ABOUT THE SIMILARITY OF THE WEBSITES AND KOLMOGOROV COMPLEXITY

Pechnikov A.
Doctor (DSc) of Technics, Assistant Professor, Chief Research Associate, Laboratory for Telecommunications Systems, Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences

О СХОЖЕСТИ ВЕБ-САЙТОВ И КОЛМОГОРОВСКОЙ СЛОЖНОСТИ

Печников А.А.
Доктор технических наук, доцент, главный научный сотрудник, лаборатория телекоммуникационных систем, Институт прикладных математических исследований Карельского научного центра РАН

Abstract

In this paper proposed approach to the study similarity of websites with the use of Kolmogorov complexity and normalized compression distance. A number of experiments show certain potential of this approach, but allow to put a number of tasks that you want to investigate before to automate the conduct of large studies. These tasks are formulated as the main directions of further research.

Аннотация

Предложен подход к исследованию схожести веб-сайтов с использованием Колмогоровской сложности и нормализованного расстояния сжатия. Ряд экспериментов показывают определенный потенциал такого подхода, но при этом ставят ряд задач, которые требуется исследовать, прежде чем автоматизировать проведение больших исследований. Эти задачи сформулированы в виде основных направлений дальнейших исследований.

Keywords: Kolmogorov complexity, normalized compression distance, web site, hierarchical clustering.

Ключевые слова: Колмогоровская сложность, нормализованное расстояние сжатия, веб-сайт, иерархическая кластеризация.

Введение

Веб-сайты похожи друг на друга, особенно сайты организаций, занимающихся одинаковой деятельностью, однако некоторые сайты более похожи, чем другие. Человек-эксперт, сравнивая различные сайты (с целью объединения аналогичных сайтов в одном разделе некоего каталога, или блокирования сайтов с неприемлемым содержимым), будет искать определенные конкретные сходства. Известные попытки автоматизировать этот процесс делают примерно то же самое [1-3]. Вообще говоря, берется сайт и из него извлекаются данные для вычисления различных числовых характеристик, например, частота встречаемости ключевых слов. Затем характеристические векторы, соответствующие различным сайтам, классифицируются или кластеризуются с использованием метода опорных векторов [4], наивного байесовского классификатора [5] и др.

Наша цель является более общей. Мы не ищем сходства во многих конкретных признаках, которые были бы актуальны для классификации веб-сайтов. Хотелось бы найти одну обобщенную характеристику, отражающую взаимодействия различных влияний примерно так, как это описывается в общей теории подобия, основанной на переходе от обычных физических величин, влияющих на моделируемую систему, к обобщённым величинам комплексного типа, зависящих от конкретной природы исследуемого процесса [6].

Нам представляется, что следует очень внимательно исследовать применимость для веб-сайтов в качестве такой обобщенной характеристики нормализованного расстояния сжатия (Normalized Compression Distance – *NCD*) [7,8], базирующегося, в свою очередь, на теоретическом понятии Колмогоровской сложности [9,10]. Исследования показали, что эта универсальная характеристика сходства хорошо работает на конкретных примерах в самых разных областях применения, таких как автоматическое конструирование дерева филогенеза на основе целых митохондриальных геномов [11], классификация музыкальных жанров [12] и анализ текстовых данных в задачах обеспечения кибербезопасности [13]. Мы провели серию экспериментов для выбранного множества веб-сайтов, при этом многие из шагов были выполнены «вручную», без разработки и использования соответствующего программного обеспечения. Основная цель этих экспериментов заключалась в том, чтобы выявить особенности использования *NCD* применительно к исследованиям веб-сайтов и наметить основные вопросы, на которые требуется ответить, и задачи, которые требуется решить, для того, чтобы проводить крупномасштабное исследование.

Статья организована следующим образом. Дается формализованное определение Колмогоровской сложности и нормализованного расстояния сжатия. Далее описывается один из нескольких проведенных небольших экспериментов и дается содержательная интерпретация полученных результатов. Затем, на основе эксперимента, формулируются основные задачи, которые требуется решить, прежде чем заняться разработкой системы

автоматизации исследований схожести веб-сайтов для больших целевых множеств.

Колмогоровская сложность и нормализованное расстояние сжатия

Дадим определение Колмогоровской сложности в точности по [10, стр. 23], этот вариант определения сложности был предложен в статье Колмогорова [9]. Отличие следующих двух абзацев (не взятых в кавычки) от [10] в том, что слово «колмогоровской» мы пишем с заглавной буквы.

