Распределение документов по степени релевантности на основе мультифрактальных свойств

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

Ландэ Д.В.
ИЦ «ЭЛВИСТИ»
dwl@visti.net

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

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

Аннотация

В рамках данной работы были исследованы распределения документов в сетевых информационных потоках по двум степеням релевантности относительно поисковых запросов. Обосновывается мультифрактальная природа распределений документов по степеням релевантности. Описаны основные алгоритмы, применяемые в процессе исследований.

Обоснована возможность использования мультифрактальных свойств для верификации репрезентативности выборок релевантных документов, которые можно рассматривать в качестве одного из вариантов представления конечных результатов поиска или документальных корпусов для дальнейших исследований.

1. Введение

В связи с наблюдающимся в последние годы ростом объемов и темпов обновления сетевой информации приобретает актуальность задача изучения статистических свойств сетевых документальных массивов [1, 2].

В Манифесте РОМИП [3] сказано: «Непрерывная эволюция информационного пространства и применение методов поиска в новых контекстах мотивирует актуальность дальнейших исследований в области теории информационного поиска». Вместе с тем, сегодня еще не произошло сколько-нибудь заметных изменений в понимании центральной задачи поиска – предоставить пользователю ту информацию, в которой он заинтересован, хотя полный список задач, ассоциируемых с ней, существенно расширился.

Основная проблема информационного поиска в настоящее время, по-видимому, связана с большими объемами выходных данных современных поисковых систем, из-за чего пользователь зачастую не в состоянии их обработать. Традиционные способы решения этой проблемы используют три основные технологии: ранжирование [4-7], кластеризацию [8, 9, 10, 11] и фильтрацию [12].

Вместе с тем, представляется, что основная проблема поиска принципиально не может быть решена путем усовершенствования традиционной технологии. Более того, такие усовершенствования технологии поиска на самом деле усугубляет ситуацию, в частности, увеличивая объемы релевантных выборок. Становится очевидным, что многие стандартные критерии эффективности поиска уже не могут применяться без известных оговорок. Вместе с тем, остается задача получения репрезентативных выборок, отражающих основные тенденции, соответствующие информационному запросу.

Для успешного решения этой, пусть частной задачи, необходимо достаточно хорошо представлять себе возможные механизмы генерации массивов релевантных документов, а также знать основные свойства этих массивов.

Сложность и многоплановость задачи поиска предполагает активное использование современных теоретических подходов, позволяющих более глубоко понять специфику данной предметной области, в том числе подходы, применяемые в теории детерминированного хаоса [13-16]. Одним из таких направлений является исследование фрактальных свойств документальных массивов. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, и могут рассматриваться как стохастические фракталы [13, 17]. Известно, что все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Ципфа, могут быть обобщены именно в рамках теории стохастических фракталов [18].

Применение фрактальных и мультифрактальных методов, рассматриваемых в настоящем исследовании, представляется тем более интересным, что сетевые документальные массивы в этом плане остаются малоизученными [2].

2. Объекты и методы исследования

Основными теоретическими методами являются методы математической статистики и фрактальный анализ числовых рядов.

Экспериментальные же методы были применены для изучения поведения реальных наборов документов, релевантных некоторым поисковым запросам, полученных в рамках технологии InfoStream [19], а также тестовые наборы, предоставленные ООО «Яндекс».

В предлагаемой работе исследовались распределения мер релевантности документов, определяемые двумя различными способами (нормированной и ненормированной по длине документа).

На основании обработки данных наблюдений были получены значения различных статистических показателей соответствующих рядов, а также показано, что они обладают мультифрактальной природой [20-22]. Показано, что мультифрактальные характеристики позволяют оценивать репрезентативность отдельных выборок на основе мультфрактальных свойств.

В классической задаче информационного поиска под релевантностью, как известно, понимается формальное соответствие документа из набора данных, в котором осуществляется поиск, информационному запросу пользователя [23]. В ряде задач возникает необходимость подходов к оценке отобранных документов, предполагающих использование количественной меры соответствия документов запросам, которая описывалась бы, достаточно широким спектром значений. Величину, используемую с такой целью, уместно назвать ***мерой или степенью релевантности***.

В настоящей работе с целью моделирования рассматривались две простейшие меры релевантности (однако предложенный подход не предполагает ограничений и может быть применен к другим мерам релевантности). Первая мера релевантности определяется частотами вхождения поисковых терминов из запроса в документ и описывается следующим соотношением:

$$R_F = \sum_k N_k \,, \qquad (1)$$

где N_k – число вхождений k-го термина в данный документ.

Вторая мера релевантности включает в себя нормировку на длину документа L:

$$R_N = \frac{1}{L}\sum_k \ln(N_k + 1)\,. \qquad (2)$$

Различие отдельных мер релевантности с точки зрения эффективности поиска в свое время широко обсуждалось в литературе [24-26]. Существенным является то, что мера (1) является однопараметрической (определяется частотами поисковых терминов), а мера (2) – двухпараметрической (определяется частотами поисковых терминов и длиной документа).

В ходе исследований обрабатывались информационные массивы, содержащие сообщения онлайновых СМИ. В дальнейшем изложении как примеры рассматриваются следующие корпусы: 1) массив из 5000 документов, опубликованных за 3 суток с 1 по 3 декабря 2006 г., предположительно по банковской тематике, удовлетворяющих запросу “банк”; 2) массив из 7380 документов, опубликованных за 4 суток с 1 по 4 марта 2007 г., удовлетворяющих запросу “Ющенко ИЛИ Янукович ИЛИ Тимошенко”; 3) массив из 279 документов из новостной дорожки, предоставленных компанией «Яндекс» для проведения исследований, относящихся ко времени отставки Шеварнадзе с поста Президента Грузии, удовлетворяющих запросу “Грузия”. Заметим, что многие статистические, фрактальные и мультифрактальные характеристики 3-го корпуса выглядит неубедительными ввиду его нерепрезентативности, однако авторы приводят их с целью возможного сравнения с результатами других исследований.

