Свойства распределения релевантности в документальных массивах♣

© Снарский А.А. Ландэ Д.В. Брайчевский С.М. Дармохвал А.Т.

НТТУ «КПИ» Информационный центр «ЭЛВИСТИ»
asnarskii@gmail.com dwl@visti.net smb@visti.net hval@visti.net

Аннотация

Исследуются распределения двух видов меры релевантности документов в документальных потоках. Выявлены устойчивые корреляции в их взаимных зависимостях. Определен показатель Херста соответствующих рядов и показано, что они обладают фрактальной природой.

1. В связи с наблюдающимся в последние годы быстрым ростом объемов сетевой информации и темпов ее обновления особую актуальность приобретает задача изучения статистических свойств сетевых документальных массивов [1-3]. Сложность и многоплановость этой задачи, в свою очередь, предполагает активное использование современных теоретических методов, позволяющих более глубоко понять специфику данной предметной области. В том числе, перспективными представляются различные методы теории детерминированного хаоса [4-7], получившие в настоящее время широкое распространение во многих областях науки. Применение таких методов в нашем случае представляется тем более интересным, что сетевые документальные массивы в этом плане остаются малоизученными.

В предлагаемой работе исследуются распределения мер релевантности документов, определяемые двумя различными способами (нормированной и ненормированной на длину документа). На основании обработки опытных данных для нормированной меры релевантности получено значение показателя Херста соответствующих рядов, а также показано, что они обладают фрактальной природой.

Исследования проводились на наборе документальных корпусов, содержащих сообщения онлайновых СМИ различных объемов, сформированные системой InfoStream [8].

2. В классической задаче информационного поиска под релевантностью, как известно, понимается формальное соответствие документа из набора данных, в котором осуществляется поиск, информационному запросу пользователя [9]. 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 [12, 13]. Существенным является то, что мера (1) является однопараметрической (определяется частотами поисковых терминов), а мера (2) – двухпараметрической (определяется частотами поисковых терминов и длиной документа).

В связи с введением такой величины как мера релевантности, возникает естественный вопрос: какими свойствами обладают ее распределения по документам, отобранным поисковой системой?

3. В ходе исследований обрабатывались информационные корпуса, содержащие сообщения онлайновых СМИ. В дальнейшем изложении будем рассматривать как пример массив из 5000 документов, опубликованных за 3 суток с 1 по 3 декабря 2006 г., предположительно по банковской тематике, удовлетворяющих запросу “банк”.

На рис. 1. приведены зависимости меры

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2007, Переславль-Залесский, Россия, 2007.

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 и следовало ожидать, эти зависимости обладают существенно различным поведением.

Однако, из приведенного графика видно, что между данными зависимостями существуют вполне отчетливые корреляции. 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.

4. На рис. 2 представлены те же самые данные, но в полулогарифмическом масштабе. Прежде всего, мы видим, что отсортированная 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 [1], но и на начальном участке.

Очевидно, что частоты поисковых терминов в документе образуют дискретный набор (термин может встречаться один раз, два раза и т. д.). Поэтому отсортированная зависимость однопараметрической меры R_F на самом деле представляет собой набор линейных участков, отвечающих каждому значению частоты появления термина-запроса в документах.

Нетрудно заметить, что зависимость R_F, изображенная на рис. 2, обладает специфической структурой. В нижней части она имеет явную регулярность, обусловленную дискретностью частот поискового термина. 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 характерные особенности, типичные для детерминированного хаоса.

[1] Речь идет о законе Ципфа, применяемого к распределению произвольных объектов.

