СТРУКТУРА НОВОСТНОГО ИНФОРМАЦИОННОГО ПРОСТРАНСТВА, САМОПОДОБИЕ И ДИРЕКТОР

[1]Жуков Д.О., [1]Лесько С.А.

[1]Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет информационных технологий, радиотехники и электроники» (МИРЭА), 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: kunaka@bk.ru

Для описания структуры новостного информационного пространства введено понятие директора - условной оси, положение которой определяется усреднением направлений векторов, задающих положение центров всех новостных кластеров. Рассмотрена разработанная авторами модель прогнозирования новостных событий на основе динамики приближения состояния информационного пространства к точке события. Для описания динамики достижения новостного события вводятся понятия понижающего и повышающего трендов изменения состояния информационного пространства.

Ключевые слова: информационное пространство, новостной кластер, директор информационного пространства, самоподобие процессов в информационном пространстве.

STRUCTURE NEWS INFORMATION SPACE, SELF-SIMILARITY AND DIRECTORS

[1]Zhukov D.O., [1] Lesko S.A.

[1]Federal State Educational Institution of Higher Education "Moscow State University of Information Technologies, Radio Engineering and Electronics" (MIREA) 119454, Russia, Moscow, Vernadsky prospect, 78,e-mail: kunaka@bk.ru

In a report to describe the structure of the news information space introduces the concept of the Director - conventional axis, whose position is determined by averaging the directions of the vectors that define the position of the centers of all news clusters. In addition, in the paper the authors developed a model predicting the news events on the basis of changes in the status information area closer to the point of the event. To describe the dynamics of achieving the notions of news events and up-down trend change of state information space.

Key words: information space, a news cluster director of information space, self-similarity of the processes in the information space.

Введение

По оценкам экспертов к 2020 году общемировой объем накопленных различных данных будет составлять 35000 Экзобайт (Экзо=10^{18}), что составит их рост по отношению к 2010 году в 44 раза. Причем до 90% будут составлять не структурированные или плохо структурированные данные.

Одним из направлений работы с этими данными является их анализ с целью прогнозирования появления новых информационных событий, что является очень сложной задачей. Прогнозирование появления новых информационных событий осложняется необходимостью поиска скрытых закономерностей в больших объемах слабоструктурированной гетерогенной информации и преодолением трудностей связанных с проблемой измеримости параметров протекающих информационных процессов. Априори все данные являются по своему характеру гетерогенными (*имеют разный формат представления и неоднозначные единицы измерения*). При создании модели 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.

В настоящее время новостное информационное пространство описывают на основе методов кластеризации текстов и их синтаксического анализа. Отметим, что задача кластеризации потоков новостных сообщений обладает высокой размерностью данных. Представление текстовых данных обычно осуществляется с помощью процедуры сопоставления каждого признака с функцией-индикатором данного

слова, и таким образом общая размерность пространства задачи R^n определяется общим количеством различных признаков [1]. В качестве признаков в информационном пространстве используются слова и их сочетания, поэтому общая размерность **n** пространства R^n может достигать величины нескольких десятков миллионов. При этом вектор признаков каждого отдельного сообщения имеет только очень небольшое количество не нулевых координат и поэтому называется сильно разреженным.

Значительная часть современных методов информационного поиска основывается на алгоритмах кластеризации текстов, которые условно по своему типу можно разделить на линейные и иерархические. В линейных алгоритмах [2-5] первоначально множество кластеров считается пустым, а для каждого нового сообщения выполняются следующие операции:

- Оцениваются расстояния от вектора нового сообщения 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.

Особенностью алгоритма является то, что не учитывается довольно большой объем мета-информации: даты публикации материала, наличие взаимных ссылок между статьями, дополнительные ссылки по теме и решение о принадлежности какой-либо точки принимается только один раз (в этом смысле алгоритм является линейным по времени).

