УДК 510.52:004.7

ПРОГРАММЫ-АРХИВАТОРЫ ДЛЯ ВЫЧИСЛЕНИЯ КОЛМОГОРОВСКОЙ СЛОЖНОСТИ

[1]Печников А.А., [2]Прусский Д.А.

[1]Институт прикладных математических исследований – обособленное подразделение Федерального исследовательского центра «Карельский научный центр Российской академии наук», Петрозаводск, e-mail: pechnikov@krc.karelia.ru;

[2]Санкт-Петербургский государственный университет, факультет прикладной математики – процессов управления, Санкт-Петербург, e-mail: dimaprusskii@mail.ru

В предыдущих работах одного из авторов был предложен подход к исследованию схожести веб-сайтов с использованием Колмогоровской сложности и нормализованного расстояния сжатия, показывающий определенный потенциал такого подхода. При этом сформулирован ряд вопросов, на которые требуется ответить, прежде чем автоматизировать проведение больших исследований. Один из этих вопросов относится к программам-архиваторам, используемым на практике в качестве так называемого способа описания, и заключается он в том, можно ли выбрать наилучший архиватор. Этот вопрос подробно исследуется в данной статье. Авторами была проведена серия экспериментов с девятью наиболее популярными и доступными архиваторами, позволившая на практике проверить выполнение свойств идемпотентности, монотонности, симметричности и дистрибутивности для этих программ. Четыре наилучших архиватора были экспериментально проверены на выполнение аксиом расстояния для нормализованного расстояния сжатия, вычисляемого с их использованием, и показано их выполнение. Все этапы экспериментов были автоматизированы с помощью разработанных вспомогательных программ для скачивания первых страниц сайтов в формате html, создания конкатенаций файлов, сжатия файлов различными архиваторами и вычисления нормализованного расстояния сжатия. Авторы остановились на программе RAR, как наиболее удобной в использовании, и с её помощью на конкретном примере провели иерархическую кластеризацию заданного множества сайтов, практически полностью соответствующую компаниям – разработчикам сайтов, что свидетельствует о выявлении стиля разработки сайтов, присущего каждой компании, с помощью предложенного подхода и «наилучшего» архиватора.

Ключевые слова: Колмогоровская сложность, нормализованное расстояние сжатия, программа-архиватор, веб-сайт, кластерный анализ

FILE ARCHIVERS TO EVALUATE THE KOLMOGOROV COMPLEXITY

[1]Pechnikov A.A., [2]Prusskiy D.A.

[1]Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Petrozavodsk, e-mail: pechnikov@krc.karelia.ru;

[2]Saint-Petersburg State University, Faculty of Applied Mathematics and Control Processes, Saint-Petersburg, e-mail: dimaprusskii@mail.ru

An approach to the research of the similarity of websites using Kolmogorov complexity and normalized compression distance was proposed in previous works by one of the authors. A certain potential of such an approach was revealed. At the same time, a number of questions that need to be answered before automating the conduct of large studies was formulated. One of these questions relates to the archiver programs used in practice as a so-called method of description, and it is whether you can choose the best archiver. In this article, the question is investigated in detail. The authors conducted a series of experiments with nine of the most popular and accessible archivers, which allowed to verify the implementation of the properties of idempotency, monotony, symmetry and distributivity for these programs in practice. The best four archivers were experimentally tested for the implementation of distance axioms for the normalized compression distance, calculated using them. Their implementation is shown. All stages of the experiments were automated using developed auxiliary programs for downloading the first pages of sites in html format, creating file concatenations, compressing files with various archivers, and calculating the normalized compression distance. The authors settled on the RAR program, as the most convenient to use, and conducted a hierarchical clustering of a given set of sites using it. This clustering almost completely corresponds to the site development companies, which indicates the identification of the site development style inherent in each company using the proposed approach and the «best» archiver.

Keywords: Kolmogorov complexity, normalized compression distance, archiver program, website, cluster analysis

В работе [1] был предложен подход к исследованию схожести веб-сайтов с использованием Колмогоровской сложности и нормализованного расстояния сжатия. Там же был сформулирован ряд вопросов, которые требуется рассмотреть, прежде чем автоматизировать проведение больших исследований, один из которых, – можно ли выбрать наилучший архиватор и какими свойствами он должен обладать.

