Теоретические вопросы информатики, прикладной математики, компьютерных наук и когнитивно-информационных технологий

УДК 519.21

Жуков Д.О., Новикова О.А., Алёшкин А.С.

Московский технологический университет МИРЭА, г. Москва, Россия

ВОЗМОЖНОСТЬ ИСПОЛЬЗОВАНИЯ МЕТОДОВ ПОЧТИ – ПЕРИОДИЧЕСКИХ ФУНКЦИЙ, ВЕЙВЛЕТ АНАЛИЗА И ТЕОРИИ САМОПОДОБИЯ ХЁРСТА ДЛЯ ПРОГНОЗИРОВАНИЯ НОВОСТНЫХ СОБЫТИЙ В ИНФОРМАЦИОННОМ ПРОСТРАНСТВЕ

Аннотация

В представленной работе рассматривается возможность применения методов почти – периодических функций, вейвлет анализа и теории самоподобия Хёрста для анализа спектров поведения с течением времени векторов, задающих положение кластеров новостных сообщений в информационном пространстве. Суть предлагаемого авторами подхода заключается в применении методов математической лингвистики (разметка текста, нормализация, комментирование) для создания словаря и коллекции новостных текстовых сообщений, привязанных к шкале времени. Это позволяет, используя стандартные методы, создать для каждой новости её векторное представление. Для всей совокупности векторов в представленной статье предлагается ввести понятие директора (условной оси, характеризующей основное направление всех векторов). Изменение с течением времени метрики (косинус угла) векторов, задающих положение центров кластеров относительно директора, образуют спектры информационных процессов, анализ которых с помощью методов почти – периодических функций, вейвлет анализа и теории самоподобия Хёрста может позволить выявить наличие повторяемости некоторых групп социальных событий, и тем самым прогнозировать их возможное проявление в будущем.

Ключевые слова

Информационное пространство, вектор новости, кластеризация новостей, директор информационного пространства, спектр информационного процесса, метод почти – периодических функций, самоподобие, метод Хёрста, вейвлет анализ.

Zhukov D.O., Novikova O.A., Alyoshkin A.S.

Moscow Technological University MIREA, Moscow, Russia

FEASIBILITY OF USING METHODS OF ALMOST PERIODIC FUNCTIONS, WAVELET ANALYSIS AND HURST SELF-SIMILARITY FOR PREDICTING NEWS EVENTS IN THE INFORMATION SPACE

Abstract

In the present paper we consider the feasibility of using methods of almost periodic functions, wavelet analysis and Hurst self-similarity to analyze in length of time the behavior spectra of vectors defining the position of the news reports clusters in the information space. The essence of the authors approach is to apply the methods of mathematical linguistics (text markup, normalization, comment) to create a dictionary and a collection of news text messages tied to the time scale. This makes it possible to create a vector representation for each newsletter, using standard methods. It is proposed to introduce the concept of a director (a conditional axis characterizing the basic direction of all vectors) for the entire set of vectors presented in the article. Time progress in metrics (cosine of the angle) of the vectors defining the position of the clusters centers relative to directors form

spectra of information processes. The analysis by the methods of almost - periodic functions, wavelet analysis and Hurst self-similarity can help identify the presence of recurrence of certain social events groups, and thus predict their possible manifestation in the future.

Keywords

Information space; news vector; news clustering; a director of the information space; information process spectrum; method of almost periodic functions; self-similarity; Hurst method; wavelet analysis.

Введение

Решение задач прогнозирования поведения систем и протекающих в них процессов является очень важным для различных областей деятельности, начиная с технических объектов и заканчивая социальными структурами. В настоящее время наименее проработанными (вследствие большей сложности) являются вопросы прогнозирования поведения социальных и экономических систем, поскольку наличие человеческого фактора создает не только стохастичность (случайность), но и неопределенность (недетерминированность возможных реализаций состояний и выборе путей их реализации).

В настоящее время существует выдвинутая Нассимом Николасом Талебом теория, в которой рассматривается природа возникновения и реализации неожиданных информационных событий.

В своей теории Нассим Талеб предполагает, что человечество не способно успешно прогнозировать своё будущее, а уверенность в своих знаниях опережает сами знания и порождает феномен «сверхуверенности».

Поскольку в реальном мире существуют причинно–следственные связи, то на наш взгляд нельзя однозначно утверждать, что данная задача не имеет решения, и в представленной нами работе мы попытаемся обозначить некоторые возможные пути её решения.

Краткий обзор существующих моделей прогнозирования процессов в экономических, социальных и технических системах

Прежде чем описать предлагаемый нами подход необходимо кратко рассмотреть некоторые существующие методы и модели, применяемые для прогнозирования развития процессов в технических, экономических и социальных системах.

Одним из важных инструментов прогнозирования является имитационное моделирование и использование математического аппарата искусственных нейронных сетей. В работе [1] это использовалось для решения вопроса оценки запасов и прогнозирования извлечения нефти, что находится в пограничной области между техническими и экономическими системами.

Нелинейные модели прогноза на основе нейронной сети и комбинации различных экономических параметров часто используются для прогнозирования развития отдельных отраслей экономики и их влияния на глобальные системы. В работе [2] этот подход был применен для прогнозирования влияния в 2012 – 2020 годах на экономику Китая в механизации сельского хозяйства. Результаты показывают, что требование полной мощности сельскохозяйственных машин для экономики будет составлять 1232298,2 МВт к 2015 году (что полностью подтвердилось) и 1560579,6 МВт к 2020 г.