На рис. 1 (а, б и в) приведены зависимости меры релевантности R_N и R_F от номера документа. При этом документы ранжированы по мере релевантности R_N, соответственно, гладкая кривая представляет собой распределение меры релевантности R_N. Каждому номеру документа соответствует также некоторое значение меры релевантности R_F. Как можно убедиться, эти зависимости обладают существенно различным поведением.

Однако, из приведенных графиков видно, что между данными зависимостями существуют некоторые корреляции. Например, существенно нелинейному участку полученных зависимостей с максимальными значениями R_N соответствует область с минимальными значениями R_F. Иными словами, мы имеем основания утверждать, что в реальных документальных массивах существует устойчивая нетривиальная взаимосвязь между частотами слов и размерами документов. По-видимому, статистически значимыми являются ситуации, когда многократное повторение термина встречается в документах, размеры которых значительно превышают среднюю по выборке длину. Напротив, малые по объему сообщения приводят к появлению в зависимости R_N резких пиков, не характерных для поведения R_F.

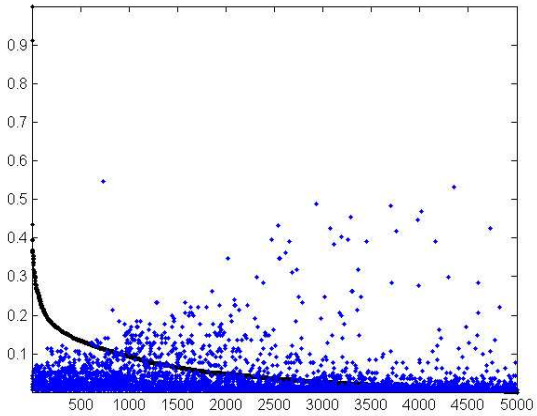
На рис. 2 представлены данные, соответствующие первому корпусу, но в логарифмическом масштабе по оси Y. Данные двух других рассматриваемых документальных корпусов имеют аналогичную структуру. На рис. 2. показано, что сортированная по мере релевантности зависимость R_N содержит линейный центральный участок, соответствующий обобщенному закону Ципфа[1]. Интересно, что отклонение от закона Ципфа имеет место не только в «хвосте» зависимости, что неоднократно отмечалось в различных лингвостатистических исследованиях [24], но и на начальном участке.

Очевидно, что частоты поисковых терминов в документе образуют дискретный набор. Поэтому сортированная зависимость однопараметрической меры R_F на самом деле представляет собой набор линейных участков, отвечающих каждому значению частоты появления терминов из запроса в документах.

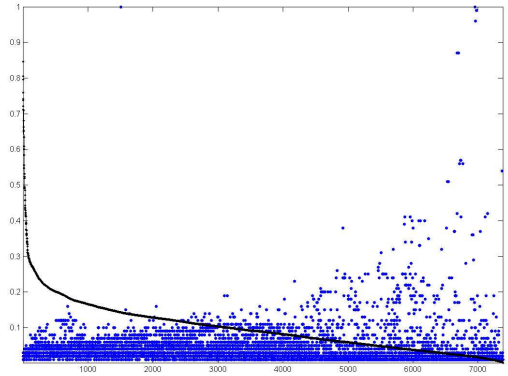
Нетрудно заметить, что зависимость R_F, изображенная на рис. 2, в нижней части имеет явную регулярность, обусловленную

[1] Закон Ципфа (предложенный для описания поведения рангов слов в документе), применяемый к распределению произвольного параметра.

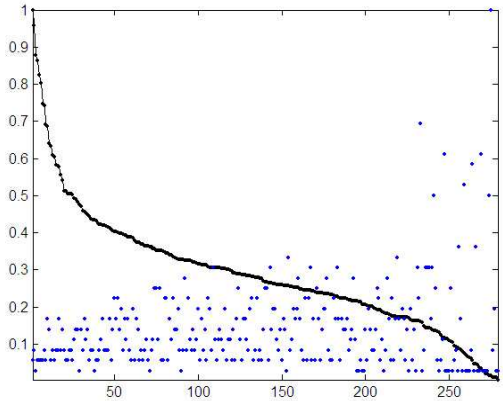
дискретностью частот появления поисковых терминов. Этим значениям частот соответствуют отчетливые горизонтальные линии в нижней части графика. Однако при более высоких частотах двухпараметрическая зависимость R_F теряет регулярность, и в ее поведении появляются особенности, характерные для детерминированного хаоса.



а) – запрос «банк»



б) – запрос «Ющенко ИЛИ Янукович ИЛИ Тимошенко»



в)) – запрос «Грузия»

Рис. 1. Значения мер релевантности (ось Y) документов по двум критериям. Документы (ось X) ранжированы по значениям R_N

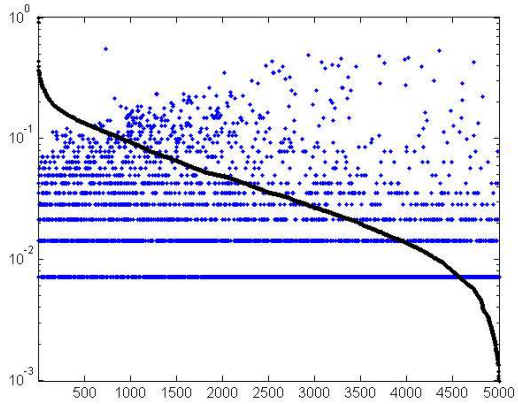
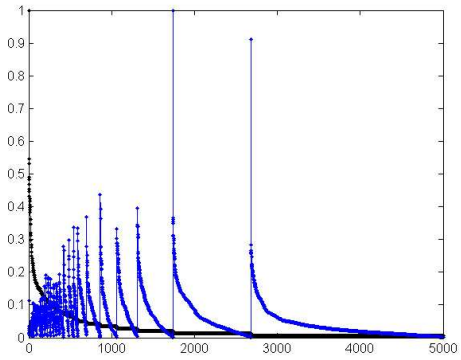