Алгоритмы иерархической кластеризации [6], состоят из двух шагов. Первый шаг — каждый вектор множества данных представляется отдельным кластером [6]. На втором шаге находятся два ближайших кластера, которые сливаются друг с другом. Этот шаг повторяется до тех пор, пока не останется один кластер, объединяющий всю обучающую совокупность. Преимуществом этого алгоритма является то, что кластеры организованы иерархически, что позволяет выбирать степень обобщения. Для агрегирования информации можно использовать так называемый алгоритм LSH. Для этого для заранее заданной метрики расстояния определяется хеш-функция, такая что если:

$d(x, y) \leq R$, то $h(x) = h(y)$ с вероятностью не менее P,

$d(x, y) \geq C * R$, то $h(x) = h(y)$, с вероятностью не более P,

где P, R и C– заданные величины.

Рассмотренные выше модели кластеризации и поиска в информационном пространстве сами по себе не позволяют осуществить прогнозирование новостных событий. Однако они решают ряд очень важных вспомогательных задач: происходит очистка и отображение в единое формализованное пространство плохо измеримых и гетерогенных данных (социально – экономические, геологические, климатические, астрономические и т.д.).

Разработка методик прогнозирования событий представляет значительный интерес и в этой связи можно упомянуть работу [7] Asela Gunawardana, Christopher Meek, Puyang Xu. A Model for Temporal Dependencies in Event Streams, в которой представлена модель для временных зависимостей в потоках событий. Авторы этой статьи вводят кусочно-постоянную модель интенсивности событий для изучение временных зависимостей в потоках событий, при этом используется Байесовский подход и распределение Пуассона к описанию выборки важности будущих событий. Это позволило разработать алгоритм, позволяющий изучать нелинейные временные зависимости для предсказания будущих событий с использованием деревья решений.

В настоящее время продолжается поиск моделей и методов возможного прогнозирования новостных событий в информационном пространстве, и на наш взгляд значительный прогресс в данной области может быть достигнут, если удастся установить, что процессы в информационном пространстве обладают свойствами самоподобия.

Структура и директор информационного новостного пространства, самоподобие процессов

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 обладают 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. Информационное пространство, так же как и реальное физическое, обладает памятью и способностью к самоорганизации.
5. Не смотря на то, что прогнозируемое новостное событие является априори неизвестным, мы можем искусственно вербально описать его в информационном пространстве, создавая некоторый новостной образ, а затем построить динамическую модель возможной трансформации уже существующих кластеров к заданному образу события.

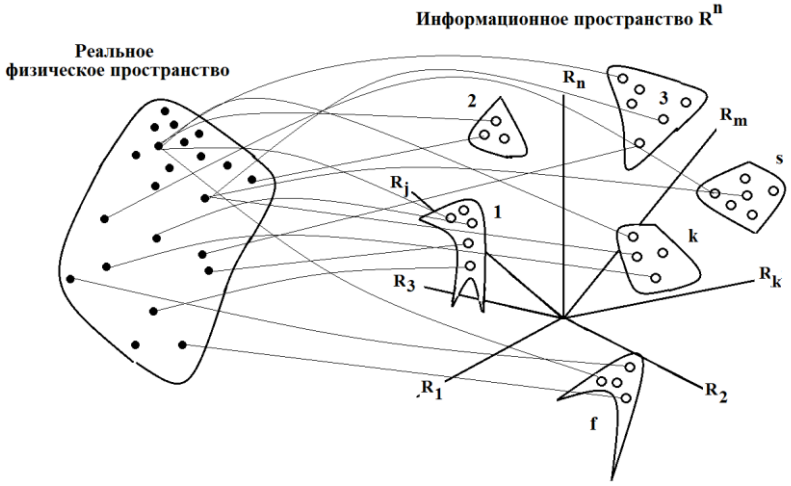

Рис. 1. Отображение событий из реального физического пространства в информационное и образование новостных кластеров

На рисунке 1 показано как происходит отображение событий из реального, физического пространства в информационное, с размерностью R^n и их кластеризация по новостным группам (1,2,3, k, f,).