Некоторую аналогию между техническими и экономическими системами можно найти, например, при рассмотрении процесса возникновения так называемой разладки. В процессе своего функционирования технические системы могут испытывать, в частности, случайные широкополосные и сложно-периодические возмущения (так называемые ритм-сигналы), не подпадающие ни под один из известных законов распределения случайных величин, что может привести к аномальному состоянию технической системы (разладке). Большинство традиционных методов и моделей для выявления разладок основываются на предположении, что вероятностные характеристики выходного ритм-сигнала у исправного технического объекта с течением времени не изменяются, т.е. измеренный стохастический сигнал является стационарным, и основаны на преобразовании Фурье (Fast Fourier Transform (FFT), помимо которого ещё достаточно часто используются методы, основанные на распределение Вигнера-Вилля, вейвлет-преобразование и преобразование Гильберта-Хуанга. Однако ритм-сигналы могут носить сложный нестационарный характер, что приводит к невозможности применения традиционных методов выявления разладок и их прогнозированию.

В некоторых экономических процессах, например биржевая торговля, часто наблюдается явление неожиданной смены тренда, что по сути также является разладкой системы, для прогнозирования которой сейчас используются методы технического анализа, основанные на многомерном статистическом анализе, а также фундаментальный анализ.

Обсуждая математическое моделирование и прогнозирование развития процессов в экономических и социальных системах можно обратить внимание на работы российской научной школы академика А.А. Петрова, чл. – корр. И.Г. Поспелова [3] и профессора Ф.Т. Алескерова [4], посвященные теории выбора (модель интервального выбора, модель коллективного выбора, основанные на рассмотрении классов бинарных отношений, таких как слабые бипорядки, простые и простейшие полупорядки, на основе которых может быть построена поведенческая модель биржи.

Для прогнозирования развития рынков и оценки их потенциала можно использовать кластерный анализ и математический аппарат цепей Маркова. В работе [5] показано, что формирующиеся рынки могут быть разделены с помощью кластерного анализа на группы в зависимости от имеющихся у них атрибутов, а переходы между группами (кластерами) могут быть описаны с помощью цепей Маркова. Это позволяет прогнозировать эволюцию развивающихся рынков.

Прогнозирование стабильности финансовых систем, устойчивости и уязвимости можно осуществить, используя стохастические имитационные модели, что в ряде случаев дает очень хорошие результаты, согласующиеся с практикой. В частности, такой подход был применен в работе [6]. Результаты исследования показывают улучшение стабильности румынской финансовой системы в период 1999-2007 гг., а рассчитанный с использованием стохастической имитационной модели совокупный индекс отражает финансовые завихрения румынского банковского кризиса 1998-1999 годов и ипотечного кризиса 2007 года.

Модели, использующие нелинейные степенные зависимости, могут быть применены для описания динамики колебания цен на высоколиквидные товары, а также демографических изменений в обществе [7].

Следует подчеркнуть, что задача прогнозирования в социальных системах и информационном пространстве существенным образом отличается от прогнозирования процессов в экономических и технических системах.

В настоящее время основные работы в области прогнозирования социальных и информационных процессов направлены на описание поведения пользователей в социальных сетях и сети интернет, и не рассматривают возможности прогнозирования реализации каких-либо новостных событий в информационном пространстве, что является очень важным.

В работе [8] авторы показывают, что существующие алгоритмы рекомендации выбора новостей (прогнозирование предпочтений пользователей) редко учитывают, что со временем характеристики поведения пользователей, просматривающих новости становятся упорядоченными. Поэтому эффективность традиционных алгоритмов рекомендации выбора новостей не достаточно хороша в предсказании следующей новостной статьи, которую выберет пользователь. Чтобы решить эту проблему, авторы работы [8] предлагают упорядоченный во времени совместный алгоритм фильтрации рекомендаций (TOCF), который принимает во внимание характеристики последовательности действий пользователей во времени. Кроме того, в [8] предложен метод, позволяющий вычислить подобие поведения среди различных пользователей.

Для выбора контента, интересного большинству активных пользователей (по сути это тоже является прогнозированием их предпочтений), социальные сети используют различные алгоритмы фильтрации (CF). Для решения этой задачи имеется широкий спектр схем CF. Однако большинство из них не может решить проблему так называемого "холодного" запуска, которая обозначает ситуацию, что социальным сетям не удается найти рекомендации для прогнозирования выбора контента новыми пользователями. С этой целью авторы [9] предлагают bi-кластеризацию и комплексный BiFu метод решения проблемы "холодного" запуска. Чтобы идентифицировать источники оценки для рекомендации, используется [9] понятие популярных элементов и экспертных оценок. Для того чтобы уменьшить размерность матрицы оценки, BiFu усиливает метод bi-кластеризации, а для преодоления разреженности данных и разнообразия оценки используется метод сглаживания.

Некоторые вопросы прогнозирования и управления действиями пользователей в социальных сетях на основе анализа моделей стохастической динамики происходящих процессов описаны в работе [10]. Авторы рассматривают социальную сеть, в которой происходит стохастический обмен знаниями на основе предпочтений, опыта, доверия и дружеских отношений при наличие узлов обмана. Используя метрику доверия, авторы оценили, как при определенных условиях можно прогнозировать поведение и манипулировать доверием в некоторых нетривиальных случаях.

Существенное распространение для моделирования и прогнозирования развития процессов в социальных сетях получило применение клеточных автоматов. В работе [11] авторы показывают, что поведение социальной системы может зависеть от свойств внешней среды и структуры поведения, которая

описывается с помощью четырех параметров в поведении: разнообразие, связность, взаимозависимость, и адаптируемость. В этой статье было показано, что поведение становится более упорядоченным и целенаправленным при увеличении взаимозависимости и адаптивности, и таким образом можно его прогнозировать и управлять им, влияя на соответствующие параметры.

Авторы работы [12] изучали влияние структуры сетей (случайные структуры, маленькие миры, цикл, колесо, звезда, иерархическая), в которых функционируют клеточные автоматы, и правил их поведения на динамику процессов в социальных сетях. Было показано, что при одинаковых правилах взаимодействия клеток динамика процессов сильно зависит от топологии сети (неупорядоченные структуры оказывают тормозящее влияние и наибольшая скорость наблюдается в регулярных структурах). Таким образом, при прогнозировании протекания процессов в социальных сетях крайне важно учитывать их топологию.