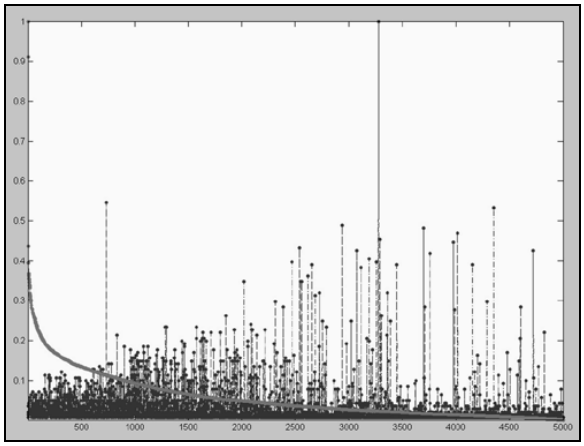

Рис. 1. Значения мер релевантности (ось Y) документов по двум критериям. Документы (ось X) ранжированы по значениям R_N

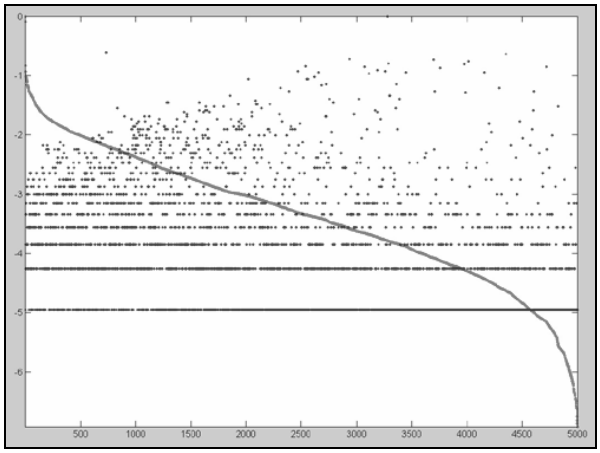

Рис. 2. Значения мер релевантности документов по двум критериям в полулогарифмической шкале

Как известно, возникновение детерминированного хаоса в динамике объектов тесно связано с наличием у него фрактальных свойств, важность которых в последние годы широко обсуждается в самых различных областях науки [4, 7].

Наиболее интересным объектом для изучения фрактальных 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.

Одним из универсальных подходов к выявлению самоподобия основывается на методе обработки временных рядов DFA (Detrended Fluctuation Analysis) [14-16]. Этот подход представляет собой вариант дисперсионного анализа, позволяющий исследовать эффекты длительных корреляций в нестационарных рядах.

В соответствии с алгоритмом 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 числового ряда ξ_i и строится обобщенное 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. 3. представлена зависимость $F(n)$ от длины участков аппроксимации в двойном логарифмическом масштабе. Наличие линейного участка на этом графике позволяет говорить о наличие локального скейлинга. Этот факт, дополнительно к работам [5, 6, 17], еще один пример, подтверждающий фрактальную природу структур, порождаемых информационными процессами.

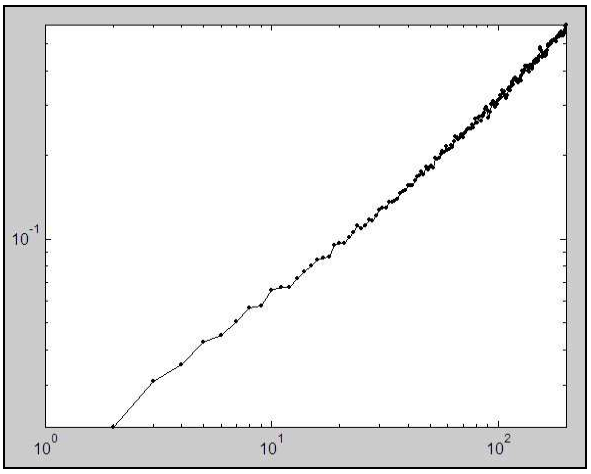

Рис. 3. Зависимость F(n) ряда наблюдений (ось Y) от длины участка аппроксимации n (ось X) в логарифмической шкале

Основной характеристикой рядов, обладающих хаотическим поведением, является, как известно, показатель Херста [4]. Для его определения был применен, так называемый, *R/S*-анализ, который хорошо зарекомендовал себя, например, в исследованиях фрактальной природы научных коммуникаций и информационных потоков [17]. Он позволяет достаточно эффективно исследовать свойства числовых рядов на основе отношения разброса значений к среднеквадратичному отклонению. Показатель Херста вычисляется по следующему алгоритму. Сначала вычисляется среднее значение измеряемой переменной:

Показатель Херста вычисляется по следующему алгоритму. Сначала вычисляется среднее значение измеряемой переменной по всему числовому ряду $<\xi>_N$ и стандартное отклонение S. Затем рассчитывается накопившееся отклонение ряда измерений $\xi(t)$ от среднего $<\xi>_N$:

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

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

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

Пусть R – размах значений числового ряда, образуемого в нашем случае набором N значений меры релевантности R_N, а S – его среднеквадратичное отклонение. Тогда для объектов фрактальной природы имеет место соотношение:

$$\frac{R}{S} = \left(\frac{N}{2}\right)^{H}$$

где H – показатель Херста, который для достаточно широкого класса рядов измерений связан с хаусдорфовой размерностью $D = 2 - H$. Численные значения H характеризуют различные типы коррелированной динамики. При $H = 0.5$ наблюдается некоррелированное поведение ряда, а значения $0.5 < H < 1$ соответствуют анти-корреляциям (чередование больших и малых величин в анализируемых данных).

На рис. 4. изображена зависимость значения показателя Херста от размерности подмножества документов. Мы видим, что при небольших объемах наборов документов показатель Херста обладает сложным апериодическим поведением, но при увеличении их его значение стремится к величине ~ 0.77, что соответствует хаусдорфовой размерности ~ 1.23.

Таким образом, проведенные исследования числового ряда распределений релевантностей подтвердили предположение о самоподобии и

фрактальной природе.

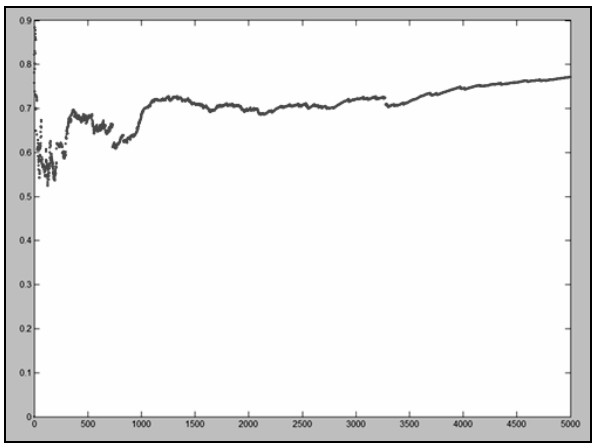

Рис. 4. Значения показателя Херста (ось Y) в зависимости от объема исследуемого массива (ось X)

5. 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 отсортируем по значениям меры релевантности R_N, поскольку такое представление данных обладает большей наглядностью.

Полученные таким образом результаты приведены на рис. 5. В известном смысле можем сказать, что картины, приведенные на нем и на Рис. 1, взаимно обратимы.

Учитывая то, что частоты слов в документах и длины самих документов в документальных потоках распределены достаточно случайно, мы можем ожидать, что распределение даже двухпараметрической характеристики по однопараметрической будет, в некотором приближении, описываться теоремой Пуассона. Именно, максимальные значения R_N для локальных отрезков R_F = const будут распределены по Пуассону. И, действительно, мы видим нечто подобное на рис. 5. Причем максимум распределения Пуассона приходится именно на центральный участок рис. 2, на котором наблюдается выполнение обобщенного закона Ципфа.

Отметим, что данное обстоятельство открывает интересный аспект, непосредственно связанный с проблемой оптимизации информационного поиска.

6. В заключение отметим, что данное обстоятельство открывает нам интересный аспект, непосредственно связанный с проблемой оптимизации результатов информационного поиска по набору критериев (например, по двум мерам релевантности, использовавшимся в данной работе). Из характера зависимости R_N от R_F, следует, что решение этой задачи может быть получено путем отбора подмножеств документов из локальных интервалов, лежащих справа от точек $\max(\{R_N\}_k)$, принадлежащих центральному участку рис. 2. 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.

Поэтому процесс построения наборов данных, приведенных на рис. 5, 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.

