УДК 004.738

МОДЕЛИРОВАНИЕ ДИНАМИКИ ИНФОРМАЦИОННЫХ ПОТОКОВ

Ландэ Д.В.

Институт проблем регистрации информации НАН Украины, Киев, e-mail: dwl@visti.net

Информационное пространство в статье рассматривается как множество связанных по смыслу единиц контента (документов). В динамике их эволюции образуются информационные потоки. Рост объемов и усложнение динамики информационных потоков определяют актуальность задачи. В статье приводится формальное определение информационных потоков. Определено также понятие тематического информационного потока. Классические методы агрегации информации не всегда способны адекватно отражать состояние динамической составляющей информационного пространства. Для исследования современных информационных потоков все чаще применяются новые подходы, основанные на методах нелинейной динамики. В работе описана логистическая модель взаимодействия информационных потоков. Структура уравнений, лежащих в основе логистической модели, является достаточно общей и позволяет моделировать эффекты конкуренции и симбиоза, а также случайные отклонения. К недостаткам моделирования можно отнести проблематичность надежной верификации результатов.

Ключевые слова: информационные потоки, моделирование, логистическая модель, информационное пространство, нелинейная динамика

MODELING THE DYNAMICS OF INFORMATION FLOWS

Lande D.V.

The Institute for Information Recording the National Academy of Sciences of Ukraine, Kiev, e-mail: dwl@visti.net

Information space in the article we consider as the set of related units of content (documents). In dynamics of their evolution formed information flows. Growth in the volume and complexity of the dynamics of information flows determine the urgency of the problem. The article provides a formal definition of information flows. Also defined the concept of thematic information flow. Classical methods of aggregation of information is not always able to adequately reflect the condition of the dynamic component of the information space. To research of modern information flows it is even more often applied the new approaches based on methods of nonlinear dynamics. This paper describes the logistic model the interactions of information flows. The structure of the equations underlying the logistic model, is quite general and allows to simulate the effects of a competition and symbiosis, and also casual deviations. The disadvantages include the difficulty of modeling a reliable verification of results.

Keywords: information flows, modeling, logistic model, information space, nonlinear dynamics

Информационное пространство можно рассматривать как множество связанных по смыслу элементов (документов), образующих в динамике своей эволюции информационные потоки [1]. При этом многолетние наблюдения свидетельствуют о том, что информационное пространство обладает устойчивыми закономерностями, в частности, показано, что параметры частотного и рангового распределений документов во многих информационных потоках остаются одинаковыми и определяются параметрами, зависящими от содержания, тематики информации [2].

Для исследования современных информационных потоков все чаще применяются новые подходы, потому что классические методы и средства агрегации информационных массивов не всегда способны адекватно отражать состояние динамической составляющей информационного пространства.

Для моделирования информационных потоков, с одной стороны, вполне подходит классическая теория информации, которую можно трактовать как математическую теорию связи, разработанную К. Шенноном [3] в 40-х годах XX столетия и существенно дополненную и расширенную в последующие годы работами Н. Винера, В.А. Котельникова и А.Н. Колмогорова. В этих работах рассматривались количественные оценки, относящиеся к передаваемой информации, было определено «количество информации». Однако сегодня понятна ограниченность такого подхода, невозможность разрешения реальных проблем, связанных с содержательной составляющей информации. Значительный вклад в исследования в области теории информации вносит нелинейная динамика, синергетика [4, 5].

Для строгости дальнейшего изложения дадим определение информационного потока, которое корреспондируется с классическим определением. Не принимая во внимание линии передачи данных, потоки данных между серверами, клиентами и т.п., остановимся лишь на факте размещения информации в информационном пространстве. Введем для этого понятие «идеального сканера», обеспечивающего считывание любого документа (будем для единообразия использовать этот термин для обозначения единицы контента, понимая его как синоним терминов «публикация, сообщение» и т.п.) в момент его помещения в информационное пространство (к таким сканерам сегодня в веб-пространстве, как фрагменте

информационного пространства, все более приближаются роботы промышленных поисковых систем типа Google).

Рассмотрим отрезок (a, τ) действительной оси (оси времени), где $\tau > a$. Допустим, что на этом отрезке времени в соответствии с некоторыми закономерностями в сети «идеальным сканером» считывается некоторое количество документов – k. На оси времени моменты публикации отдельных документов обозначим как $\tau_1, \tau_2, ..., \tau_k$ $(a \le \tau_1 \le \tau_2 \le ... \le \tau_k \le \tau)$. Информационным потоком будем называть процесс $N_a(\tau)$, реализация которого характеризуется количеством документов, сосканированным в интервале (a, τ), как функцию правого конца отрезка τ. В соответствии с этим определением реализация информационного потока является неубывающей ступенчатой всегда целочисленной функцией $N_a(\tau)$.