Основная цель данной статьи заключается в разработке методов, подходов и критериев определения наилучшего архиватора из заданного набора наиболее популярных и доступных архиваторов.

Для девяти архиваторов проведена серия экспериментов по проверке свойств так называемой «нормальности» и были оставлены четыре с наилучшими показателями выполнения этих свойств. Для этих четырех

архиваторов было показано экспериментально выполнение аксиом расстояния для нормализованного расстояния сжатия.

Все этапы экспериментов были автоматизированы с помощью разработанных вспомогательных программ. С использованием программы RAR, вошедшей в четверку «нормальных», была проведена иерархическая кластеризация тестового множества сайтов компаний – разработчиков вебсайтов, имеющая хорошую содержательную интерпретацию.

Основные понятия и инструменты

Способом описания называется произвольное вычислимое частичное отображение D из множества двоичных слов Ξ в себя [2]. Если $D(y) = x$, говорят, что y является описанием x при способе описания D. Для каждого способа описания D сложность относительно этого способа описания равна длине кратчайшего описания $l(y)$: $KS_D(x) = \min\{l(y)|D(y) = x\}$.

Чтобы определить Колмогоровскую сложность, необходимо ввести понятие оптимального способа описания. Закономерно определить, что способ описания D_1 не хуже способа описания D_2, если $KS_{D1}(x) \le KS_{D2}(x) + c$ при некотором c и для всех x.

Теорема Соломонова – Колмогорова говорит, что существует такой способ описания D, что для любого другого способа описания D' найдется такая константа c, что $KS_D(x) \le KS_{D'}(x) + c$ для любого слова x [2]. Будем называть оптимальным такой способ описания, который обладает приведенным в теореме свойством. Колмогоровской сложностью слова x будем называть $KS_D(x)$, где D – оптимальный способ описания.

Теперь зафиксируем некоторый (не обязательно оптимальный) способ описания, и сложность слова x относительно этого способа описания обозначим $K(x)$. Пусть y – еще одно двоичное слово. Обозначим $K(x|y)$ минимальное количество битов, необходимых для восстановления x из y. Для любой пары строк x и y можно определить нормализованное расстояние сжатия (normalized compression distance, *NCD*) как в [3]:

$$NCD(x,y) = \frac{\max\{K(x\,|\,y), K(y\,|\,x)\}}{\max\{K(x), K(y)\}}. \quad (1)$$

Простыми словами, если два объекта достаточно похожи, то мы можем более кратко описать один из них, учитывая информацию о другом.

Вводя расстояние *NCD* на множестве двоичных слов Ξ, мы тем самым определяем метрическое пространство (Ξ, *NCD*), если расстояние удовлетворяет трем аксиомам:

– $NCD(x, y) = 0$ тогда и только тогда, когда $x = y$ (аксиома тождества),

– $NCD(x, y) = NCD(y, x)$ (аксиома симметрии),

– $NCD(x, y) \le NCD(x, z) + NCD(z, y)$ (аксиома или неравенство треугольника).

В [3] показано, что *NCD*, заданная формулой (1), удовлетворяет указанным аксиомам.

Колмогоровская сложность является невычислимой по Тьюрингу [2]. Поэтому на практике в качестве отображения D используются программы-архиваторы [4]. Свойства *NCD*, очевидно, зависят от особенностей используемого архиватора.

В соответствии с [5] определим понятие «нормального архиватора». Авторы в [5] используют термин «compressor», который мы переводим как «архиватор» (archiver), тем более что в [5] компрессор определяется как кодировщик без потерь, что более свойственно архиваторам, а не компрессорам [6]. Итак, архиватор C является нормальным, если он удовлетворяет следующим свойствам:

1) идемпотентность: $C(xx) = C(x)$,

2) монотонность: $C(xy) \ge C(x)$,

3) симметричность: $C(xy) = C(yx)$,

4) дистрибутивность: $C(xy) + C(z) \le C(xz) + C(yz)$,

с точностью до аддитивного члена $O(log(n))$, где n – максимальная двоичная длина элемента из Ξ, включенного в рассматриваемые (не)равенства.