Способом описания, или декомпрессором, мы называли произвольное вычислимое частичное отображение D из множества двоичных слов Ξ в себя. (Вычислимость отображения D означает, что есть алгоритм, который применим к словам из области определения отображения D и только к ним; результат применения алгоритма к слову x есть $D(x)$.) Если $D(y)=x$, говорят, что y является описанием x при способе описания D. Для каждого способа описания D мы определяем сложность относительно этого способа описания, полагая её равной длине кратчайшего описания:

$$KS_D(x) = min\{l(y)|D(y)=x\}.$$

При этом минимум пустого множества считается равным ∞. Говорят, что способ описания $D1$ не хуже способа описания $D2$, если найдётся такая константа c, что

$$KS_{D1}(x) \le KS_{D2}(x)+c$$

для всех слов x. Способ описания называют оптимальным, если он не хуже любого другого способа описания.

Теперь зафиксируем некоторый (не обязательно оптимальный) способ описания, и сложность слова x относительно этого способа описания обозначим $K(x)$. Для простоты скажем сразу, что в дальнейшем мы будем ее считать равной числу битов в сжатой версии x. Более того, как уже было сказано, в дальнейшем в качестве отображения D мы будем использовать стандартные программы-архиваторы.

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 как:

$$\mathrm{NCD(x,\,y)} = \frac{\max\{\mathrm{K(x \mid y), K(y \mid x)}\}}{\max\{\mathrm{K(x), K(y)}\}}.$$

Грубо говоря, два объекта считаются близкими, если мы можем значительно "сжать" один из них, учитывая информацию, содержащуюся в другом; идея заключается в том, что если два объекта достаточно похожи, то мы можем более кратко описать один из них, учитывая другой.

В [7] доказывается весьма изощренная теорема о симметрии алгоритмической информации, следствием из которой является примерное равенство $K(x|y)\approx K(yx)–K(y)$, здесь yx обозначает конкатенацию двоичных строк y и x.

Тогда так же как в [8], с учетом того, что опять-таки на практике $K(xy) \approx K(yx)$, для приближенных вычислений, проводимых в дальнейшем, NCD можно записать так:

$$\mathrm{NCD(x,\,y)} = \frac{\mathrm{K(xy)} - \min\{\mathrm{K(x), K(y)}\}}{\max\{\mathrm{K(x), K(y)}\}} \qquad (1).$$

Понятно, что расстояние $NCD(x,y)$ симметрично, а в [14] показано, что это действительно метрика, поскольку аксиомы тождества и треугольника также выполняются. Более того, там же показано, что метрика NCD универсальна в том смысле, что для каждой нормализованной метрики f и каждого x и y мы имеем $d(x,y)\le f(x,y)+O(log(k)/k)$, где $k=max\{K(x),K(y)\}$. Неформально говоря, это означает, что если x и y близки по любой «хорошо ведущей себя» метрике f, то x и y также близки по универсальной метрике NCD. Именно метрику $NCD(x,y)$ мы будем использовать далее.

Простой эксперимент

Опишем результаты одного из ряда очень простых проведенных экспериментов, результаты которого хорошо интерпретируемы в содержательном плане. Здесь множество состоит из 10 известных нам веб-сайтов, сведенных в таблицу 1.

Сравнение сайтов было проведено по их первым (начальным) страницам, для чего они в обычном браузере были сохранены в виде файлов в формате *.txt. Это самый простой и интуитивно понятный вариант, посокльку текстовый файл содержит последовательность символов, объединённых в строки, и в этом смысле наиболее близок к бинарному файлу, который, в свою очередь, может напрямую интерпретироваться как двоичное слово (см. раздел выше).

Табл. 1

Множество веб-сайтов.