Рис. 2. Значения мер релевантности документов по двум критериям в логарифмическом масштабе

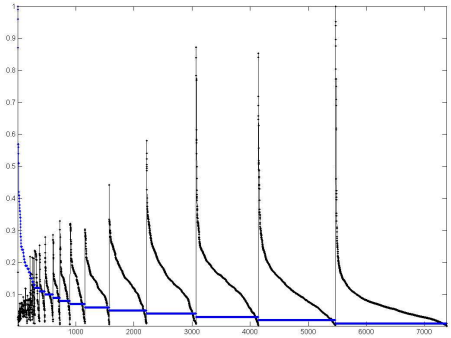
Построим теперь зависимость, при которой документы отсортированы по значениям частот поисковых терминов R_F. Зависимость меры релевантности R_F от номеров документов представляет собой набор горизонтальных участков, отвечающих определенным значениям частот. При этом наборы документов, принадлежащие каждому такому участку, дополнительно сортируются по значениям меры релевантности R_N, поскольку такое представление данных обладает большей наглядностью.

Полученные таким образом результаты приведены на рис. 3. Учитывая то, что частоты слов в документах и длины самих документов в документальных потоках распределены достаточно случайно, можно ожидать, что распределение двухпараметрической характеристики по однопараметрической будет, в некотором приближении, описываться распределением Пуассона, а именно, максимальные значения R_N для локальных отрезков R_F = const будут распределены по Пуассону. И, действительно, мы видим нечто подобное на рис. 3. Нкоторые отклонения от регулярности в 3-м случае обусловлены недостаточной репрезентативностью выборки.

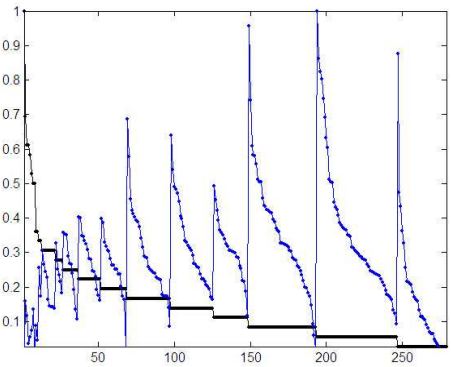
В случае, когда ставится задача оптимизации результатов поиска по некоторому набору критериев (например, по двум мерам релевантности), очевидно, что решение задачи может быть получено путем отбора подмножеств документов из локальных интервалов, лежащих справа от точек $\max(\{R_N\}_k)$, принадлежащих центральным участкам на рис. 3 а), б). Длины этих интервалов могут определяться с учетом ограничений на полный объем выборки. Действительно, в начальном участке графика большие значения R_F компенсируются малыми значениями R_N, а «хвосты» обоих распределений по понятным причинам интереса не представляют.



а) – запрос «банк»



б) – запрос «Ющенко ИЛИ Янукович ИЛИ Тимошенко»



в) – запрос «Грузия»

Рис. 3. Значения мер релевантности документов по двум критериям. Документы ранжированы по значениям R_F

Заметим, что данное обстоятельство открывает нам интересный аспект, непосредственно связанный с проблемой оптимизации поиска. Процесс построения наборов данных, приведенных на рис. 3 а), б), в определенном смысле может играть роль графического метода решения задачи оптимизации выборок, получаемых с помощью многопараметрических критериев. Мы проиллюстрировали его на простейшем примере, однако он может быть без труда обобщен на произвольные задачи, в том числе и с большим числом параметров. Таким образом, статистические исследования, проведенные еще на предварительном этапе, позволили найти подход к решению важной практической задачи.

3. Фрактальные характеристики потока документов

Как известно, возникновение детерминированного хаоса в динамике объектов тесно связано с наличием фрактальных свойств.

Наиболее интересным объектом для изучения фрактальных и мультифрактальных свойств распределения документов по степени релевантности по мнению авторов оказались распределения мер релевантностей R_F в последовательностях документов, ранжированных по R_N, которые проиллюстрированы на рис. 1 и 2. ***Именно эти числовые ряды рассматривались как исходные данные для всех последующих экспериментов и исследований, проводимых в рамках данной работы.***

3.1 Размер страницы Метод DFA

Одним из универсальных подходов к выявлению самоподобия основывается на методе DFA (Detrended Fluctuation Analysis) [20, 27, 28] – универсальном методе обработки временных рядов. Этот подход представляет собой вариант дисперсионного анализа, позволяющий исследовать эффекты длительных корреляций в нестационарных рядах. В рамках алгоритма DFA анализируется среднеквадратическая ошибка линейной аппроксимации в зависимости от размера аппроксимируемого участка. Этот метод в был применен к рядам значений релевантностей.

В рамках этого алгоритма вначале осуществляется приведение данных рядов релевантностей к нулевому среднему (вычитание среднего значения $<\xi>$ из временного ряда ξ_i и строится случайное блуждание:

$$y(k) = \sum_{i=1}^{k} [\xi_i - <\xi>].$$

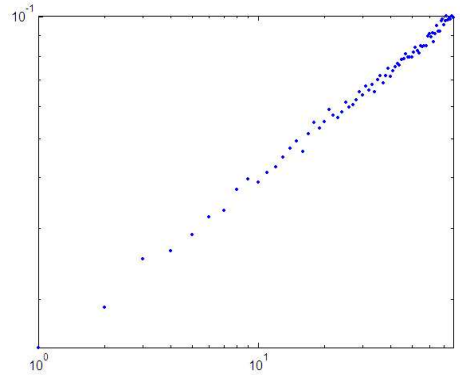
Затем ряд значений $y(k)$, $k = 1,...,N$ разбивается на неперекрывающиеся отрезки (участки) длины n, в пределах каждого из которых определяется уравнение прямой, аппроксимирующей последовательность $y(k)$.