Приведеное определение на локальных временных областях соответствует действительности, но не учитывает такой эффект, как старение информации, которое противоречит «накопительной» способности информационного потока $N_a(\tau)$ на больших промежутках времени. Этот недостаток можно компенсировать, введя дополнительные поправки, базирующиеся на модели старения информации Бартона-Кеблера [6].

Такой подход позволяет рассматривать информационные потоки как временные ряды; учитывая то, что отдельные документы из информационных потоков отражают процессы, происходящие в реальном мире, что дает возможность прогнозировать их динамику, выявлять скрытые корреляции, циклы. Сегодня для решения названных задач все чаще применяются корреляционный, дисперсионный, фрактальный, вейвлет-анализ временных рядов.

Основным объектом современного моделирования информационных потоков являются тематические информационные потоки, последовательности документов, соответствующих определенной тематике. Многочисленные факты свидетельствуют о том, что в действительности динамика тематических информационных потоков определяется комплексом внутренних нелинейных механизмов, которые, как правило, коррелируют с реальностью.

Количество документов в общем информационном потоке, состоящем из тематических потоков, является величиной относительно стабильной. Изменяются во времени лишь объемы потоков, соответствующих той или иной тематике, тому или иному информационному сюжету. Другими словами, увеличение количества документов по одной теме сопровождается уменьшением документов по другим темам, так что для каждого промежутка времени T имеем [6]:

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

где $n_i(t)$ – количество документов в единицу времени по теме i, а M – общее количество всех возможных тем. Таким образом для локальных временных промежутков можно наблюдать так называемый «тематический баланс». Основной интерес при этом представляет изучение динамики отдельного тематического потока, который описывается плотностью $n_i(t)$. При этом общие политематические потоки являются стационарными по количеству документов, динамика же в основном определяется «конкурентной борьбой» отдельных тематик.

Еще сложнее выглядит синхронное изменение количества документов, относящихся к нескольким тематическим информационным потокам. Их поведение четко напоминает процессы взаимодействия популяций в биоценозе. Так, например, в ряде случаев увеличение числа документов по одной теме сопровождается сокращением числа документов по другим темам. Общая динамика в этом случае может описываться системой уравнений, каждое из которых относится к отдельному монотематическому потоку.

Вместе с тем в практическом плане часто оказывается полностью удовлетворительным упрощеное понимание информационного потока как некоторой зависимой от времени величины $n(t)$, которая описывается уравнением:

$$\frac{dn(t)}{dt} = F\left(n(t), t\right).$$

В самом простом виде такие уравнения могут иметь следующий вид:

$$\frac{dn_i(t)}{dt} = p_i \cdot n_i(t) - \sum_{j=1}^{N} r_{ij} \cdot n_i(t) \cdot n_j(t),$$

где N – количество тематик; p_i – вероятность появления в единицу времени публикации по теме i, r_{ij} – коэффициент взаимосвязи тематик i и j.

Классические модели информационных потоков, линейные и экспоненциальные, мало пригодны для изучения реальной динамики сетевых информационных потоков в течение длительных интервалов времени. Как обобщение экспоненциальной модели, предусматривающей пропорциональность скорости роста функции $n(t)$ в каждый момент времени ее значению, можно рассмаривать логистическую модель. Главная идея логистической модели заключается в том, что для ограничения скорости роста на функцию $n(t)$ накладывается дополнительное условие, в соответствии с которым

ее значением не должно превышать некоторую величину. Для этого выберется множитель $k(t)$ такого вида:

$$k(t) = k \cdot [P - rn(t)],$$

где P – некоторое предельное значение, которое функция $n(t)$ не может превышать ($rn_0 \le P$)(); r – коэффициент, описывающий негативные для данной тенденции процессы; k – коэффициент пропорциональности. В результате получаем логистическое уравнение:

$$\begin{cases} \dfrac{dn(t)}{d(t)} = kn(t)\left[P - rn(t)\right], \\ n(t_0) = n_0. \end{cases}$$

Приведенное уравнение можно считать феноменологическим: исследователям не обязательно знать, как действуют конкретные механизмы, по мере роста $n(t)$ снижающие скорость ее изменения.

В случае информационных потоков, которые ассоциируются с конкретными темами, 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 документов по всем тематикам распространяется и на совокупность всех монотематических потоков. В случае изучения общего информационного потока наблюдается явление «перетекания» документов из одних тематик, в другие, более актуальные.