№	Организация	Сайт
1	Карельский научный центр РАН (КарНЦ РАН)	www.krc.karelia.ru
2	Институт прикладных математических исследований КарНЦ РАН	mathem.krc.karelia.ru
3	Институт языка, литературы и истории КарНЦ РАН	illhportal.krc.karelia.ru
4	Институт геологии КарНЦ РАН	ig.krc.karelia.ru
5	Институт геологии КарНЦ РАН-2	igkrc.ru
6	Петрозаводский государственный университет	petrsu.ru
7	Санкт-Петербургский государственный университет	spbu.ru
8	Московский государственный университет	www.msu.ru
9	Интернет-газета Столица на Онего	stolicaonego.ru
10	Гостиница "Онего Палас"	onegopalace.com

Сжатие этих файлов выполнено архиватором RAR 5.50 (https://www.rarlab.com). Объем RAR-архива сайта с номером *i* принимается в качестве *K(i)*. Конкатенация файлов для сайтов с номерами *i* и *j* произведена простым объединением содержимого файлов с помощью редактора текстов, объем RAR-архива объединенного файла равен *K(ij)*. По формуле (1) вычисляется *NCD(i,j)* – расстояние между сайтами *i* и *j*.

Например, для сайта 1 (www.krc.karelia.ru) объем текстового файла начальной страницы равен 55592 байт, его архив – 10657 байт, а для сайта 2 (mathem.krc.karelia.ru) 41130 и 7263 байт соответственно. Для сайтов с номерами 1 и 2 *K(12)*=15031 байт. Тогда по формуле (1) расстояние *NCD*(1,2) = *NCD*(2,1) = 0,728910575. Числовые значения элементов матрицы {*NCD(i,j)*}, *i,j*=1,..10 приведены в таблице 2 (матрица симметрична относительно главной диагонали).

Табл. 2

Матрица расстояний.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	0,729	0,740	0,758	0,883	0,929	0,897	0,927	0,955	0,898
2		0	0,555	0,586	0,855	0,937	0,918	0,940	0,962	0,913
3			0	0,536	0,867	0,941	0,940	0,947	0,965	0,937
4				0	0,780	0,943	0,941	0,947	0,963	0,940
5					0	0,945	0,942	0,944	0,963	0,937
6						0	0,915	0,900	0,949	0,927
7							0	0,935	0,963	0,907
8								0	0,943	0,893
9									0	0,955
10										0

Дендрограммы, построенные по методу ближайшего соседа и методу Варда [15] применительно к данной матрице расстояний, приводятся на рисунках 1 и 2.

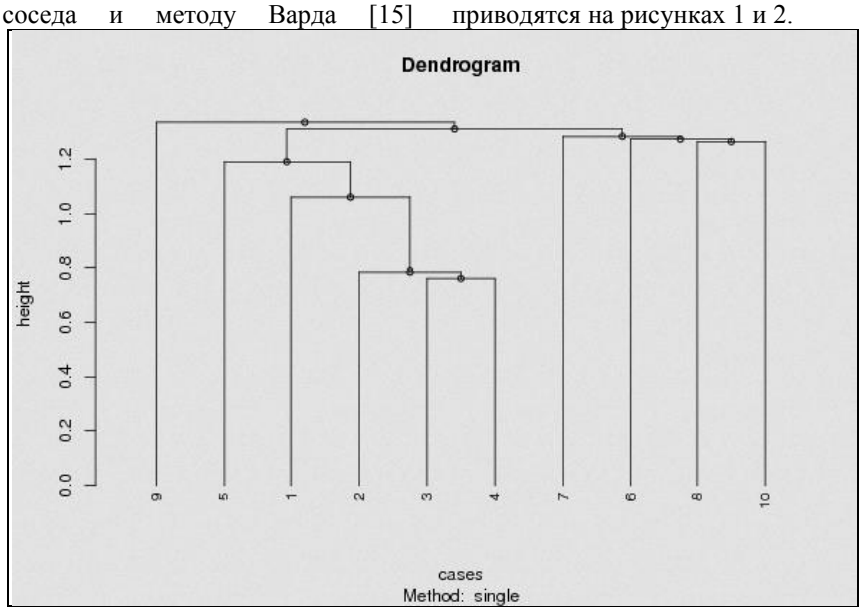

Рис. 1. Дендрограмма по методу одиночной связи для {NCD(i,j)}, i,j=1,..10.

Очевидно объединение в том и другом случае в кластеры четырех сайтов КарНЦ РАН (1-4) и трех университетских сайтов (6-8), к которым (по неведомым причинам) примкнул сайт гостиницы (10). Однако в первом случае сайт с номером 5 (igkrc.ru) попадает в кластер с другими сайтами КарНЦ РАН, а во втором – весьма далек от него.