Далее вычисляется среднеквадратическая ошибка линейной аппроксимации $F(n)$ и соответствующие расчеты проводятся в широком диапазоне значений n. Считается, что зависимость $F(n)$ часто имеет степенной характер $F(n) \sim n^{\alpha}$, а наличие линейного участка в двойном логарифмическом масштабе $\lg F(\lg n)$ позволяет говорить о существовании скейлинга. На практике величина α (называемая скейлинговой экспонентой DFA-метода) может отличаться для разных n, что свидетельствует об изменении свойств скейлинга при увеличении масштаба. В данной ситуации целесообразно проводить анализ локальных показателей α.

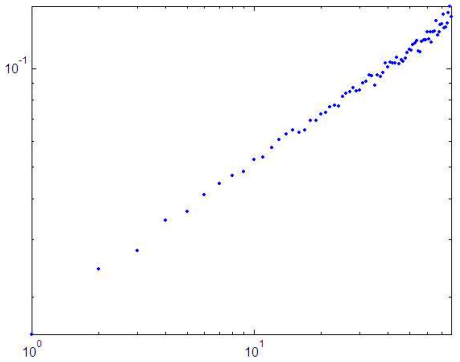
На рис. 4. представлена зависимость $F(n)$ от длины участков аппроксимации в двойном логарифмическом масштабе. Наличие линейного тренда на этих графиках (даже в случае 3-го

корпуса) позволяет говорить о наличие локального скейлинга в исследуемых числовых рядах.

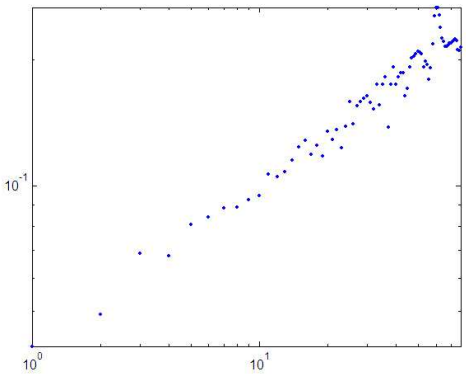
Численные значения скейлинговой экспоненты α характеризуют различные типы коррелированной динамики, если $\alpha \neq 0.5$ и некоррелированное поведение при $\alpha = 0.5$. Например, диапазон $0 < \alpha < 0.5$ соответствует антикорреляциям; $0.5 < \alpha < 1$ определяет коррелированную последовательность. В исследуемых случаях параметр α попадал именно в этот диапазон.



а)) – запрос «банк»



б) – запрос «Ющенко ИЛИ Янукович ИЛИ Тимошенко»



в) – запрос «Грузия»

Рис. 4. Зависимость F(n) ряда наблюдений (ось Y) от длины участка аппроксимации n (ось X)

3.2 Фактор Фано

Для подтверждения самоподобия числовых рядов принято использовать еще один показатель – индекс разброса дисперсии (IDC), так называемый, фактор Фано [29]. Эта величина определяется как отношение дисперсии исследуемого числового ряда на заданном окне наблюдений k к соответствующему математическому ожиданию:

$F(k) = \sigma^2(k)/m(k)$.

Для самоподобных числовых рядов выполняется:

$F(k) \cong Ck^{\alpha}$,

где C и α – константы. При этом в случае фрактальной структуры числовых рядов выполняется соотношение:

$\alpha = 2H - 1$,

где H – показатель Херста (см. п. 3.4), непосредственно связанный с фрактальной размерностью.

На рис. 5 приведен график значений $F(k)$ в двойном логарифмическом масштабе для запроса *«Ющенко ИЛИ Янукович ИЛИ Тимошенко»*. При этом рассматривались окна наблюдений от 2 до 16, соответственно, скейлинговый коэффициент α оказался равным ~ 0.538, и коэффициент H равным ~ 0.77, что вполне согласуется с данными, которые были получены при непосредственном вычислении параметра Херста. Следует отметить, что при значениях окон наблюдений, превышающих 20 и доходящих до $N/10$ (N – размер документального корпуса), наклон аппроксимирующей линии значительно уменьшается, что свидетельствует об отсутствии у исследуемых числовых рядов «долговременной памяти».

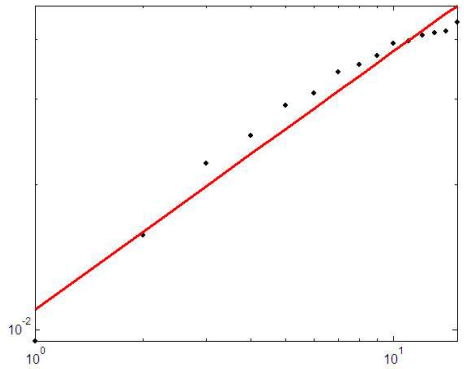


Рис. 5. Зависимость фактора Фано (ось Y) от ширины окна наблюдений k (ось X)

3.3 Корреляционная функция

Как известно, коэффициенты корреляции для ряда измерений рассчитываются по формуле:

$R(k) = < (\xi_{k+t} - <\xi>)(\xi_k - <\xi>>/\sigma^2$,

где $R(k)$ - коэффициент корреляции; X_k - ряд измерений; $<\xi>$ – математическое ожидание, σ^2 – дисперсия.

На рис. 6 приведены значение коэффициентов корреляции исследуемого ряда и ряда, который был получен из исходного путем его псевдослучайного перемешивания. Графическое представление коэффициентов корреляции для исследуемых рядов наблюдений и «перемешанных» рядов еще раз подтверждает гипотезу о существенной

автокорреляции исследуемых числовых рядов распределений релевантностей. У «перемешанного» ряда корреляция убывает быстрее, и ее значения как правило меньшие по сравнению с исходным рядом.

