляет собой таблицу, ячейки которой закрашены оттенками серого цвета, в зависимости от значений объемов публикации по выбранному объекту. Wordlet-диаграммы позволяют визуально выявлять группы наиболее связанных по датам и интенсивностям публикаций объектов. Для большого количества объектов предлагается кластеризация Wordlet-диаграммы в соответствии с алгоритмом k-means [19].

В рамках данной работы в качестве объектов, рассматривались персоны, упоминаемые в текстах сообщений. На рис. 23 представлен график информационного потока по теме «Выборы в Украине». Система контент-мониторинга InfoStream на основании анализа свыше 2500 источников информации в сети Интернет позволила построить зависимость суточных объемов тематических публикаций за 3 года (1096 суток, свыше 320 тысяч сообщений). На рис. 24 представлена Wordlet-диаграмма, объектами для которой выбраны персоны – участники избирательных процедур.

В результате проведенных экспериментов, есть основания предположить, что использование таких средств визуализации, как Wordlet-диаграммы,

позволяет «разлагать» исходные временные ряды в соответствии с объектами, выявлять взаимосвязи объектов в разрезе дат.

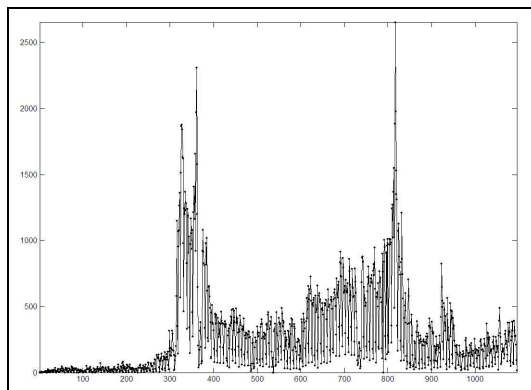


Рис. 23. ТИП «Выборы в Украине»

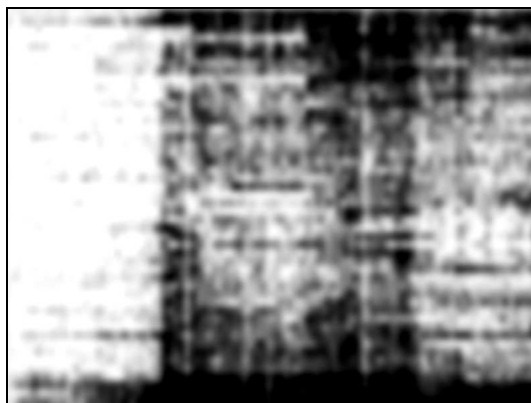


Рис. 24. Пример Wordlet-диаграммы. Ось X – дни выбранного периода, ось Y – объекты

7. Заключение

Итак, построена достаточно компактная модель (с небольшим числом параметров), которая в некотором приближении позволяет описывать временную зависимость числа документов, соответствующих определенной теме, в сетевых документальных потоках.

Модель в том виде, в котором она приведена в данной работе, пригодна для описания общих тенденций в динамике потоков. Полная картина может быть получена с учетом дополнительного набора факторов, большинство которых являются случайными и потому не воспроизводятся во времени. В какой мере необходим их явный учет – зависит в первую очередь от поставленной задачи. Структура уравнений, лежащих в основе модели, позволяет вносить соответствующие коррективы, например моделировать случайные отклонения.

В связи со сказанным отметим также, что воспроизведение результатов во времени, вообще говоря, является в данном случае крайне серьезной проблемой. В динамике сетевых информационных потоков точная повторяемость динамики если и

встречается, то крайне редко. Поэтому в нашем распоряжении в данный момент не оказалось надежного способа верификации результатов. По-видимому, дальнейшие исследования внесут в этот вопрос большую ясность.

8. Литература

- [1] Gianna M. Del Corso, Antonio Gullí, Francesco Romani. Ranking a stream of news. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. - P. 97 - 106.
- [2] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1. Вып. 11. – 2005. - С. 21-33.
- [3] Burton R.E. and Kebler R.W. The "half-life" of some scientific and technical literatures. American Documentation 1960;1:98—109.
- [4] Чурсин Н.Н. Популярная информатика. -К.: Техника, 1982. -158 с.
- [5] Ландэ Д.В. Основы интеграции информационных потоков - К.: Инжиниринг, 2006. - 240 с. (<http://dwl.kiev.ua/art/monogr-osnov/spusk3.pdf>)
- [6] Григорьев А.Н., Ландэ Д.В. и др. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис. – К.: «Старт-98», 2007. – 40 с. (<http://dwl.kiev.ua/art/booklet/booklet.pdf>)
- [7] Y.W. Lee, T.P. Cheatham, J.B. Wiesner, Application of correlation analysis to the detection of periodic signals in noise, PJRE 38, 1165, 1950.
- [8] Вольтерра В. Математическая теория борьбы за существование - М.: Наука, 1976.
- [9] В.И. Арнольд. Аналитика и прогнозирование: математический аспект. // Научно-техническая информация. Сер. 1. Вып. 3. - 2003. - С. 1-10.
- [10] Verhulst P.F. Notice sur la loi que la population suit dans son accroissement Corr. Math. Et Phys. 10, 113-121, 18.
- [11] Pearl R. The Introduction to Medical Biometry and Statistics. Philadelphia, 1930; Ibid. The Natural History of Population. L., 1939.
- [12] Malthus T.R. An essay on the principal of Population, as it affects the future improvement of society. - 1798 (<http://etext.lib.virginia.edu/toc/modeng/public/MalPopu.html>).
- [13] Гаузе Г. Ф. Борьба за существование. – М: УРСС, 2002. - 160 с.
- [14] Гаузе Г. Ф. Экология и некоторые проблемы происхождения видов. В кн.: Экология и эволюционная теория - Л.: Наука, 1984, с. 5–108.
- [15] Ландэ Д.В., Фурашев В.Н., Брайчевский С.М., Григорьев А.Н. Основы моделирования и оценки электронных информационных потоков - К.: Инжиниринг,

2006. – 176 с. (<http://dwl.visti.net/art/inf-potok/inf-potok.pdf>)

- [16] M.M. Knepper, R. Killam, K.L. Fox O. Frieder. Information Retrieval and Visualization using SENTINEL / TREC 1998: 336-340.
- [17] Ландэ Д.В. Присмотритесь внимательно или "Изюминки" поисковой визуализации // hiTech Pro - К., 2006. - декабрь, - С. 94-95. (<http://dwl.kiev.ua/art/hitech/>)
- [18] Григорьев А.Н., Ландэ Д.В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Труды Международного семинара «Диалог'2005». – М.: Наука, 2005. – С. 109-111. (<http://dwl.kiev.ua/art/dialog/>)
- [19] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.

Modelling of dynamics of textual news streams

Lande, D.V., Snarskii, A.A.,
Brajchevskiy, S.M., Darmokhval, A.T.

In behaviour of the text information streams , generated on network of the Internet, two tendencies: permanent growth of volumes and a complication of a dynamic structure are observed. In connection with it the problem of modeling of dynamics of the informational streams becomes topical. The given work has been dedicated to just this question.

In the given work results of modelling of frequency characteristics of the thematic information streams, executed are resulted within the framework of the generalized logistical model.

Are submitted both theoretical conclusions, and results of the experimental analysis of dynamics of the information streams, processable within the framework of content-monitoring technology InfoStream.