В работе [13] была рассмотрена модель поведения клеточного автомата, на основе модели Изинга. Данный подход был применен к прогнозированию социального поведения людей двух типов: конформистов (идущих за большинством) и нонконформистов (демонстрируют индивидуальные предпочтения). По мнению авторов, такая модель может быть полезна при прогнозировании поведения игроков на бирже. Авторы используют понятие информационной энтропии для состояний клеточного автомата как коллективной переменной, позволяющей сравнивать как детерминированное поведение, так и вероятностное поведение (на основании модели среднего поля и бифуркационных диаграмм).

В работе [14] для решения задачи прогнозирования авторы изучают временные зависимости в потоках событий и вводят кусочно-постоянную аппроксимацию их интенсивности, применяя Байесовский подход и распределение Пуассона к описанию выборки важности будущих событий. Это позволяет построить нелинейные временные зависимости для предсказания будущих событий с использованием деревьев решений. Однако, авторы не рассматривают наличие памяти информационного пространства и возможность его самоорганизации, что является, на наш взгляд, очень важным для прогнозирования редких событий.

В работе [15] для описания структуры новостного информационного пространства вводится понятие директора - условной оси, положение которой определяется усреднением направлений векторов, задающих положение центров всех новостных кластеров. Направление директора может быть рассчитано с помощью метода наименьших квадратов для отклонения углов векторов, задающих положение центров информационных кластеров от условного направления, которое принимается за директор. Авторы [15] предлагают прогнозировать возможные события в социальном информационном пространстве на основе анализа самоподобия в поведении директоров с использованием модели Хёрста, на основании этого выявлять наличие периодичности в их поведении, и таким образом прогнозировать повторяемость.

Отмечая существенные успехи, достигнутые в прогнозировании поведения процессов для технических и экономических систем, нельзя не обратить внимание на то, что для социальных систем и новостных событий существенный прогресс в данной области отсутствует, а механистическое перенесение методов и подходов технических и экономических наук не дает желаемых результатов.

Априори все данные, которые можно использовать для прогнозирования новостных событий, являются по своему характеру гетерогенными (имеют разный формат представления и единицы измерения), и применение существующих методов и подходов интеллектуального анализа данных (Big Data) не приносит ощутимых результатов, поскольку имеющиеся инструменты очистки, извлечения и анализа данных, скорее всего, не позволяет извлекать сами знания, уничтожая их скрытые (латентные) связи.

Постановка задачи исследования

При создании модели прогнозирования новостных событий необходим математический аппарат, который позволил бы формализовать характер данных и привести их к единой шкале измерений. Очевидно, что нельзя в одной модели проводить вычислительные операции, например, над лингвистическими оценками и величинами метрической шкалы без использования процедур отображения на формальное безразмерное множество.

Общая схема предлагаемых нами методов для прогнозирования новостных событий заключается в следующем:

1. Первоначально создается коллекция текстовых документов (новости, статьи научного, культурного и т.д. характера, и иные документы), составляющая информационное пространство. Каждый документ коллекции по описываемому в нем событию имеет привязку к шкале времени (база документов является темпоральной). На основе существующих методик математической лингвистики (разметка текста, нормализация, ком-

ментирование) и словаря коллекции можно создать формализованное векторное описание информационного пространства, представляющего в исходном виде набор текстов на естественных языках. Это позволит решить проблему разнородности данных и единиц измерения параметров различных процессов, и все данные формализуются единым образом в виде векторов.

2. Для прогнозирования событий, которые могут произойти в будущем, можно провести анализ самоподобия и повторяемости с течением времени изменений в векторном информационном пространстве описываемых событий с помощью теории Хёрста и модели почти-периодических функций.

Методика исследования

При создании модели прогнозирования новостных событий необходим математический аппарат, который позволил бы формализовать характер данных и привести их к единой шкале измерений. Очевидно, что нельзя в одной модели проводить вычислительные операции, например над лингвистическими оценками и величинами метрической шкалы, без использования процедур отображения на формальное безразмерное множество. Для этого можно использовать существующие методы математической лингвистики (разметка текста, нормализация, комментирование), что позволяет формализовать (в виде векторов) описание информационного пространства, исходно представляющего набор текстов на естественных языках. Разметка текста (линеаризация) необходима для распознавания и извлечения слов-терминов (термов) и составных терминов (термов) из исходного текста. При разметке учитывается пунктуация и переносы слов, и игнорируются стоп-слова (малоинформативные, но часто встречающиеся в текстах слова, например предлоги, союзы, причастия, междометия и т.д.).

При нормализации из исходного текста удаляется грамматическая информация (падежи, числа, глагольные виды и времена, залоги причастий, род и так далее).

Комментирование текста необходимо, потому что, одно и то же слово в тексте может быть употреблено в различных значениях. Комментирования заключаются в добавлении в документ метаданных с информацией о части речи, о наличии у слова других значений и синтаксического разбора предложения.

После разметки, нормализации и комментирования текстов каждый документ коллекции информационного пространства представляется набором термов (терминов), которые в него входят, и их частотами вхождения, что позволяет создать векторную модель каждого документа. Векторная модель основывается на гипотезе (statistical semantics hypothesis) о том, что: статистические зависимости употребления слов и терминов (термов) могут быть использованы для нахождения заложенного в текст смысла [16]. Согласно гипотезе о скрытых связях, пары слов, которые встречаются в похожих моделях, стремятся иметь близкую семантическую зависимость [17].