Анализ данных, описывающих информационное пространство показывает, что можно выделить две взаимосвязанные подсистемы:

- “медленную”, в которой обрабатывается информация, медленно изменяющуюся или почти не изменяющаяся с течением времени (научные, культурные, религиозные, художественные и т.д. знания);
- “быструю”, к которой можно отнести информацию политического, экономического, спортивного и т.д. характера.

Каждая из подсистем содержит определенные наборы кластеров информации, со своими векторами задающими их положение.

В рамках нашей модели мы предлагаем ввести понятие директор. Директором будем называть условное направление в информационном пространстве, которое определяется взаимным усреднением направлений всех рассматриваемых векторов. 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. Используя данную методику можно получить для информационного пространства “медленный” и “быстрый” директоры.

При наличии в зависимости от времени величин рассматриваемых директоров самоподобия, для его описания может быть применена теория, суть которой заключается в том, что непрерывный стохастический процесс X(t), считается статистически самоподобным с параметром Харста (Hurst) $\boldsymbol{H}$ ($0{,}5 \le \boldsymbol{H} \le 1$), если для любого положительного числа **a**, случайные процессы X(t) и $\mathbf{a}^{-H}X(at)$ будут иметь одинаковые статистические свойства. Значение H=0.5 показывает отсутствие самоподобности, а значения $\boldsymbol{H}$ близкие к единице показывают её большую степень.

Следует отметить, что теория самоподобия широко применяется для исследования информационных процессов, в частности поведении трафика при передачи данных [8].

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: $R(X_m) = R(X_n)$, а для асимптотически самоподобных: $R(X_m) \to R(X_n)$, при $m \to n$. Дисперсия для обоих классов процессов определяется одинаковым образом: $D(X_m) = {D(X_n)}/{m^{\beta}}$, где β – параметр самоподобия ($0<\beta<1$), связанный с параметром Харста следующим соотношением: $\beta = 2(1-H)$, а m – величина блока разбиения исходных данных.

Коэффициент Харста находится по зависимости логарифма среднего значения дисперсии от логарифма величины блоков m разбиения исходной выборки данных. При наличие 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 Харста (Hurst) H.

Важным вопросом является решение задачи выбора параметров информационных процессов, в поведении которых можно определить самоподобие. В частности при исследовании самоподобия процессов в информационном новостном пространстве можно определять зависимость от времени угла между “быстрым” и “медленным” директорами, и определить интервал или период самоподобия процессов в информационном пространстве.

Кроме того для определения самоподобия процессов, приводящих к реализации интересующего информационного события можно исследовать зависимость от времени углов между каждым из директоров и вектором, описывающим в информационном пространстве данное прогнозируемое событие.

Вектор прогнозируемого события и тренды процессов в информационном пространстве

Проведем в какой-то момент времени t кластеризацию информационного пространства с размерностью R^n по различным новостным событиям и определить значения величин векторов (z_1, z_2, z_3, z_k, z_j,), задающих положение центров этих кластеров в данный момент времени (см. рис. 2). Отметим, что соответствующие компоненты R_1, R_2, R_3, R_m, R_n, векторов z_1, z_2, z_3, z_k, z_j, являются не отрицательными.

Далее проведем вербальное описание прогнозируемого новостного события и таким образом зададим его вектор $\mathbf{X_{bs}}$ в информационном пространстве (см. рис. 2).

Поскольку мы предполагаем, что в информационном пространстве уже имеются некоторые данные о предстоящем новостном событии, то должно существовать и отображение имеющихся групп новостных событий, на событие которое мы пытаемся прогнозировать. Априори мы не можем точно указать математическую формулу такого отображения, но оно должно иметь однозначный характер и сохранять причинно-следственные связи и последовательности событий. Однако при этом возможно искажение масштабов интервалов между событиями. Выберем в качестве отображения, нахождение проекций x_j векторов, задающих положение центров информационных кластеров в данный момент времени z_1, z_2, z_3, z_k, z_j, на направление вектора $\mathbf{X_{bs}}$, определяющего появление прогнозируемого события. В данном случае мы предполагаем, что данные проекции являются источником формирования редкого события на оси X, задающей его появление. Каждая из проекций x_k определяется как произведение величины соответствующего вектора z_k и косинуса угла между направлениями векторов z_k и $\mathbf{X_{bs}}$; $x_k=z_k*\cos(\alpha_k)$.