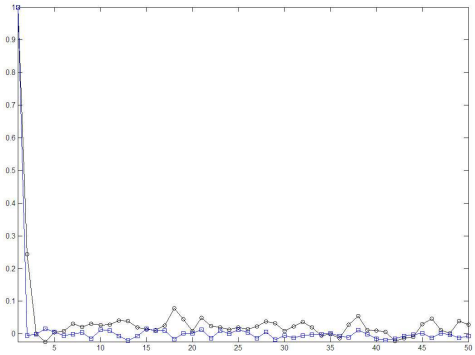


Рис. 6. Коэффициенты корреляции ряда наблюдений (ось Y) в зависимости от окна наблюдений k (ось X): o *– исследуемый ряд наблюдений;* □ *- перемешанный ряд наблюдений*

3.4 Показатель Херста

Показатель Херста (*H)* связывают с коэффициентом нормированного размаха (*R/S*), где *R* - вычисляемый определенным образом «размах» соответствующего временного ряда, а *S* - стандартное отклонение.

Показатель Херста вычисляется по следующему алгоритму. Сначала вычисляется среднее значение измеряемой переменной:

Показатель Херста вычисляется по следующему алгоритму. Сначала вычисляется среднее значение измеряемой переменной $<\xi>_N$ и стандартное отклонение *S*.

Затем рассчитывается накопившееся отклонение ряда измерений $\xi(t)$ от среднего $<\xi>_N$:

$$X(t,N) = \sum_{u=1}^{t} (\xi(u) - <\xi>_N) \cdot$$

После этого определяется разность максимального и минимального накопившегося отклонения, которая и называется «размахом»:

$$R(N) = \max_{1 \le t \le N} X(t,N) - \min_{1 \le t \le N} X(t,N).$$

Для фрактальных числовых рядов справедливо:

$$R/S = (N/2)^H,$$

где *H* - показатель Херста.

На рис. 7. изображена зависимость нормированного размаха *(R/S)* от размерности подмножества документов в двойной логарифмической шкале. Мы видим, что его значение хорошо аппроксимируется прямой. При этом показатель Херста (рис. 8) для трех рассматриваемых рядов измерений составлял ~ 0.78, ~ 0.67 и ~ 0.68, что соответствует хаусдорфовым размерностям ~ 1.22, ~ 1.33 и ~1.32.

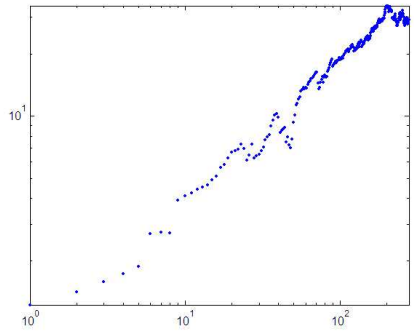
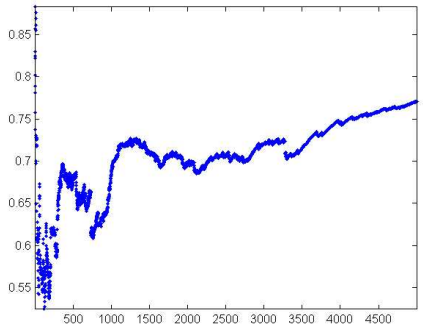
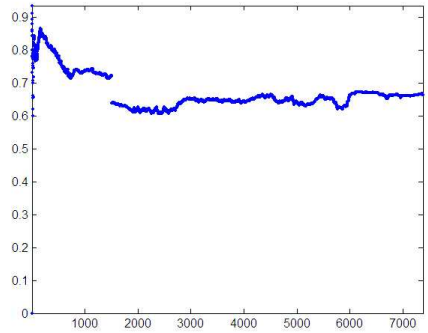


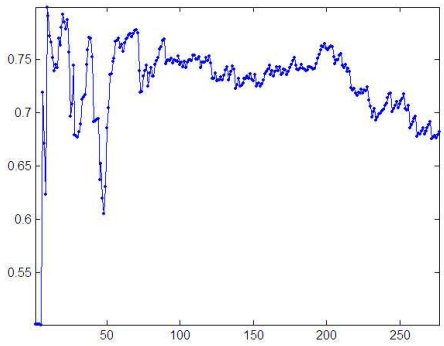
Рис. 7. Значения показателя нормированного размаха (ось Y) в зависимости от объема исследуемого массива (ось X)



а) – запрос «банк»



б)) – запрос «Ющенко ИЛИ Янукович ИЛИ Тимошенко»



в) – запрос «Грузия»

Рис. 8. Значения показателя Херста (ось Y) в зависимости от объема исследуемого массива (ось X) в двойной логарифмической шкале

Таким образом, проведенные исследования числовых рядов распределений релевантностей подтвердили предположение о самоподобии и

фрактальности.

4. Мультифрактальные характеристики

4.1 Мультифрактальный формализм

Наиболее общее описание природы самоподобных объектов позволяет дать теория мультифракталов, характеризуемых бесконечной иерархией размерностей, и позволяющая отличить однородные объекты от неоднородных [14]. Концепция мультифрактального формализма [19-21, 30-32] дает эффективный инструмент для изучения и количественного описания широкого многообразия сложных систем. В данной работе изучается возможность мультифрактального описания ряда релевантностей документов R_F, ранжированных по мере релевантности R_N.

Носителем мультифрактальной меры является множество L – объединение фрактальных подмножеств L_α. Т.е. мультифрактал можно понимать как объединение различных однородных фрактальных подмножеств L_α исходного множества L, каждое из которых имеет собственное значение фрактальной размерности.