Пусть общее количество имеющихся различных уникальных термов равно m (термы могут быть пронумерованы от 1 до m), а число имеющихся в коллекции текстовых документов равно n (документы могут быть пронумерованы от 1 до n). Тогда данному документу (обозначим его $\mathbf{X_i}$, i – принимает значения от 1 до n), который содержит некоторое число термов (терминов), их порядок в наборе не важен) можно поставить в соответствие вектор $\mathbf{X_i}=\{\omega_1, \omega_2, \omega_3, \cdots \omega_j \cdots \omega_m\}$, где первый элемент соответствует числу (частоте) вхождений в документ первого термина, второй – второго, третий – третьего и т.д. Каждая ω_j ω_j может принимать любое положительное значение от 0. Все $\mathbf{X_i}$ образуют матрицу, большинство значений элементов которой равны 0 (матрица является разреженной). Таким образом, каждый документ коллекции является точкой в многомерном пространстве (вектора в векторном пространстве, размерности $\mathbf{R^m}$). Близко лежащие друг к другу точки соответствуют семантически схожим документам, что позволяет сделать кластеризацию информационного пространства по смысловым группам близких по значению документов (провести тематическое разделение). При этом каждая такая группа (кластер) текстовых документов может быть описана вектором, задающим положение центра. Данный подход используется, например, в задачах классификации и отнесении сообщений новостной ленты к определенной группе и весьма эффективен с точки зрения экономии вычислительных ресурсов.

Для кластеризации текстовых документов в основном используются следующие виды алгоритмов: алгоритм k-средних, Scatter-Gather, BIRCH и алгоритмы иерархической кластеризации.

В линейных алгоритмах [18] первоначально множество кластеров считается пустым, а для каждого нового сообщения выполняются следующие операции:

- оцениваются расстояния от вектора нового документа до центров всех кластеров;
- если минимальное расстояние больше некоторого наперед заданного числа, то новое сообщение помещается в отдельный кластер;
- если нет, то в один (или несколько ближайших);
- пересчитываются центры измененных кластеров.

Для классификации текстовых документов может быть использован латентно–семантический анализ (LSA) [16], который предназначен для извлечения контекстно-зависимых значений лексических единиц на основе факторного анализа и статистической обработки больших корпусов текстов. В качестве метрик при кластеризации можно использовать, как косинус угла между векторами, так и работать напрямую с расстоянием между векторами.

Отметим, что коллекция текстовых документов и описывающие их вектора, должны являться темпоральной базой данных (каждый вектор по заключенному в нем информационному смыслу привязан к шкале времени (дате описываемого события)).

Для всей совокупности векторов мы предлагаем ввести понятие директора (условной оси, характеризующей основное направление всех векторов) или внешней оси (которые задаются один раз и не изменяются с течением времени). Появление новых информационных событий и исчезновение возникших ранее, будет приводить к изменению положений центров смысловых кластеров. Изменение с течением времени метрики (косинус угла) векторов, задающих положение центров кластеров относительно директора (или произвольно заданной внешней оси), будет создавать спектры информационных процессов, анализ которых с помощью методов почти-периодических функций, вейвлет анализа и самоподобия Хёрста может позволить выявить наличие повторяемости некоторых групп социальных информационных событий, и тем самым сделать возможным прогнозирование их проявления в будущем.

Структура новостного информационного пространства

Используя формализованные представления данных в информационном пространстве, попробуем создать метод прогнозирования новостных событий, основанный на предположении о том, что процессы, протекающие в пространстве информационных событий, могут обладать свойством самоподобия.

В представленной работе мы описываем разработанный нами для прогнозирования информационных событий подход, суть которого состоит в следующем:

1. Учитывая, что в реальном мире существуют множественные причинно-следственные связи, то при отображении событий в информационное пространство эти связи также должны хотя бы частично сохраняться (*правило сохранения причинно-следственных связей при любых отображениях*).
2. Любое событие может быть описано в информационном пространстве некоторым новостным кластером, имеющим свои собственные характеристики (*правило кластеризации информации*). В любой момент времени в информационном пространстве существует множество различных новостных кластеров (см. рис. 1). Информационное пространство является “зеркалом” физического мира, отображающего его основные свойства и взаимосвязи событий.
3. С течением времени новостные кластеры могут изменяться или исчезать, и эти изменения могут быть описаны в рамках динамических моделей.
4. Информационные процессы могут быть описаны с помощью спектров, которые по своей сути, характеризуют изменение его метрики с течением времени. Помимо метрики, определяемой как угол между директором и вектором, задающим положение центра информационного кластера, в данный момент времени можно использовать и иные метрики. Например, изменение объема (площади поверхности) фигуры кластера в информационном пространстве или ещё энтропийную метрику. Важно отметить, что любая из используемых метрик будет являться одномерной функцией от времени.

В зависимости от выраженности структуры для анализа спектров может быть использована теория почти периодических функций, вейвлет анализ и теория самоподобия Хёрста. Применение метода Фурье для анализа спектров информационных процессов является ограниченным в силу того, что реальные процессы по механизмам их формирования могут не являться суперпозицией гармонических колебаний, а иметь нелинейный характер, что не учитывается в существующих методиках обработки данных и приводит к потери информации о реальных процессах или получению ошибочных результатов. Несоответствие исследуемого процесса гармоническим колебаниям компенсируется в спектральном анализе простым увеличением числа гармонических компонент, что только еще больше затрудняет интерпретацию получаемых результатов. В результате разложение в ряд Фурье может давать значимую систематическую ошибку из-за несоответствия методов обработки свойствам реальных данных. В данном случае необходимо использовать методы анализа на основе почти-периодических функций [19], вейвлет преобразованиях [20] и теории самоподобия Хёрста [21].

Для анализа информационных процессов, спектр которых близок или хотя бы имеет отдаленное подобие с гармоническими процессами (степень сходства в данном случае заслуживает отдельного обсуждения), можно использовать метод почти-периодических функция или вейвлет анализ, а всех остальных теорию самоподобия.