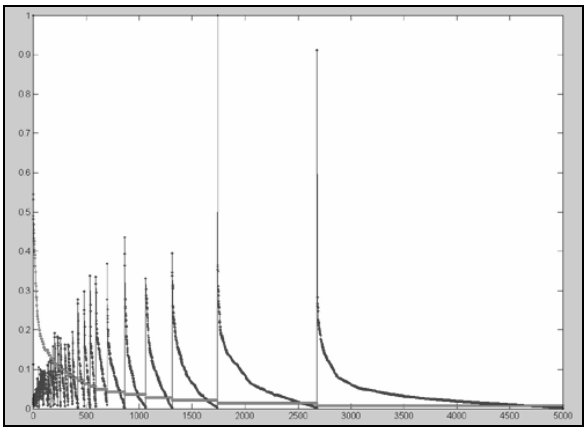

Рис. 5. Значения мер релевантности документов по двум критериям. Документы ранжированы по значениям R_F

Результаты, полученные в данной работе, позволили вплотную подойти к исследованию не только фрактальных, но и мультифрактальных свойств распределения документов по степени релевантности, что, в свою очередь, как представляется авторам, позволит решить важную практическую задачу выявления репрезентативных массивов документов, а в дальнейшем и задачу кластеризации массивов документов по мультифрактальному спектру.

Литература

[1] Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing - Cambridge, Massachusetts: The MIT Press, 1999.

[2] Gianna M. Del Corso, Antonio Gullí Univerisity, Francesco Romani. Ranking a stream of news. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. - P. 97 - 106.

[3] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. - Сер. 1. - Вып. 11. – 2005. - С. 21-33.
[4] Федер Е. Фракталы -М.: Мир, 1991. - 254 с.
[5] Van Raan A. F. J. Fractal geometry of information space as represented by cocitation clustering // Scientometrics. -1991. – Vol. 20, № 3. - P. 439-449.
[6] Стохастические фракталы в Информатике / Иванов С.А. // Научно-техническая информация. Сер. 2, 2002. - № 8. - С. 7-18.
[7] Гринченко В.Т., Мацыпура В.Т., Снарский А.А. Введение в нелинейную динамику. Хаос и фракталы. Изд. 2.–М: УРСС, 2007. - 263 с.
[8] Григорьев А.Н., Ландэ Д.В., и др. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: Старт-98, 2007. – 40 с.
[9] ГОСТ 7.73-96 SU. Поиск и распространение информации.
[10] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска // Программирование. - 28(4), 2002 - С. 226-242 .
[11] Российский семинар по Оценке Методов Информационного Поиска. Труды четвертого российского семинара РОМИП'2006. - Санкт-Петербург: НУ ЦСИ, 2006, 274 с.
[12] Baeza-Yates, R., Ribeiro-Neto, B. 1999. Modern Information Retrieval. Addison Wesley, New York, NY.
[13] Singhal, A., C. Buckley, and M. Mitra, “Pivoted Document Length Normalization,” ACM SIGIR 96.
[14] 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;
[15] 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.
[16] Pavlov A.N., Ebeling W., Molgedey L., Ziganshin A.R., Anishchenko V.S., Scaling features of texts, images and time series // Physica A, vol. 300, pp. 310-324 (2001).
[17] Павлов А.Н., Сосновцева О.В., Зиганшин А.Р., Мультифрактальный анализ хаотической динамики взаимодействующих систем // Изв. вузов, Прикладная нелинейная динамика. - Т. 11, - № 2. С. 39-54. -2003.
[18] Ландэ Д.В., Григорьев А.Н., Брайчевский С.М. Стабильность источников информации как один из параметров информационных потоков // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2006 – Москва: Наука, 2006. - С. 332-334.

The properties of relevance distribution in documentary arrays

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

Distributions of two kinds of the measure of relevance of the documents in the documentary streams are investigated.

Stable correlations in their reciprocal dependence are revealed. Hurst index of the corresponding rows is defined. It is shown that they possess fractal nature.

♣ Исследование является частью НИР, поддержанной компанией «Яндекс» в рамках конкурса «Интернет-математика»