Теперь в формуле (1) заменим $K(\cdot)$ на $C(\cdot)$, получая формулу

$$NCD'(x,y) = \frac{\max\{C(x\,|\,y), C(y\,|\,x)\}}{\max\{C(x), C(y)\}}. \quad (2)$$

Это следует понимать в том смысле, что если два файла близки в соответствии с «теоретическим» расстоянием *NCD*, определяемым через Колмогоровскую сложность, то они также близки в смысле нормального архиватора. Очевидно, что для любых x и y имеем $0 \le NCD' \le 1 + \varepsilon$, представляющее, насколько различны эти два файла. Меньшие числа представляют более похожие файлы. Значение ε в верхней границе связано с деталями методов сжатия, что мы и увидим далее. В [5] доказана теорема, утверждающая, что если при вычислении *NCD* был использован нормальный архиватор, то такое *NCD* удовлетворяет аксиомам расстояния. Значит, чтобы максимально точно аппроксимировать *NCD*, вначале необходимо проверить существующие реальные архиваторы на «нормальность» и выбрать те из них, которые удовлетворяют его свойствам.

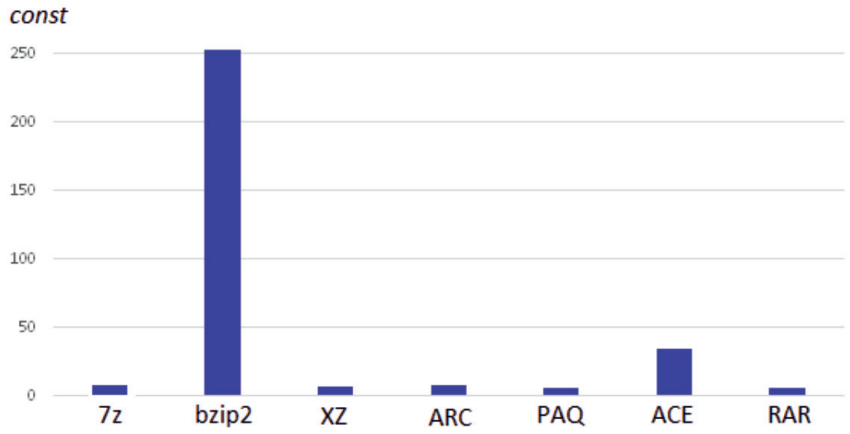

Рис. 1. Результаты проверки идемпотентности для семи архиваторов

Проверка свойств нормального архиватора

Были рассмотрены программы-архиваторы bzip2, gzip, XZ, 7z, ACE, ARC, PAQ, RAR и ZIP, выбор которых объясняется их популярностью и доступностью. Не углубляясь в детали работы архиваторов из-за недостатка места, скажем, что они могут быть разбиты на два семейства, – потоковые и блочные, – алгоритмы работы которых существенно различны [6].

В качестве объектов для экспериментов были выбраны первые страницы 30 различных сайтов в формате html. Среди них есть сайты вузов, факультетов СПбГУ, научных, коммуникационных и торговых площадок, а также сайты коммерческих компаний: www.amursu.ru, jf.spbu.ru, arxiv.org, www.amazon.com, amdm.ru, www.instagram.com, ok.ru и др.

Все этапы экспериментов были автоматизированы с помощью разработанных вспомогательных программ для скачивания первых страниц сайтов в формате html, создания конкатенаций файлов, сжатия файлов различными архиваторами и вычисления нормализованного расстояния сжатия.

В разделе 1 сказано: свойство идемпотентности для нормального архиватора должно выполняться с требуемой точностью, а именно $C(xx)$-$C(x) = O(log)n))$. Если это свойство выполняется, то разность между длиной сжатой версии файла и длиной сжатой версии конкатенации этого файла с самим собой должна быть ограничена сверху функцией $const*log(n)$, где $const$ – константа, а n – размер файла: $C(xx)$-$C(x) \leq const*O(log)n))$. Поэтому для проверки идемпотентности и сравнения результатов будем рассматривать эту разность для каждого объекта и будем делить её на $log(n)$:

$$const = \frac{C(xx) - C(x)}{\log(n)}.$$

Тогда чем больше эта константа, тем хуже выполняется проверяемое свойство. На рис. 1 представлена диаграмма средних значений *const*, вычисленных по описанному выше способу, по которой можно судить о выполняемости идемпотентности различными архиваторами; zip и gzip отсутствуют на диаграмме (значения $const > 1500$).