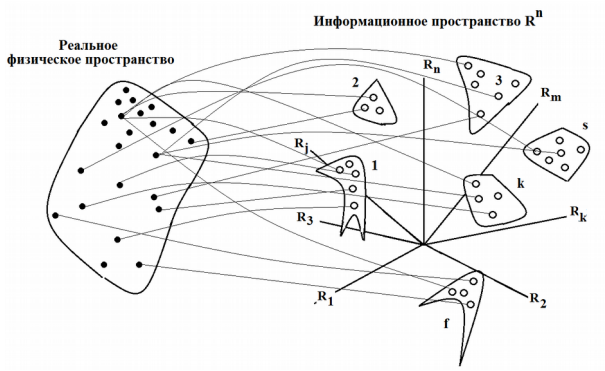

Рис. 1. Отображение событий из реального физического пространства в информационное и образование новостных кластеров

Использование метода почти – периодических функций в анализе информационных процессов

Исследование спектра информационного процесса (зависимости метрики, задающей положение смыслового информационного процесса от времени) можно провести, используя метод почти-периодических функций [19], в котором определение периодических компонент в спектре достигается преобразованием исходного ряда в новый ряд по формуле (1): $\chi(t_k)=\ln\{y(t_k-\Delta t_m)*y(t_k+\Delta t_m)/y^2(t_k)\}$, где $y(t_k-\Delta t_m)$, $y(t_k)$, $y(t_k+\Delta t_m)$ – значения метрики процесса в соответствующие моменты времени, t_k – моменты регистрации измеряемой характеристики, Δt_k – фиксированный пробный временной интервал. Результатом преобразования является ряд $\chi(t_k)$ с близким к нулю значением математического ожидания. Для определения наиболее близких к периодам значений параметров колебаний воспользуемся методом сдвиговых функций. Применение сдвиговых функций, основанных на метриках функционального анализа и теории почти-периодических функций, позволяет эффективно определять значения близких к периодам (почти-периодам). Частным случаем таких функций является (2): $a(\tau_k)=1/(n-k)*\Sigma\left|\chi(t_j+\tau_k)-\chi(t_j)\right|$ (вычисляется от j=1 до (n-k)), где n- количество точек исходного ряда, τ_k – пробная сдвижка.

Система почти-периодов τ функции χ(t) может быть определена как совокупность локальных минимумов сдвиговой функции: $\tau=\arg\min a(\tau)$ с условием: $\tau_{min}\le\tau\le\tau_{max}$, где τ_{min} и τ_{max} - естественные пределы поиска периода, выбираемые таким образом, что, с одной стороны, отбрасываются $\tau < \tau_{min}$, при которых функция $a(\tau)$ может принимать малые значения из-за инерционности функции $\chi(t)$, и, с другой стороны, отбрасываются $\tau > \tau_{max}$, при которых определение средней $a(\tau)$ становится ненадежным из-за малого числа членов суммирования в выражении (2).

Для выявления трендовой составляющей в наблюдаемых данных, соответствующих исходному ряду, необходимо использовать методы, позволяющие, во-первых, избежать навязывания определенной зависимости предполагаемому тренду, во-вторых, согласовать методы выделения тренда с уже найденными характеристиками процесса.

Целесообразно за основу выделения тренда взять метод, при котором из исследуемых данных исключаются колебания, присутствующие в исходном ряду. В качестве способа выделения трендовой составляющей можно использовать метод скользящей средней (3): $y^*(t)=[1/\tau]\cdot\Sigma y(t+\tau)$, где суммирование производится от $-\tau/2$ до $+\tau/2$, а $y^*(t)$ – значения трендовой составляющей y(t), соответствующее середине интервала усреднения, τ – количество элементов, по которым ведется усреднение.

Помимо метода почти-периодических функций мы предполагаем, что повторяемость процессов в информационном пространстве, а значит и их прогнозирование может быть исследовано на основе теории вейвлетов и самоподобия Хёрста.

Вейвлет анализ информационных процессов

Преимущество вейвлет анализа перед анализом Фурье состоит в том, что он позволяет проследить за изменением спектра процесса с течением времени и определить преобладающие частоты [20]. Вейвлет анализ дает хорошие результаты для исследования временных спектров с изменяющимся спектральным составом. Вейвлеты объединяют в себе свойства подобия и локализации по частоте и времени. Семейство вейвлет функций должно удовлетворять следующим требованиям:

1. Допустимость. Анализирующий вейвлет ψ(t) (материнский вейвлет) должен иметь нулевое среднее значение: $\int_{x_0-0}^{x_0+0} G(x,p)dx \int_{-\infty}^{+\infty}\psi(t)dt=0$.

2. Подобие. Всё семейство вейвлет функций может быть получено из анализирующего вейвлета путем масштабного преобразования и сдвига: $\psi_{a,b}(t)=\psi\left[\frac{t-b}{a}\right]$. Семейство вейвлет функций является двухпараметрическим: параметр a – масштаб (растяжение/сжатие) функции, параметр b – положение (сдвиг) функции.

3. Обратимость. Существует обратное преобразования, однозначно восстанавливающее исходную функцию по ее вейвлет – преобразованию.

4. Регулярность. Функция ψ(t) должна быть локализована.

Для целей анализа (лучшего разрешения) спектров информационных процессов, спектральный состав которых может меняться с течением времени может быть использован в качестве анализирующего (материнского)

комплексный вейвлет Морле: $\psi(t)=\exp\left\{\frac{-t^2}{2}+i\omega t\right\}$, который сжимаем или растягиваем функцию как целое, не нарушая подобия отдельных функций семейства.