Для характеристики мультифрактального множества используют так называемую функцию мультифрактального спектра $f(\alpha)$ (спектр сингулярностей мультифрактала), к которой вполне подходил бы термин «фрактальная размерность». Величина $f(\alpha)$ фактически равна хаусдорфовой размерности однородного фрактального подмножества L_α из исходного множества L, которое дает доминирующий вклад в некоторую статистическую сумму, называемую моментом порядка q для рядов мер релевантностей.

Кроме того, для описания мультифрактала используют обобщенные фрактальные размерности D_q, которые определяются соотношением:

$$D_q = \lim_{r\to 1} \frac{1}{q-1} \frac{\ln \sum_{i=1}^{N} p_i^q}{\ln r},$$

где p_i – вероятность того, что случайная величина (или элемент нормированного числового ряда по общей сумме) попадет в некоторый диапазон r.

Соотношение между D_q и показателями q и τ выглядит следующим образом:

$\tau(q) = (1 - q) D_q$.

Функции $f(\alpha)$ и $\tau(q)$ вязаны друг с другом соотношением:

$\tau(q) = f(\alpha) - q\alpha$,

где α как функция от q определяется из решения уравнения:

$d/d\alpha\ (q\alpha - f(\alpha)) = 0$.

И наоборот, если известен показатель мультифрактального скейлинга $\tau(q)$, то мультифрактальный спектр может быть найден по формуле:

$f(\alpha(q)) = \tau(q) + q\alpha(q)$,

где $\alpha(q) = - d/dq\ \tau(q)$.

Эти соотношения задают кривую $f(\alpha)$ параметрически (как функцию от параметра q) и представляют собой преобразование Лежандра от переменных q и τ к переменным α и f.

4.2 Метод вычислений

При анализе рассматриваемого нами ряда релевантностей использовался следующий метод расчета мультифрактальных характеристик. Значения исследуемого ряда нормировались $Z_i = \xi_i / \sum_i \xi_i$ и ассоциировались с вероятностями p_i появления случайного события в момент t_i в рамках приведенной в [30] модели.

После нормирования весь диапазон $[0; N]$ разбивался на $n = N/m$ ячеек (участков) длиной m. Затем определялась следующая сумма:

$$S_m^Z(q) = \sum_{k=1}^{n} (\overline{Z_k^{(m)}})^q,$$

где

$$\overline{Z}_k^{(m)} = \sum_{l=1}^{m} Z_{(k-1)m+l}.$$

Как оказалось для рассматриваемых числовых рядов, значение $\log S_m^Z(q)$ хорошо аппроксимируется линейной зависимостью от $\log m$, в результате чего появилась возможность говорить [30], что числовой ряд демонстрирует мультифрактальный скейлинг, или короче: Z_i - мультифрактал. Наклон аппроксимирующей линии, полученный методом наименьших квадратов - $\tau^Z(q)$, или короче $\tau(q)$ определялся по формуле:

$$\log S_m^Z(q) \cong \tau^Z(q) \log m + const.$$

4.3 Расчет мультифрактальных характеристик

На рис. 9. показана трехмерная поверхность – зависимость $\tau(q,m)$ от q и m для распределений документов 1-го корпуса по степени релевантности. В соответствии с формулой:

$f(\alpha(q)) = \tau(q) - q\ \tau'(q)$,

был определен мультифрактальный спектр исследуемого ряда. На рис. 10 показана трехмерная поверхность – зависимость $f(\alpha(q))$ от q и m.

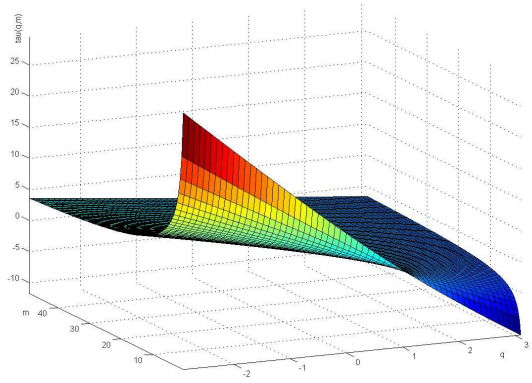


Рис. 9. Значения τ(q,m) для исследуемого ряда (запрос «банк»)

На рис. 11 показана зависимость $f(\alpha(q))$, которая при больших m уже не зависит от этой переменной.

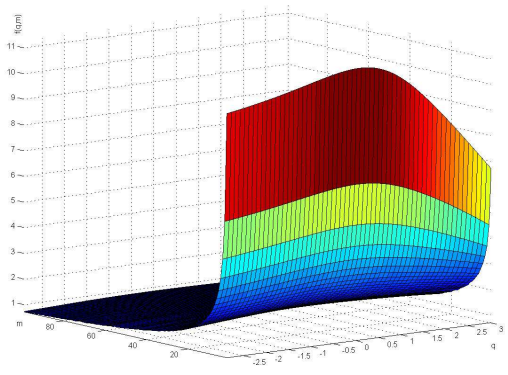


Рис. 10. Значения мультифрактального спектра для исследуемого ряда (запрос «банк»)

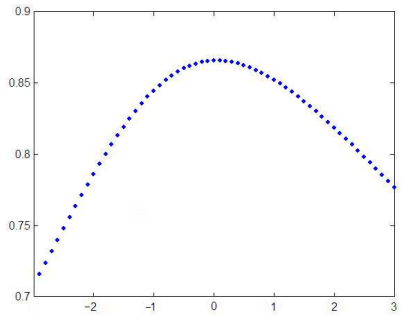


Рис. 11. «Стабилизированные» по m значения мультифрактального спектра f (ось Y) от q (ось X)

Во многих мультифрактальных исследованиях основным объектом анализа является зависимость мультифрактального спектра f от индекса сингулярности (показателя Липшица-Гельдера) a, а не от q. В рамках данной работы также была оценена эта зависимость (рис. 12).