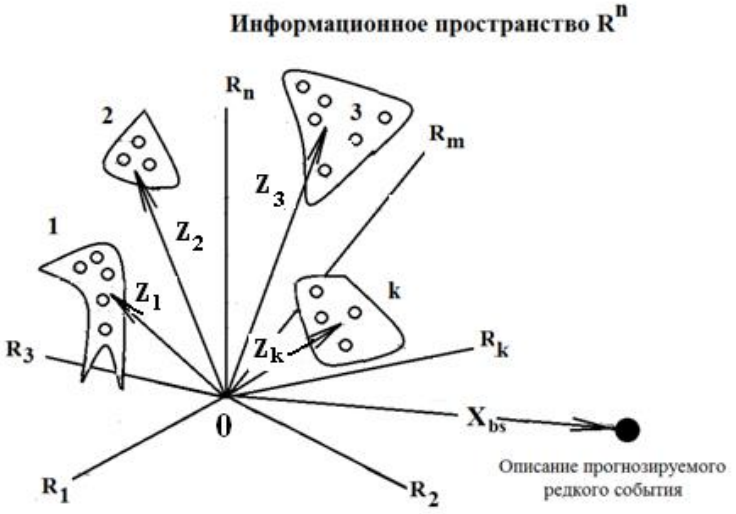

Рис. 2. Кластеризация информационного пространства по новостным группам и определение в нем положения редкого события

Спустя некоторый интервал времени (назовем его интервал измерения τ_0) величины векторов, задающие положения центров информационных кластеров изменяются на какие-то случайные значения Δ_j (j – обозначает рассматриваемый вектор). На рисунке 3 в качестве примера показано изменение за время τ_0 для векторов z_2 и z_3, задающих положение центров кластеров номер 2 и номер 3. Вектора z_{02} и z_{03} определяют положение центров новостных кластеров в информационном пространстве в момент времени t, а вектора z_2 и z_3 спустя интервал времени τ_0 (момент времени $t+\tau_0$). В данных случаях $\Delta_2=z_2-z_{02}$ и $\Delta_3=z_3-z_{03}$. Аналогичным образом определяются изменения положения центров для всех кластеров в информационном пространстве.

Величины x_{02}, x_{03}, x_2 и x_3 будут задавать значения соответствующих проекций векторов, определяющих положение центров новостных кластеров 2 и 3, на направление прогнозируемого события, в моменты времени t и $t+\tau_0$.