Учитывая, что спектр информационного процесса описывается одномерной метрикой от времени, то мы предлагаем использовать следующее вейвлет преобразование: $\omega(a,b)=a^k\int_{-\infty}^{+\infty} f(t)\psi\left[\frac{t-b}{a}\right]dt$. Для данного преобразования должно выполняться обращение, имеющее следующий вид:

$$f(t)=a^k\frac{1}{C_\psi}\int_0^{+\infty}\int_{-\infty}^{+\infty}\frac{\omega(a,b)}{a^{k+3}}\psi\left[\frac{t-b}{a}\right]da\,db\,.$$

Показатель степени масштабного множителя к выбирается в зависимости от целей анализа. При к=-1 равные значения вейвлет-коэффициентов ω(a,b) соответствуют равным амплитудам пульсаций сигнала, независимо от масштаба пульсаций. Соответственно для C_ψ должно выполняться условие: $C_\psi=\int_{-\infty}^{+\infty}\frac{\widehat{\psi(\omega)}}{|\omega|}d\omega<\infty$, где $\psi(\omega)=\int_{-\infty}^{+\infty}\psi(t)e^{-i\omega t}dt$, ψ(t) - комплексная функция Морле, удовлетворяющая условиям (1)–(4). Вейвлет – анализ позволяет восстановить полную эволюцию спектрального сигнала во времени.

Для анализа спектров информационных процессов и на этапе разложения сигнала по вейвлетам, и на этапе восстановления исходного сигнала по его вейвлет образу условие $C_\psi=\int_{-\infty}^{+\infty}\frac{\widehat{\psi(\omega)}}{|\omega|}d\omega<\infty$ можно заменить более "мягким" условием вида: $C_\psi=\int_{-\infty}^{+\infty}\frac{\widehat{\psi(\omega)}\star\widehat{\phi(\omega)}}{|\omega|}d\omega$ где вместо одного из вейвлетов можно использовать сингулярную функцию (например, δ – функцию), не являющуюся вейвлетом. Это приводит к повышению точности анализа.

Выявление самоподобия информационных процессов с помощью теории Хёрста

При анализе временной зависимости метрики информационного процесса, для его описания может быть применена теория самоподобия Хёрста [21], суть которой заключается в том, что непрерывный стохастический процесс Y(t) считается статистически самоподобным с параметром Хёрста (Hurst) H (0,5≤H≤1), если для любого положительного числа **a**, случайные процессы Y(t) и $\mathbf{a}^{-H}$Y(at) будут иметь одинаковые статистические свойства. Значение H=0.5 показывает отсутствие самоподобности, а значения H, близкие к единице, показывают её большую степень.

Для выборочного случайного набора X_j (j=1, N) можно определить выборочное среднее $M=[1/N]\cdot\Sigma X_j$, (суммирование ведется по j от 1 до N); выборочную дисперсию $S^2_N=[1/N]\cdot\Sigma(Xj-M)^2$ (суммирование ведется по j от 1 до N) и и интегральное отклонение $D_j=\Sigma X_k-jM$ (суммирование ведется по k от 1 до j). Изменчивость самоподобного процесса СП на интервале N определяется как неубывающая функция длины интервала $R_n=\max D_j-\min D_j$ (1<j<N)

Хёрстом было показано, что для большинства естественных процессов при больших значениях N выполняется соотношение: $R_N/S_N\approx\{N/2\}^H$ или $\log(R_N/S_N)\approx H\cdot\log(N/2)$, где H – параметр Хёрста.

Степень самоподобия процесса можно оценить путем построения графика отношения $\log(R_N/S_N)$ в зависимости от log(N/2) при разных N и вычислить величину H как тангенс угла наклона полученной линии. Следует заметить, что полученное множество точек не будут лежать на одной линии, поэтому их следует аппроксимировать линией, например, по методу наименьших квадратов. Данная методика определения параметра Хёрста получила название R/S-метод. R/S-метод дает лишь приближенное значение показателя Хёрста, поэтому для его вычисления целесообразно пользоваться несколькими методиками и сравнения полученных результатов. Например на основе периодограммного анализа. Для самоподобного процесса $X=\{x_j\}$ вычисляется периодограмма:

$$I_N(\omega)=[1/(2\pi N)\}\cdot\Sigma x_k\cdot\exp(jk\omega),\ \omega[0;\pi],$$

(суммирование ведется по k от 1 до j), где N - количество отсчетов временного ряда. Учитывая, что самоподобие влияет на характер спектра S(ω), должен получаться график зависимости спектральной плотности вида: $I_N(\omega)\approx[\omega]^{1-2H}$, при ω→0. Из последнего выражения следует, что множество случайных точек $\{\log[I_N(\omega)];\ \log(\omega)\}$ будет располагаться линейно с коэффициентом наклона линии 1–2H. На практике для вычисления оценки должны использоваться только нижние 10% частот, т.к. описанное выше поведение справедливо только для области частот, близких к нулю. Основным недостатком данного метода является большой объем вычислений при построении оценки показателя Хёрста.

Самоподобные процессы принято разделять на два класса: точно самоподобные и асимптотически самоподобные. Различие между этими двумя классами заключается в том, что для функции автокорреляции точно самоподобных процессов выполняется условие: $R(Y_k)=R(Y_q)$, а для асимптотически самоподобных: $R(Y_k)\to R(Y_q)$, при k→q. Дисперсия для обоих классов процессов определяется одинаковым образом: $D(Y_k)=D(Y_q)/k^\beta$ $D(Y_k)=\frac{D(Y_q)}{k^\beta}$, где β – параметр самоподобия ($0<\beta<1$), связанный с параметром Хёрста следующим соотношением: $\beta=2(1-H)$, а k – величина блока разбиения исходных данных.