Любой архиватор должен обладать свойством монотонности, по крайней мере с требуемой точностью. Это свойство очевидно для потоковых архиваторов и лишь немного менее очевидно для блочных архиваторов, поэтому отдельную проверку проводить нецелесообразно.

Для проверки симметричности будем по аналогии с предыдущей проверкой вычислять константу, определяющую точность выполнения свойства как отношение разности между длинами симметричных конкатенаций файлов к логарифму от длины одного из этих файлов.

Известен теоретический факт, что потоковые компрессоры не являются абсолютно симметричными, что с особенностями сжатия: исходный файл может иметь закономерности, к которым адаптируется архиватор; однако после пересечения границы между файлами внутри конкатенации он должен «отучиться» от этих закономерностей и приспособиться к закономерностям второго файла [5–7]. Это определяет в свойстве симметрии неточность, которая асимптотически уменьшается с длиной

файла. Для архиваторов, основанных на блочном кодировании, симметричность выполняется более строго, поскольку они анализируют каждый входной блок, учитывая все особенности внутри него для получения сжатой версии. Сказанное подтверждается нашими экспериментами. На рис. 2 показаны средние константы по всем конкатенациям файлов для рассматриваемых архиваторов.

Из диаграммы видно, что худшей симметричностью обладают потоковые архиваторы gzip и zip. Наилучший (практически нулевой) результат у блочного архиватора bzip2, что подтверждает теоретический факт, описанный выше.

Свойство дистрибутивности архиватора не сразу интуитивно понятно. Рассмотрим сначала случай, при котором все три файла, участвующие в неравенстве, различны. В этом случае эксперименты показали, что все рассматриваемые архиваторы удовлетворяют неравенству дистрибутивности при всех возможных комбинациях файлов, на которых мы проверяем данное свойство. Теперь предположим, что какие-либо два из трех файлов в неравенстве одинаковы. При этом возможны три варианта:

1) $x = y$ – в этом случае неравенство получается следующим: $C(xx) + C(z) \leq C(xz) + C(xz)$ и неравенство справедливо, если компрессор C удовлетворяет свойствам идемпотентности и монотонности;

2) $y = z$ – в этом случае получившееся неравенство справедливо при условии идемпотентности компрессора C: $C(xy) + C(y) \leq C(xy) + C(yy)$;

3) $x = z$ – при этом получившееся неравенство аналогично предыдущему при условии симметричности компрессора C: $C(xy) + C(x) \leq C(yx) + C(xx)$.

Сведем результаты экспериментов в табл. 1 с условными обозначениями: « + » обозначает хорошее выполнения свойства архиватором, « + /–» удовлетворительное, а «–» – невыполнение свойства. Свойствами нормального архиватора обладают четыре архиватора: 7z, XZ, PAQ и RAR.