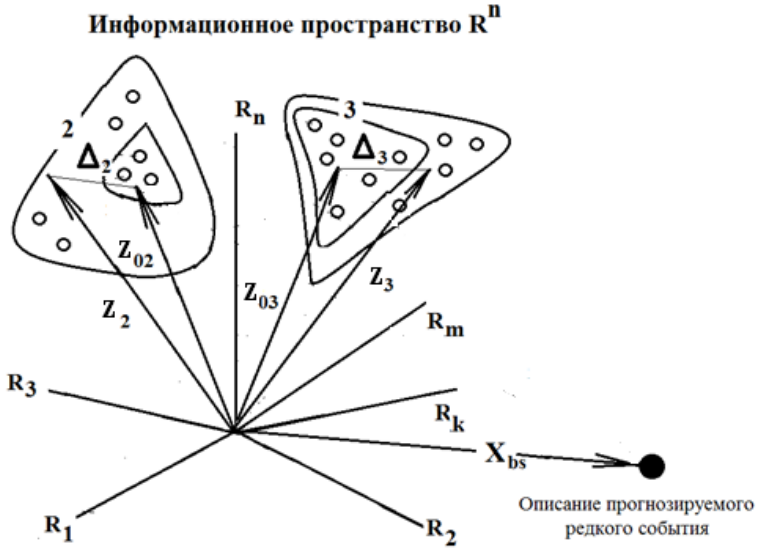

Рис. 3. Изменение положения центров кластеров за интервал времени измерения τ_0

Следует отметить, что некоторые величины проекций x_k могут оказаться больше предыдущих значений x_{0k} (для одной и той же группы новостных данных), а некоторые меньше, т.е. в информационном пространстве одновременно сосуществуют два тренда поведения. Один на увеличение значений проекций, другой на их уменьшение, что необходимо учесть в рамках разрабатываемой модели. Введем для любого момента времени понятие среднего значения $\overline{x_i}$ всех величин проекции векторов, задающих положение центров новостных кластеров в информационном пространстве на направление оси прогнозируемого события. Для группы из K новостных кластеров в момент времени t среднее значение $\overline{x_t}$ определяется следующим образом: $\overline{x(t)} = \frac{\sum_{j=1}^{K} x(t,j)}{K}$, где $x(t,j)$ обозначают соответствующие значения проекций векторов, задающих положение центров новостных кластеров, на направление оси прогнозируемого события, в данный момент времени t. Спустя интервал времени τ_0: $\overline{x(t+\tau_0)} = \frac{\sum_{j=1}^{K} x(t+\tau_0,j)}{K}$. Чтобы учесть тренды увеличения и уменьшения значений величин проекций векторов, задающих положение центров новостных кластеров, на направление оси прогнозируемого события можно поступить следующим образом. На основании анализа значений $x(t,j)$ и $x(t+\tau_0,j)$ разделим множество $x(t,j)$ на две подгруппы, в одной $x(t,j)_{\text{I}}$ будут все новостные кластеры, для которых за интервал времени τ_0 произошло уменьшение значений величин проекций $x(t+\tau_0,j)$ (обозначим число таких кластеров как R), а во второй $x(t,j)_{\text{II}}$ – увеличение (обозначим число таких кластеров как K-R), и найдем для каждой из них средние значения $\overline{x(t,j)}_I = \frac{\sum_{j=1}^{R} x(t,j)_I}{R}$ и $\overline{x(t,j)}_{II} = \frac{\sum_{j=1}^{K-R} x(t,j)_{II}}{K-R}$) проекций векторов, задающих положение центров этих новостных кластеров. Далее, мы предлагаем следующий подход к учету трендов увеличения и уменьшения значений величин проекций векторов, задающих положение центров новостных кластеров, на направление оси прогнозируемого события. Поскольку при учете трендов имеет смысл говорить об усредненных величинах, то будем рассматривать переход за интервал времени τ_0 в точку $\overline{x(t+\tau_0)}$ из точки $\overline{x(t,j)}_I$, которая находится на оси прогнозирования события правее точки $\overline{x(t+\tau_0)}$ и точки $\overline{x(t,j)}_{II}$ которая находится левее $\overline{x(t+\tau_0)}$. Сами по себе переходы являются случайными событиями, а их величины можно определить следующим образом: $\xi(t) = \overline{x(t,j)}_I - \overline{x(t+\tau_0)}$ и $\varepsilon(t) = \overline{x(t+\tau_0)} - \overline{x(t,j)}_{II}$. После следующего шага τ_0 определяем новые значения $\xi(t+\tau_0)$ и $\varepsilon(t+\tau_0)$: $\xi(t+\tau_0) = \overline{x(t+\tau_0,j)}_I - \overline{x(t+2\tau_0)}$ и $\varepsilon(t+\tau_0) = \overline{x(t+2\tau_0)} - \overline{x(t+\tau_0,j)}_{II}$ и т.д.

На любом шаге **n** величины $\xi_{t+n\tau}$ и $\varepsilon_{t+n\tau}$ могут принимать случайные (или почти случайные) значения, однако в их поведении могут наблюдаться характерных особенностей (например, зависимости $\xi_{t+n\tau}$ и $\varepsilon_{t+n\tau}$ от времени, могут обладать самоподобием).