Коэффициент Хёрста находится по зависимости

логарифма среднего значения дисперсии от логарифма величины блоков m разбиения исходной выборки данных. При наличии самоподобия, полученная зависимость должна иметь линейный вид. Таким образом, если аппроксимировать логарифмические зависимости линейной функцией, то с помощью метода наименьших квадратов можно вычислить коэффициенты данного линейного уравнения и коэффициент корреляции данных с линейной зависимостью. Тангенс угла наклона линейной зависимости связан с параметром Хёрста (Hurst) H.

Теория самоподобия Хёрста находит применение во многих областях науки и техники, например [22] таких, как анализ флуктуации скорости в высокой турбулентности и числа Рейнольдса при течениях, а также коэффициента диффузии, трафика передачи данных в компьютерных сетях [23].

Работа выполнена за счет финансирования Министерством образования и науки Российской Федерации конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов, номер проекта 28.2635.2017/ПЧ, наименование «Разработка моделей стохастической самоорганизации слабоструктурированной информации и реализации памяти при прогнозировании новостных событий на основе массивов естественно-языковых текстов».

Литература

1. Bansal Y., Ertekin T.,. Karpyn Z., Ayala L., Nejad A., Suleen F., Balogun O., Liebmann D., Sun Q. Forecasting well performance in a discontinuous tight oil reservoir using Artificial Neural Networks. Society of Petroleum Engineers - SPE USA, Unconventional Resources Conference 2013, Pages 239-250, SPE USA Unconventional Resources Conference 2013; The Woodlands, TX; United States; 10 April 2012 through 12 April 2012; Code 98227 (2013).
2. Ju J.Y, Zhu R.X., Geng, L. Forecasting and analysis the demand of agricultural mechanization for economic development Advanced Materials Research. Volume 694 697, Pages 3512-3515, 4th International Conference on Manufacturing Science and Engineering, ICMSE 2013; Dalian; China; 30 March 2013 through 31 March 2013; Code 97229, (2013).
3. Петров А.А., Поспелов И.Г. Математические модели экономики России. // Вестник РАН, Т.79, № 6, – С. 492-506, (2009).
4. Алескеров Ф. Т. Индексы влияния, учитывающие предпочтения участников по созданию коалиций, Доклады РАН.– т. 414, № 5. – P. 594—597, (2007).
5. Huo D. Cluster analysis of market potential in emerging markets: A dynamic research based on Markov chain. Romanian Journal of Economic Forecasting, Volume 16, Issue 4, Pages 218-231, (2013).
6. Albulescu C.T. Forecasting the Romanian financial system stability using a stochastic simulation model. Romanian Journal of Economic Forecasting, Volume 13, Issue 1, Pages 81-98, (2010).
7. Akaev A.A., Sadovnichy V.B., Korotayev A.C. On the dynamics of the world demographic transition and financial-economic crises forecasts., European Physical Journal: Special Topics, Volume 205, Issue 1, Pages 355-373, (2012).
8. Yingyuan Xiao, Pengqiang Ai, Ching-hsien Hsu, Hongya Wang, Xu Jiao. Time-ordered Collaborative Filtering for News Recommendation. // China Communication. – Vol. 12, № 12. – P. 53-62.
9. Daqiang Zhang, Ching-Hsien Hsu, Min Chen, Quan Chen, Naixue Xiong, Jaime Lloret. Cold-Start Recommendation Using Bi-Clustering and Fusion for Large-Scale Social Recommender Systems. // IEEE Transactions on Emerging Topics in Computing.– Vol. 2, № 2. – P.239-250, (2014).
10. Allodi, L., Chiodi, L., Cremonini, M. Modifying trust dynamics through cooperation and defection in evolving social networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), (2011).
11. Hay, J., Flynn, D. How external environment and internal structure change the behavior of discrete systems. Complex Systems., 25 (1), pp. 39-49, (2016).
12. Hay, J., Flynn, D. The effect of network structure on individual behavior. Complex Systems., 23 (4), pp. 295-311, (2014).
13. Bagnoli, F., Rechtman, R. Topological bifurcations in a model society of reasonable contrarians. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 88(6), 062914, (2013).
14. Asela Gunawardana, Christopher Meek, Puyang Xu. A Model for Temporal Dependencies in Event Streams. / Microsoft Research, – P. 1-8, (2011).
15. Zhukov D.O., Lesko S.A. Trends, self-similarity and forecasting of news events in the information domain, its structure and director. // 2015 IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015, December 19-21, Chengdu, Sichuan, China, P. 870-873, 978-1-5090-1893-2/15, DOI 10.1109/SmartCity.2015.178, (2015).
16. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press, (2007).
17. Turney, P. D. The latent relation mapping engine: Algorithm and experiments. // Journal of Artificial Intelligence Research, 33, pp. 615–655, (2008).
18. Costa G., Mango G., and Ortale R. An incremental clustering scheme for data de-duplication // Data Mining and Knowledge Discoverym, Vol. 20. No. 1. pp. 152-187, (2010).
19. Левитан Б. М., Почти-периодические функции, М., (1953).
20. Grossman A., Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape // SIAM J. Math. 1984. Vol. 15, No. 4, P. 723-736.
21. Цыбаков Б.С. Модель телетрафика на основе самоподобного случайного процесса. Москва, – 31 с., (1999).
22. Anvari M.A., Aghamohammadi C.A, Dashti-Naserabadi H.A., Salehi E.A., Behjat E.A., Qorbani M.A., Khazaei Nezhad M.A., Zirak M.A., Hadjihosseini A.B., Peinke J.B., Tabar M.R.R. Stochastic nature of series of waiting times. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, Volume 87, Issue 6, Article number 062139, (2013).
23. Walter Willinger, Murad S. Taqqu, Robert Sherman and Daniel V. Wilson, “Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level”. IEEE/ACM Transactions on Networking, Vol. 5, No. 1, (1997).