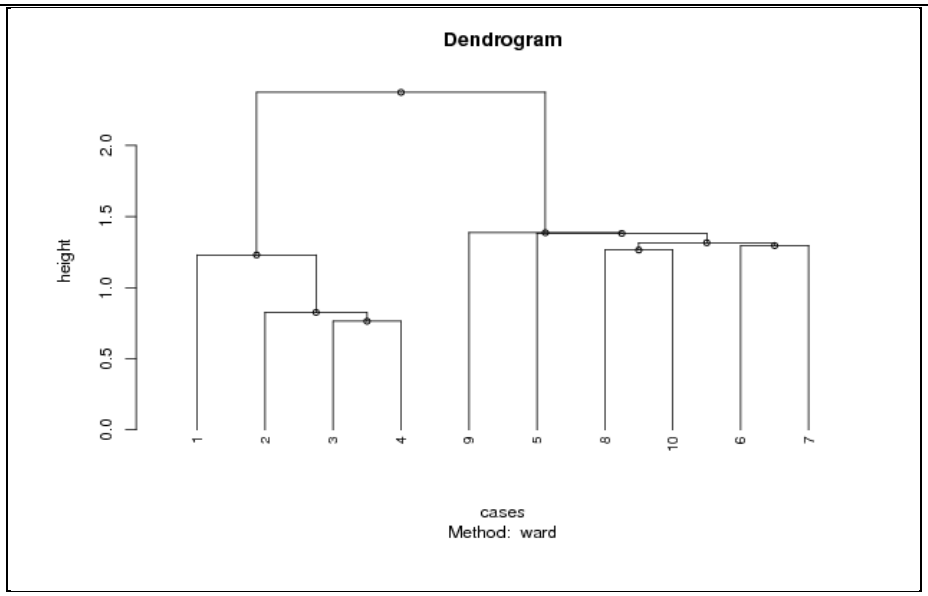

Рис. 2. Дендрограмма по методу Варда для {NCD(i,j)}, i,j=1,..10.

Интересно, что содержательное объяснение можно дать как для первого, так и для второго случая. Дело в том, что сайты 1-4 разработаны Лабораторией информационных компьютерных технологий Института прикладных математических исследований КарНЦ РАН, а сайт 5 (второй сайт Института геологии КарНЦ РАН) создан Студией Медиавеб (https://mediaweb.ru). И первая дендрограмма характеризует принадлежность сайтов к одному виду деятельности, а вторая – идентифицирует разработчиков. В любом случае можно сказать, что обе дендрограммы показывают схожесть сайтов достаточно близко к «интуитивному» представлению. Поэтому описанный подход представляется перспективным и требующим дополнительных исследований на гораздо большем целевом множестве сайтов.

Например, в качестве такового можно было бы взять множество официальных веб-сайтов научных учреждений Федерального агентства научных организаций и вузов в ведении Министерства образования России. «Ручное» исследование в этом случае провести затруднительно, поскольку множество содержит более 900 сайтов. Поэтому необходимо автоматизировать этапы сбора данных о сайтах, сжатия этих данных, построения матрицы расстояний и кластерного анализа. И здесь возникает ряд вопросов, появившихся в процессе проведения «простых» экспериментов, которые требуют исследований и ответов на них, прежде чем заниматься автоматизацией. На них мы остановимся в следующем разделе.

Вопросы для дальнейших исследований

Первый вопрос. Что должна представлять собой (первая) скачанная страница сайта? – Или в более общем варианте: что должны представлять собой данные о сайте, собираемые для дальнейшего решения задачи схожести сайтов? Стандартные браузеры и специализированные программы скачивания (офф-лайн браузеры) предлагают на выбор различные варианты. Например, кроме текстового представления возможно сохранение веб-страницы в виде файла с расширением *.htm и связанной с ним папки, в которую будут помещены относящиеся к странице изображения и стили.

Но тогда надо исследовать ***второй вопрос***: каким образом папку с файлом *.htm и связанную с ним папку изображений и стилей интерпретировать как двоичное слово?