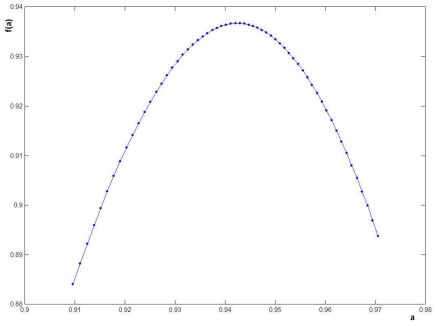


Рис. 12. Зависимость мультфрактального спектра (ось Y) от индекса сингулярности (ось X)

Как видно из рис. 12, функция $f(a)$ практически симметрична, что свидетельствует о приблизительно одинаковом вкладе документов с малой и большой релевантностью в сумму $S_m^z(q)$, а соответственно во всю мультифрактальную структуру исследуемого ряда,. Вместе с тем, зависимость мультифрактального спектра f от q позволяет, как будет показано ниже, получать более или менее реперезентативные выборки именно из документов, обладающих достаточно однородным (высоким) уровнем релевантности.

5. Верификация мультифрактальной репрезентативности выборок

В качестве простейшего примера приведем мультифрактальный спектр следующих рядов: исходного исследуемого ряда; ряда полученного путем отбора первой половины значений исходного ряда; ряда, полученного отбором каждого второго элемента из исходного ряда. Как видно на рис. 13, ряд полученный отбором каждого второго элемента обладает сходной по изгибу кривой мультифрактального спектра с исходным рядом, что, по мнению авторов, подтверждает его репрезентативность. С другой стороны, кривая, соответствующая половине исходного ряда имеет существенно отличные параметры кривизны, что может свидетельствовать о нерепрезентативности такой выборки.

В качестве экспериментов по выявлению репрезентативных подборок рассматривалось два подхода:

1) выбирались подборки, куда входил каждый 2-й документ, каждый 4-й, 8-й и 16-й. Как видно порис. 14, изгиб кривой мультифрактального спектра при этом уменьшался, выборки постепенно теряли свою репрезентативность.

2) значения релевантностей документов делилось на M равных частей. В соответствии со значениями релевантностей формируются M выборок документов. Как видно по рис. 15 при $M = 3$, наиболее репрезентативной с точки зрения изгиба кривой мультифрактального спектра является выборка с наибольшими релевантностями. Мультифрактальный спектр, соответствующий средним уровням релевантностей хорошо аппросксимируется горизонтальной прямой, что свидетельствует о том, что данная выборка – монофрактал.

Конечно, возможны и другие алгоритмы отбора документов для формирования выборок. При этом к ним также может быть применен данный подход.

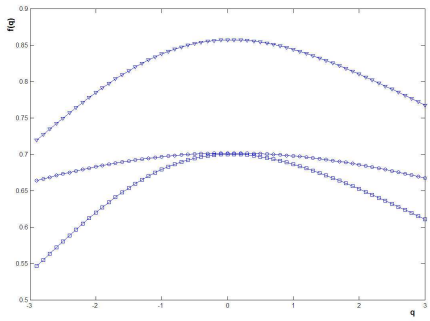


Рис. 13. Мультифрактальные спектры: исходного ряда (∇), половинного ряда (o) и ряда, полученного отбором каждого второго элемента (□)

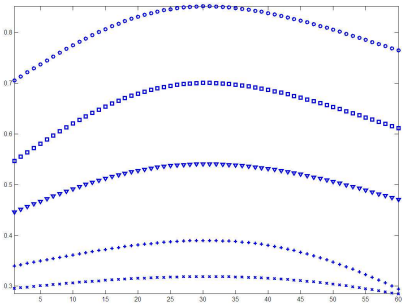


Рис. 14. Мультифрактальные спектры: исходного ряда (o), рядов, полученных отбором каждого 2-го (□), 4-го (∇), 8-го (+), 16-го (x) элементов

6. Заключение

Итак, мы видим, что распределения документов по степени релевантности в рассмотренных случаях обладают мультифрактальной природой. Однако их общие свойства достаточно близки к фрактальным, и это обстоятельство позволяет нам воспользоваться рядом важных и полезных вытекающих из него следствий.

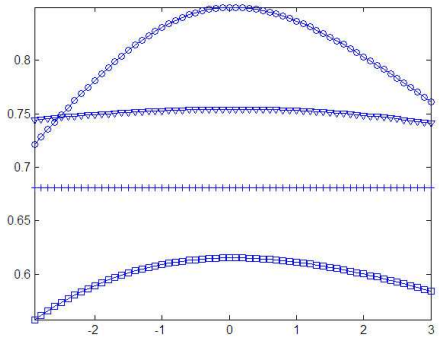


Рис. 15. Мультифрактальные спектры: исходного ряда (o), рядов, полученных отбором документов с максимальной (□), средней (+) и минимальной (∇) релевантностями

Степени релевантности документов, зависящие от количества входжений в них слов из запроса, при том, что документы ранжированы по другой степени релевантности, образуют множество, обладающее фрактальными и даже мультифрактальными свойствами. Эти свойства предоставляют механизмы верификации подобных по мультифрактальной природе подмножеств. Варьируя рассматриваемые выборки, мы можем строить различные наборы документов, репрезентативных исходному корпусу.

Похоже, что сам факт наличия у большого набора документов мультифрактальных свойств, является достаточным обоснованием предполагаемой репрезентативности отдельных выборок. При этом окончательную репрезентативность выборки следует понимать в том смысле, что эксперт, изучивший входящие в нее документы, даст им такую же оценку, как и всему исходному набору. Вместе с тем, очевидные случаи репрезентативных выборок (типа выбора каждого n-го документа при n = 2, 4, 8 и 16 из исходного числового ряда значений мер релевантностей) исследованы в ходе экспериментов и убедительно свидетельствуют в пользу данной методики.