References

1. Bansal Y., Ertekin T.,. Karpyn Z., Ayala L., Nejad A., Suleen F., Balogun O., Liebmann D., Sun Q. Forecasting well performance in a

discontinuous tight oil reservoir using Artificial Neural Networks. Society of Petroleum Engineers - SPE USA, Unconventional Resources Conference 2013, Pages 239-250, SPE USA Unconventional Resources Conference 2013; The Woodlands, TX; United States; 10 April 2012 through 12 April 2012; Code 98227 (2013).
2. Ju J.Y, Zhu R.X., Geng, L. Forecasting and analysis the demand of agricultural mechanization for economic development Advanced Materials Research. Volume 694 697, Pages 3512-3515, 4th International Conference on Manufacturing Science and Engineering, ICMSE 2013; Dalian; China; 30 March 2013 through 31 March 2013; Code 97229, (2013).
3. Petrov A.A., Pospelov I.G. Mathematical models of the economy of Russia // Herald of the RAS, T.79, number 6, - pp 492-506, (2009).
4. Aleskerov F.T. Indexes of influence considering preferences of the participants to build coalitions Reports RAN.- t 414, № 5. -. P. 594-597, (2007).
5. Huo D. Cluster analysis of market potential in emerging markets: A dynamic research based on Markov chain. Romanian Journal of Economic Forecasting, Volume 16, Issue 4, Pages 218-231, (2013).
6. Albulescu C.T. Forecasting the Romanian financial system stability using a stochastic simulation model. Romanian Journal of Economic Forecasting, Volume 13, Issue 1, Pages 81-98, (2010).
7. Akaev A.A., Sadovnichy V.B., Korotayev A.C. On the dynamics of the world demographic transition and financial-economic crises forecasts., European Physical Journal: Special Topics, Volume 205, Issue 1, Pages 355-373, (2012).
8. Yingyuan Xiao, Pengqiang Ai, Ching-hsien Hsu, Hongya Wang, Xu Jiao. Time-ordered Collaborative Filtering for News Recommendation. // China Communication. – Vol. 12, № 12. – P. 53-62.
9. Daqiang Zhang, Ching-Hsien Hsu, Min Chen, Quan Chen, Naixue Xiong, Jaime Lloret. Cold-Start Recommendation Using Bi-Clustering and Fusion for Large-Scale Social Recommender Systems. // IEEE Transactions on Emerging Topics in Computing.– Vol. 2, № 2. – P.239-250, (2014).
10. Allodi, L., Chiodi, L., Cremonini, M. Modifying trust dynamics through cooperation and defection in evolving social networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), (2011).
11. Hay, J., Flynn, D. How external environment and internal structure change the behavior of discrete systems. Complex Systems., 25 (1), pp. 39-49, (2016).
12. Hay, J., Flynn, D. The effect of network structure on individual behavior. Complex Systems., 23 (4), pp. 295-311, (2014).
13. Bagnoli, F., Rechtman, R. Topological bifurcations in a model society of reasonable contrarians. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 88(6), 062914, (2013).
14. Asela Gunawardana, Christopher Meek, Puyang Xu. A Model for Temporal Dependencies in Event Streams. / Microsoft Research, – P. 1-8, (2011).
15. Zhukov D.O., Lesko S.A. Trends, self-similarity and forecasting of news events in the information domain, its structure and director. // 2015 IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015, December 19-21, Chengdu, Sichuan, China, P. 870-873, 978-1-5090-1893-2/15, DOI 10.1109/SmartCity.2015.178, (2015).
16. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press, (2007).
17. Turney, P. D. The latent relation mapping engine: Algorithm and experiments. // Journal of Artificial Intelligence Research, 33, pp. 615–655, (2008).
18. Costa G., Mango G., and Ortale R. An incremental clustering scheme for data de-duplication // Data Mining and Knowledge Discoverym, Vol. 20. No. 1. pp. 152-187, (2010).
19. Levitan B.M. Almost periodic functions, M., (1953).
20. Grossman A., Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape // SIAM J. Math. 1984. Vol. 15, No. 4, P. 723-736.
21. Tsybakov B.S. Teletraffic model based on self-similar random process. Moscow, - 31 (1999).
22. Anvari M.A., Aghamohammadi C.A, Dashti-Naserabadi H.A., Salehi E.A., Behjat E.A., Qorbani M.A., Khazaei Nezhad M.A., Zirak M.A., Hadjihosseini A.B., Peinke J.B., Tabar M.R.R. Stochastic nature of series of waiting times. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, Volume 87, Issue 6, Article number 062139, (2013).
23. Walter Willinger, Murad S. Taqqu, Robert Sherman and Daniel V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level". IEEE/ACM Transactions on Networking, Vol. 5, No. 1, (1997).

Поступила: 2.03.2017

Об авторах:

Жуков Дмитрий Олегович, доктор технических наук, профессор, заместитель директора по научной работе института комплексной безопасности и специального приборостроения, Московский технологический университет МИРЭА, zhukovdm@yandex.ru

Новикова Ольга Александровна, соискатель, заместитель заведующего аспирантурой, Московский технологический университет МИРЭА, ol-novikova@bk.ru

Алёшкин Антон Сергеевич, кандидат технических наук, доцент кафедры автоматизированных систем управления института комплексной безопасности и специального приборостроения, Московский технологический университет МИРЭА, antony@testor.ru

Note on the authors:

Zhukov Dmitriy, doctor of technical sciences, professor, Deputy Director for Research of The Institute of Comprehensive Security and Special Instrumentation, Moscow Technological University MIREA, zhukovdm@yandex.ru

Novikova Olga, Competitor, deputy head of postgraduate study, Moscow Technological University MIREA, ol-novikova@bk.ru

Alyoshkin Anton, Candidate of technical sciences, Associate Professor of the Department of Automated Control Systems of the Institute of Comprehensive Security and Special Instrumentation, Moscow Technological University MIREA, antony@testor.ru