Третий вопрос следует из первых двух: сколько страниц сайта надо скачивать для того, чтобы исследование было адекватным? – Понятно, что речь не идет об одной странице, но, наверное, и не обо всем сайте (поскольку это может быть слишком затратно). Тут представляется перспективным следующий подход. Пусть *K(s(M))* – объем архива сайта *s*, с которого скачано *M* страниц, а *K(s(M+1))* – объем архива этого же сайта, с которого скачана *M+1* страница. Тогда можно вычислить *NCD(s(M),S(M+1))* – расстояние между двумя версиями сайтов, отличающихся на 1 страницу. Необходимо исследовать поведение ряда *NCD(s(1),S(2)), NCD(s(2),S(3)), …, NCD(s(M),S(M+1))*, … на стремление его членов к нулю с ростом *M* и вычислению некоторой оптимальной точки останова скачивания. Результаты эксперименты, вручную проведенные для первых страниц сайта www.krc.karelia.ru, (0.122, 0.015, 0.025, 0.013, …) выглядят обнадеживающе.

Четвертый вопрос: можно ли выбрать наилучший архиватор, и какими свойствами он должен обладать? По поводу «наилучшего» архиватора ответ далеко не очевиден. По поводу его свойств надо, по крайней мере, исследовать свойства идемпотентности, монотонности, симметричности и дистрибутивности.

Пятый вопрос: очевидно, что методы иерархической кластеризации применимы для относительно небольшого числа объектов. Необходимо провести анализ методов как с точки зрения их применимости для больших множеств объектов, так и с точки зрения интерпретации получаемых результатов.

Заключение

В статье описан подход к исследованию проблемы схожести веб-сайтов, основанный на Колмогоровской сложности и нормализованном расстоянии сжатия. Небольшие эксперименты показывают определенный потенциал такого подхода, но при этом обязывают поставить ряд задач, которые требуется исследовать, прежде чем автоматизировать проведение больших исследований. Эти задачи сформулированы в виде основных направлений дальнейших исследований.

Благодарности

Автор приносит глубокую благодарность своей любимой жене, ***Печниковой Галине Васильевне***, которая не только стойко терпела его работу над статьей в праздничные новогодние дни 2018 года, но и всячески поощряла его.

СПИСОК ЛИТЕРАТУРЫ:

1. Маслов М.Ю., Пяллинг А.А., Трифонов С.И. Автоматическая классификация веб-сайтов // Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия. 2008. С. 230-235.
2. Kriegel H.-P., Schubert M. Classification of Websites as Set of Feature Vectors // Proceedings of the International Conference Databases and Applications. – IASTED, Innsbruck, Austria, February 17-19. 2004. P. 127-132.
3. Ajay S. Patil, Pawar B.V. Automated Classification of Web Sites using Naive Bayesian Algorithm // Proceedings of the International MultiConference of Engineers and Computer Scientists. - IMECS 2012, Hong Kong, March 14-16, 2012. Vol 1. P. 519-523.
4. Ben-Hur A., Horn D., Siegelmann H., Vapnik V. Support vector clustering // Journal of Machine Learning Research. 2001. Vol. 2. P. 125–137.
5. Domingos P., Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss // Machine Learning. 1997. Vol. 29, Iss. 2-3, November 1997. P. 103-137.
6. Гухман А. А. Введение в теорию подобия – М.: Высшая школа, 1973. – 296 с.
7. Ming Li, Vitanyi P. An Introduction to Kolmogorov Complexity and Its Applications. – 3rd ed. New York: Springer-Verlag, 2008. – 809 p.
8. Cilibrasi R., Vitanyi P. Clustering by compression // IEEE Transactions on Information Theory. 2005. Vol. 51, Iss.4, P. 1523-1545.
9. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. 1965. т. 1, вып. 1. с. 3-11.
10. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. – М.: МЦНМО. – 2013. – 576 с.
11. Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, Haoyong Zhang An information-based sequence distance and its application to whole mitochondrial genome phylogeny // Bioinformatics. 2001. Vol. 17, no 2. P. 149-154.
12. Cilibrasi R., Vitanyi P., de Wolf R. Algorithmic Clustering of Music Based on String Compression // Computer Music Journal. Vol. 28, Iss. 4, Winter 2004. P. 49-67.
13. Суркова А.С. Анализ и моделирование текстовых данных в задачах обеспечения кибербезопасности // Системы управления и информационные технологии. №3.1(61). 2015. С. 178-182.
14. M. Li, X. Chen, X. Li, B. Ma, P. Vitanyi The similarity metric // IEEE Transactions on Information Theory. 2004. Vol.50, Iss. 12. P. 3250-3264.
15. Жамбю М. Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988. – 345 с.