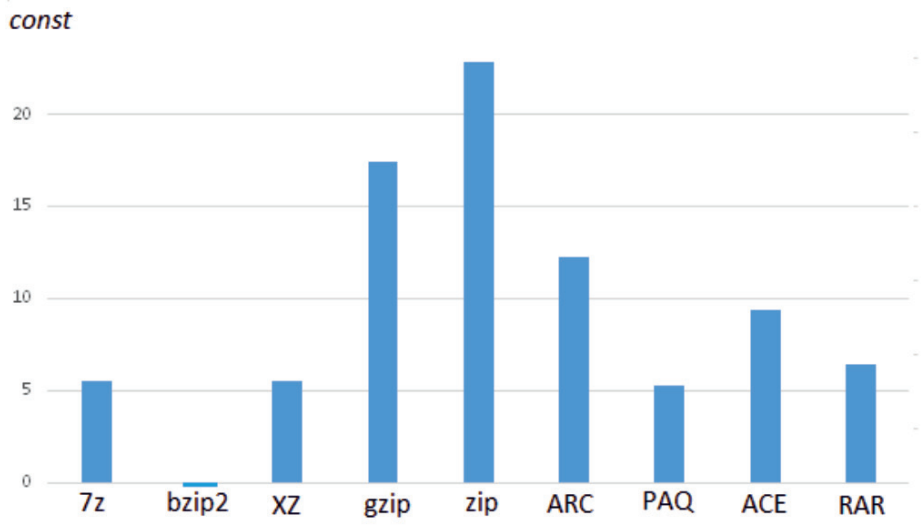

Рис. 2. Результаты проверки симметричности

Выполнение свойств нормального архиватора

	7z	Bzip2	XZ	Gzip	Zip	ARC	PAQ	ACE	RAR
Идемпотентность	+	–	+	–	–	+	+	+ /–	+
Монотонность	+	+	+	+	+	+	+	+	+
Симметричность	+	+	+	+ /–	+ /–	–	+	+ /–	+
Дистрибутивность	+	+	+	+	+	+	+	+	+

Проверка аксиом расстояния для NCD

Аксиомы расстояния были проверены для NCD, вычисляемого по формуле (2) с использованием четырех «нормальных» архиваторов. Для всех четырех случаев все три аксиомы выполняются с требуемой точностью. Покажем это на примере второй аксиомы расстояния, суть которой заключается в том, что расстояние от объекта *x* до объекта *y* должно быть равно расстоянию в обратном направлении. Для проверки этой аксиомы рассматривалась разность между этими расстояниями для каждой пары объектов из выборки, использованной в разделе 2, и вычислялось по модулю среднее отклонение по каждому архиватору. На рис. 3 представлена диаграмма, отображающая результаты экспериментов, из которой видно, что все рассматриваемые архиваторы удовлетворяют второй аксиоме расстояния с приблизительно равной точностью.

В практическом плане можно говорить о правильности результатов экспериментов для нормальных архиваторов 7z, XZ, PAQ и RAR, поскольку уже упомянутая теорема из [5] утверждает, что если при вычислении *NCD* был использован нормальный архиватор, то такое *NCD* удовлетворяет аксиомам расстояния.

Пример кластеризации сайтов с помощью NCD

Здесь приводится пример кластеризации сайтов на основе матрицы расстояний, построенной для *NCD* с использованием нормального архиватора RAR. Взято 15 сайтов четырех популярных компаний по созданию сайтов в Рунете (преимущественно интернет-магазины и корпоративные сайты). Проверялось предположение о том, что у каждой компании по созданию сайтов есть свой стиль, и тогда это будет видно по результатам кластеризации. На рис. 4 приводится дендрограмма результатов кластеризации.

На рисунке указаны порядковые номера сайтов из выборки, их доменные имена и компании-разработчики. Заметно, что все сайты разделились на четыре кластера, причем сайты, разработанные одной компанией, преимущественно содержатся в одном кластере. Исключения подтверждают правило – сайт под номером 11 компании «Медиасфера» по своей структуре действительно больше похож на сайты компании ARTW, в чем можно убедиться визуально, посмотрев сайты.

Результаты кластеризации позволяют сделать вывод, что у каждой компании, разрабатывающей сайты, есть некоторый свой стиль построения сайта. Возможно, в каждой компании используются свои шаблоны, на основе которых в последующем разрабатывается сайт.

Заключение

В статье подробно исследован вопрос о том, какими свойствами должен обладать архиватор (как практический аналог Колмогоровской сложности) и каким образом выбрать наилучший из множества архиваторов. Сформулированы четыре свойства такого нормального архиватора и три аксиомы *NCD*, вычисляемого с помощью такого архиватора.

Проведена серия экспериментов, позволивших из девяти популярных и доступных архиваторов выбрать четыре, обладающих требуемыми свойствами нормальности. Показано, что аксиомы расстояния выполняются для *NCD*, вычисляемого этими четырьмя архиваторами.