Анализ самоподобия в трендах увеличения и уменьшения может позволить выявить наличие периодичности в приближении или удалении от прогнозируемого события в информационном пространстве.

Методика экспериментальной проверки предлагаемых моделей

Экспериментальная проверка разработанных нами моделей может быть проведена на большом массиве текстовых данных с использованием следующего алгоритма:

a) Проводим представление совокупности текстовых данных в информационном пространстве с помощью процедуры сопоставления каждого признака с функцией-индикатором данного слова, и осуществляем в какой-то момент времени t=0 кластеризацию по различным новостным группам, используя алгоритм k-средних, Scatter-Gather, BIRCH или алгоритмы иерархической кластеризации. Выделяем быстро и медленно изменяющиеся подсистемы данных в информационном пространстве, и определяем для них направления директоров.
b) Задаем вектор события в информационном пространстве посредством описания его вербального образа. Величина вектора прогнозируемого события задает на оси этого события порог его достижения (***I***). Определяем углы между директорами и вектором прогнозируемого события.
c) Проводим отображение векторов, определяющих положение центров новостных кластеров на направление оси, задаваемой вектором прогнозируемого события момент времени t=0 и находим их среднее значение x_0.
d) Спустя некоторый интервал времени измерения (назовем его τ_0) определяем новые вектора, задающие положение центров новостных кластеров и их отображения на ось прогнозируемого события. Разделяем отображения на две группы. В первой будут все вектора, для которых значения отображений увеличились, во второй для которых уменьшились, по сравнению с предыдущими значениями. Находим средние значения по группам и определяем значения величин трендов увеличения (ε) и уменьшения (ξ) по отношению к начальному состоянию x_0. Величины ε и ξ определяются разностью текущего среднего значения по группе и предыдущего общего состояния x_0. Выделяем быстро и медленно изменяющиеся подсистемы данных в информационном пространстве, определяем для них новые направления директоров и углы между директорами и вектором прогнозируемого события.
e) Через новый интервал времени измерения τ_0 повторяем процедуры, описанные в пунктах b) – e).
f) По полученной зависимости от времени углов между директорами и вектором прогнозируемых событий определяем наличие или отсутствие самоподобия в процессах, протекающих в информационном пространстве, и при его наличии определяем параметры самоподобия (период или интервал, что важно для прогнозирования интересующего события).

Заключение

Для описания структуры новостного информационного пространства введено понятие директора - условной оси, положение которой определяется усреднением направлений векторов, задающих положение центров всех новостных кластеров. Рассмотрена разработанная авторами модель прогнозирования новостных событий на основе динамики приближения состояния информационного пространства к точке события. Для описания динамики достижения новостного события вводятся понятия понижающего и повышающего трендов изменения состояния информационного пространства.

Список литературы

1. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press, 2007.
2. Allan J. Topic detection and tracking: event-based information organization. Kluwer Academic Press, 2002.
3. Beringer J., Hullermeier E. Online Clustering of Parallel Data Streams // Data & Knowledge Engineering. 2006. No. 58. pp. 180-204.
4. Costa G., Mango G., and Ortale R. An incremental clustering scheme for data de-duplication // Data Mining and Knowledge Discovery. Jan 2010. Vol. 20. No. 1. pp. 152-187.
5. Moerchen F., Brinker K., and Neubauer C. Any-Time Clustering of High Frequence News Streams // 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007. pp. 23-31.
6. Zhang Y.J., Liu Z.Q. "Self-Splitting competitive learning: a new on-line clustering paradigm," IEEE Transactions on Neural Networks, No. 13(2), 2002. pp. 369-380.
7. Asela Gunawardana, Christopher Meek, Puyang Xu. A Model for Temporal Dependencies in Event Streams. URL: http://ceur-ws.org/Vol-962/paper08.pdf
8. Clegg R.G. A practical guide to measuring the hurst parameter. // Computing science technical report. – 2005. – № CS–TR–916. – P. 125-138.