Обоснование мультифрактальной природы распределения релевантностей дает возможность решить основную практическую часть исследования, а именно предоставить критерий того, что выборка документов из исследуемого ряда является репрезентативной по некоторому латентному свойству - мультифрактальному спектру. Такой критерий репрезентативности может использоваться наряду с другими, более очевидными статистическими критериями, например, близости математических ожиданий, дисперсий или корреляционных функций.

Практическая ценность задачи выявления репрезентативных выборок может быть выражена в таких приложениях, как предъявление пользователю обозримых результатов поиска, отражающих весь спектр документального корпуса или выделение локальных массивов документов (корпусов) для дальнейших детальных исследований.

В рамках настоящей работы авторы вплотную подошли к альтернативному способу объединения документов в кластеры по близости мультифрактальных спектров.

Однако, если задачу выявления репрезентативных массивов документов с помощью такого подхода можно считать теоретически решенной, то задача кластеризации исходных документов по мультифрактальному спектру лишь обозначена и ждет своей детальной проработки.

7. 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] Манифест РОМИП (http://romip.narod.ru/ru/manifest.html)

[4] G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval. // In Information Processing and Management: an International Journal, Volume 24, Issue 5, pages: 513 – 523, 1988.

[5] Haveliwala T. Efficient computation of PageRank. Technical report, Stanford Database Group, Oct. 1999.

[6] Lifantsev M. Voting Model for Ranking Web Pages. In Proc. of the IC'00, pp. 143-148, 2000.

[7] Zhang D., Dong Y. An efficient algorithm to rank web resources. In Proc. of the WWW9, pp.

449-455, 2000.
[8] Douglas L. Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In Proceedings of the SIGIR'98, pages 96-103, 1998.
[9] Ron Papka and James Allan. Document classification using multiword features. In Proc. of the CIKM'98, pages 124-131, New-York, November 1998.
[10] Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In Proc. of the SIGIR'2000, 2000.
[11] P.W. Foltz. Using latent semantic indexing for information filtering. In ACM Conference on Office Information Systems (COIS), pages 40-47, 1990.
[12] J. Callan. Learning while filtering documents. In Proc. of SIGIR'98, pages 224-231, Melbourne, Australia, 1998.
[13] Fractal geometry of information space as represented by cocitation clustering / Van Raan A. F. J. // Scientometrics. -1991. – Vol. 20, № 3. - P. 439-449.
[14] Федер Е. Фракталы -М.: Мир, 1991. - 254 с.
[15] Гринченко В.Т., Мацыпура В.Т., Снарский А.А. Введение в нелинейную динамику. Хаос и фракталы. Изд. 2.–М: УРСС, 2007. - 263 с.
[16] Олемской А.И., Флат А.Я. Использование концепции фрактала в физике конденсированной среды // Успехи физ. наук. -1993. - Т. 163. - №12. С. 1-50.
[17] Ландэ Д.В. Фрактальные свойства тематических информацийнных потоков из Интернет // Регистрация, хранение и обработка данных. -К., 2006, Т. 8, № 2. - С. 93 - 99.
[18] Иванов С.А. Стохастические фракталы в Информатике // Научно-техническая информация. Сер. 2, 2002. - № 8. - С. 7-18.
[19] Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: «Старт-98», 2007. – 40 с.
[20] Божокин С.В., Паршин Д.А. Фракталы и мультифракталы. Ижевск: НИЦ «Регулярная и хаотическая динамика». 2001. 128 с.
[21] Павлов А.Н., Сосновцева О.В., Зиганшин А.Р. Мультифрактальный анализ хаотической динамики взаимодействующих систем // Изв. вузов, Прикладная нелинейная динамика, т. 11, No. 2, стр. 39-54 (2003).
[22] B.B. Mandelbrot, Fractals and Multifractals: Noise, Turbulence and Galaxies, Selecta Vol. 1 (Springer-Verlag, New York, 1989.
[23] ГОСТ 7.73-96 SU. Поиск и распространение информации.
[24] Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing / - Cambridge, Massachusetts: The MIT Press, 1999.
[25] Singhal, A., C. Buckley, M. Mitra, “Pivoted Document Length Normalization”. ACM SIGIR 96.
[26] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска / Программирование. - 28(4), 2002. - С. 226-242.
[27] C.-K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. // CHAOS. 1995. Vol. 5, P. 82.
[28] H.E. Stanley, L.A.N. Amaral, A.L. Goldberger, S. Havlin, P.Ch. Ivanov, C.-K. Peng, Statistical physics and physiology: monofractal and multifractal approaches. // Physica A. 1999. Vol. 270, P. 309.
[29] Шелухин О.И., Тенякшев А.М., Осин А.В. Фрактальные процессы в телекоммуникациях. - М.: Радиотехника, 2003.- 480 с.
[30] Rudolf H. Riedi, Jacques Levy Vehel. Multifractal Properties of TCP traffic: a numerical study. Technical Report № 3128 INRIA Rocquencourt. - March 1997.
[31] Светова Н.Ю. Условные и взаимные мультифрактальные спектры. Определение и основные свойства // Труды Петрозаводского государственного университета. Сер. «Математика». Петрозаводск: Изд-во ПетрГУ. Вып. 10. 2003. С. 41-58.
[32] Иудин Д.И., Гелашвили Д.Б., Розенберг Г.С. Мультифрактальный анализ видовой структуры биотических сообществ // Докл. Акад. Наук. – 2003. – Т. 389. - №2, - С.279-282.

Documents distribution on a degree of relevance on a basis multifractal properties

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

Documents distributions in network information streams on two degrees of relevance concerning search inquiries are investigated. The multifractal nature of documents distributionson degrees of relevance is proved. The basic algorithms used during researches are described.

The opportunity of use multifractal properties for verification relevant documents representativeness samples which can be considered as one of variants of representation of end results of search or documentary corpses for the further researches is proved.