Общая динамика должна описываться системой уравнений, каждое из которых относится к отдельному монотематическому потоку. Приведенную выше систему уравнений «конкурентной борьбы» в рамках обобщенной логистической модели можно представить в таком виде:

$$\frac{dn_i(t)}{dt} = \left(p_i + D_i(t)\right) \cdot \left(n_i(t) - \sum_j r_{ij} \cdot n_i(t) n_j(t) \right),$$

где $D_j(t)$ – параметр актуальности темы.

Изучение взаимодействия тем является достаточно сложной задачей, так как на практике тематические информационные потоки охватывают большое количество зависимостей, уровень взаимозависимостей которых зачастую неизвестен. Если же говорить о системе логистических уравнений, то в рамках данной модели доминируют две основные темы взаимодействия – конкуренция и симбиоз. Конкуренции соответствуют положительные значения коэффициентов r_{ij}, соответсвующих i-й и j-й темам, т.е. взаимодействие происходит таким образом, что увеличение количества документов по одной из тем приводит к сокращениею второго информационного потока. Симбиоз возникает при отрицательных значениях коэффициентов r_{ij}, т. е. при условиях, когда тематические потоки не только потребляют определенные ресурсы, но и «подпитывают» друг друга.

Структура приведенных выше уравнений (лежащих в основе логистической модели) является достаточно общей и, например, позволяет моделировать случайные отклонения. К недостаткам такого моделирования можно отнести тот факт, что воспроизведение результатов (т.е. надежная верификация результатов) в данном случае является очень проблематичным.

Вместе с тем развитие методов математического моделирования, так называемого «мягкого моделирования» [5], в котором модели строятся, опираясь не на строгие количественные законы, а на качественные закономерности, позволили подойти к новой точке зрения в области исследования информационных потоков, что позволяет корректно использовать методы нелинейной динамики, теорий клеточных автоматов, перколяции, самоорганизованной критичности [8].

Список литературы

1. Додонов А. Г., Ландэ Д.В. Живучесть информационных систем. – Киев : Наук. думка, 2011. – 256 с.

2. Иванов С.А. Стохастические фракталы в Информатике // Научно-техническая информация. Сер. 2. – 2002. – № 8. – С. 7–18.

3. Шеннон К. Работы по теории информации и кибернетике. – М.: Изд. иностр. лит., 1963. – 830 с.

4. Хакен Г. Информация и самоорганизация. Макроскопический подход к сложным системам. 2-е изд., доп. – М.: Либроком (Editorial URSS), 2005. – 248 с.

5. Арнольд В.И. Аналитика и прогнозирование: математический аспект // Научно-техническая информация. – Сер. 1. – Вып. 3. – 2003. – С. 1–10.

6. Ландэ Д.В. Основы интеграции информационных потоков. – Киев: Инжиниринг, 2006. – 240 с.

7. 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.

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

References:

1. Dodonov A.G., Landeh D.V. Zhivuchest' informacionnykh sistem. K.: Nauk. dumka, 2011. 256 p.

2. Ivanov S.A. Stokhasticheskie fraktaly v Informatike // Nauchno-tekhnicheskaja informacija. Ser. 2. 2002. no. 8. pp. 7–18.

3. Shennon K. Raboty po teorii informacii i kibernetike. M.: Izd. inostr. lit., 1963. 830 p.

4. Khaken G. Informacija i samoorganizacija. Makroskopicheskijj podkhod k slozhnym sistemam. Izd.2, dop. – M.: Librokom (Editorial URSS), 2005. 248 p.

5. Arnol'd V.I. Analitika i prognozirovanie: matematicheskijj aspekt // Nauchno-tekhnicheskaja informacija. Ser. 1. Vyp. 3. 003. p. 1–10.

6. Landeh D.V. Osnovy integracii informacionnykh potokov. K.: Inzhiniring, 2006. 240 p.

7. Landeh D.V., Snarskijj A.A., Brajjchevskijj S.M., Darmokhval A.T. Modelirovanie dinamiki novostnykh tekstovykh potokov // Internet-matematika 2007: Sbornik rabot uchastnikov konkursa. – Ekaterinburg: Izd-vo Ural. un-ta, 2007. pp. 98–107.

8. Landeh D.V., Snarskijj A.A., Bezsudnov I.V. Internetika: Navigacija v slozhnykh setjakh: modeli i algoritmy. M.: Librokom (Editorial URSS), 2009. 264 p.

Рецензенты:

Матов А.Я., д.т.н., профессор, и.о. заведующего отделом Института проблем регистрации информации НАН Украины, г. Киев;

Калиновский А.Я., д.т.н., старший научный сотрудник Института проблем регистрации информации НАН Украины, г. Киев.

Работа поступила в редакцию 28.05.2012.