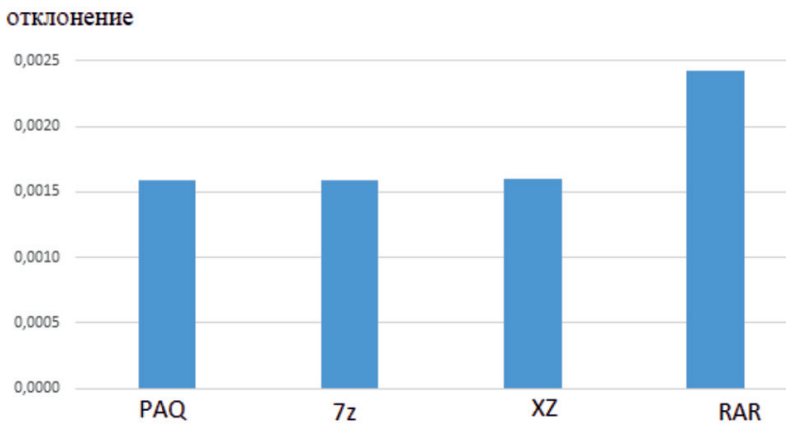

Рис. 3. Результаты проверки аксиомы симметрии

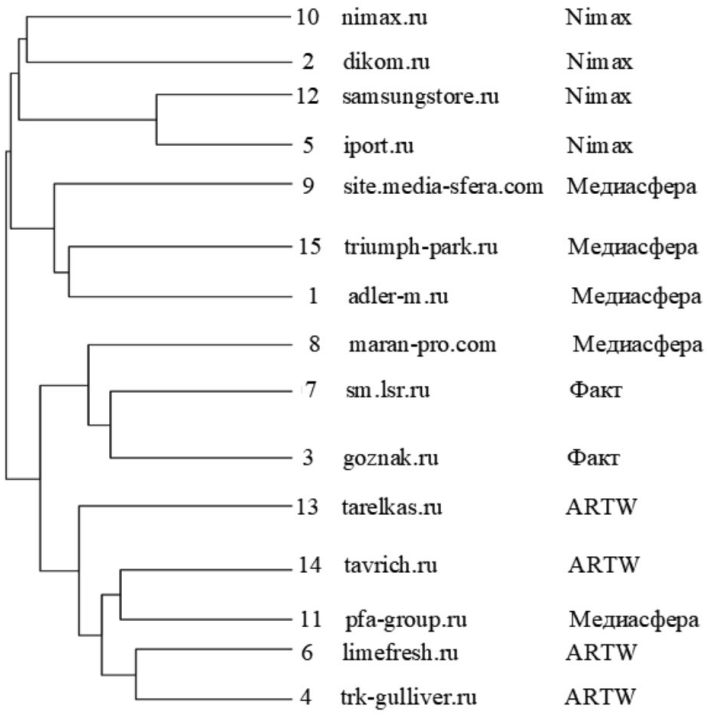

Рис. 4. Дендрограмма кластеризации по методу ближайшего соседа

С использованием одного из нормальных архиваторов, а именно RAR, показана возможность кластеризации выборочного множества сайтов, имеющая хорошую содержательную интерпретацию.

Список литературы

1. Печников А.А. О схожести сайтов и Колмогоровской сложности. Norwegian Journal of development of the International Science. 2018. № 14. Vol. 1. P. 25–29.

2. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. М.: МЦНМО, 2013. 576 с.

3. Chen X., Li M., Li X., Ma B., Vitanyi P. The similarity metric. IEEE Transactions on Information Theory, 2004. vol. 50, no 12. P. 3250–3264.

4. Архиватор. [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Архиватор (дата обращения:6 14.06.2019).

5. Cilibrasi R., Vitanyi P. Clustering by Compression. IEEE Transactions On Information Theory. 2005. vol. 51. no 4. P. 1523–1545.

6. Кириченко В.В. Аналитический обзор алгоритмов сжатия цифровой информации // Проблемы физики, математики и техники. Сер.: Информатика. 2016. № 2 (27). С. 77–83.

7. Спиваковский А.М. Управляемое сжатие данных. СПб.: НИЦ АРТ, 2018. 258 с.