О сходстве веб-сайтов

А.А. Печников[1], А.С. Головин[1]

[1] *Федеральный исследовательский центр "Карельский научный центр Российской академии наук"*

Аннотация. Рассматривается подход к исследованию сходства веб-сайтов с использованием Колмогоровской сложности и нормализованного расстояния сжатия. Описано применение данного подхода к анализу так называемого «множества сайтов оболочки» на предмет сходства его сайтов с сайтами целевого множества концептуальной модели тематического фрагмента Веба. Проведенные эксперименты показывают определенный потенциал такого подхода, но при этом ставят ряд вопросов, определяющих дальнейшие направления исследований.

Ключевые слова: Колмогоровская сложность, нормализованное расстояние сжатия, веб-сайт, сходство объектов, концептуальная модель фрагмента Веба.

On the similarity of the websites

A.A. Pechnikov[1], **A.S. Golovin**[1]

[1] *Karelian Research Centre of the Russian Academy of Sciences*

Abstract. We consider an approach to the study of the similarity of websites using Kolmogorov complexity and normalized compression distance. The article describes the application of this approach to the analysis of the so-called "set of shell sites" for the similarity of its sites with the sites of the target set of the conceptual model of the thematic fragment of the Web. The experiments show a certain potential of this approach, but at the same time raise a number of questions that determine the future direction of research.

Keywords: Kolmogorov complexity, normalized compression distance, web site, the similarity of objects, conceptual model of the Web fragment.

CEUR-WS.org/Vol-2260/2_417-427.pdf

Введение

Многие веб-сайты похожи друг на друга, особенно сайты организаций, занимающихся одинаковой деятельностью, однако некоторые сайты более похожи, чем другие. Мы не ищем сходства во многих конкретных признаках, которые были бы актуальны для классификации веб-сайтов [1-3]. 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 [4].

Возможным направлением таких исследований представляется использование нормализованного расстояния сжатия (Normalized Compression Distance – *NCD*) [5], базирующегося, в свою очередь, на теоретическом понятии Колмогоровской сложности [6]. Показано, что эта характеристика сходства работает в самых разных областях применения, таких как автоматическое конструирование дерева филогенеза на основе целых митохондриальных геномов, классификация музыкальных жанров, анализ текстовых данных в задачах обеспечения кибербезопасности [7-9].

Применение такого подхода в вебометрике пока порождает гораздо больше вопросов, чем имеющихся у нас ответов, хотя и даёт определенные надежды [10].

Мы пошли по пути создания программы, реализующей этот подход, что позволяет проводить эксперименты и пытаться более чётко формулировать возникающие вопросы и ответы на них.

Здесь мы опишем применение данного подхода к анализу так называемого «множества сайтов оболочки» на предмет сходства его сайтов с сайтами целевого множества концептуальной модели тематического фрагмента Веба.

1. Множество сайтов оболочки концептуальной модели фрагмента Веба

Краткое описание концептуальной модели фрагмента Веба дано, например, в [11]. Регламентируемый сайт – это официальный веб-сайт организации, для которого существует нормативный акт, определяющий цели его создания, структуру, основные разделы, правила наполнения и обновления информации, ответственных лиц и т.д.

Регламентируемое тематическое целевое множество T – множество регламентируемых сайтов одной тематики (например, сайты университетов России). Сопутствующее множество U – это множество сайтов, не входящих в целевое множество, на которые существуют ссылки с целевого множества (и не факт, что существуют обратные ссылки).

Таким образом, концептуальная модель фрагмента Веба – это веб-граф, вершинами которого являются сайты множеств T и U, а дугами – гиперссылки, соединяющие эти вершины.

В свою очередь U разбивается на три непересекающихся подмножества:

B – множество сайтов ближайших окрестностей сайтов целевого множества (например, сайт факультета по отношению к сайту университета),

K – множество сайтов-коммуникаторов (сайты из U, на которые сделано много ссылок с сайтов T и/или много ссылок с них на сайты T),

S – множество сайтов оболочки, не являющиеся сайтами ближайших окрестностей или веб-коммуникаторами, то есть $S=U\setminus(B\cup K)$.

Одна из проблем реализации концептуальной модели для заданного целевого множества заключается в том, что сайтов оболочки очень много, в сотни и тысячи раз больше, чем сайтов целевого множества, и далеко не все они имеют отношение к концептуальной модели. Следует сказать, что в большинстве случаев это, как и в [12] «… сайты менее известные, менее наполненные содержанием», а «… в интернет-каталогах редакторы описывают в основном известные и/или богатые содержанием сайты». Как правило, такие сайты в рамках концептуальной модели являются «висячими» вершинами, не имеющими ссылок на сайты целевого множества и не представляющими предметного интереса.

Поэтому хотелось бы уже на начальном этапе построения оставить только те сайты S, которые «похожи» на сайты целевого множества, что позволяет существенно экономить вычислительные ресурсы в дальнейшем.

2. О Колмогоровской сложности и нормализованном расстоянии сжатия

Дадим в следующих двух абзацах без кавычек определение Колмогоровской сложности в точности по [13, стр. 23].

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 $KS_{D1}(x) \leq KS_{D2}(x)+c$ 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 $K(x)$. В дальнейшем мы будем ее считать равной числу битов в сжатой версии x, а в качестве отображения D будем использовать стандартные программы-архиваторы.

Пусть y – еще одно двоичное слово. Следуя [5] 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

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

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 [14] доказывается 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 как в [5], с учетом того, что на практике $K(xy) \approx K(yx)$, для приближенных вычислений, проводимых в дальнейшем, NCD можно записать так

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

Расстояние $NCD(x,y)$ симметрично, аксиомы тождества и треугольника также выполняются.

3. Об одном способе вычисления нормализованного расстоянии сжатия для сайтов

Опишем один из способов вычисления расстояния (1) между двумя сайтами. Возьмем два сайта и скачаем у каждого по несколько страниц, которые объединим в файлы в формате *.txt. Это самый простой и интуитивно понятный вариант, поскольку текстовый файл содержит последовательность символов, объединённых в строки, и в этом смысле наиболее близок к бинарному файлу, который, в свою очередь, может интерпретироваться как двоичное слово.

Выполним сжатие этих файлов, например, архиватором RAR 5.50 (https://www.rarlab.com). Объем RAR-архива сайта с номером i принимается в качестве $K(i)$. Конкатенацию файлов для сайтов с номерами i и j произведем простым объединением содержимого файлов с помощью редактора текстов, объем RAR-архива объединенного файла равен $K(ij)$. По формуле (1) вычислим $NCD(i,j)$ – расстояние между сайтами i и j.

В качестве примера возьмем первые страницы сайтов *www.krc.karelia.ru* (Карельский научный центр РАН), *keldysh.ru* (Институт прикладной математики им. М.В. Келдыша РАН) и ресурса с URL *agora.guru.ru/abrau2018* (Конференция «Научный сервис в сети Интернет-2018»). Расстояния между ними проиллюстрированы на рис. 1.

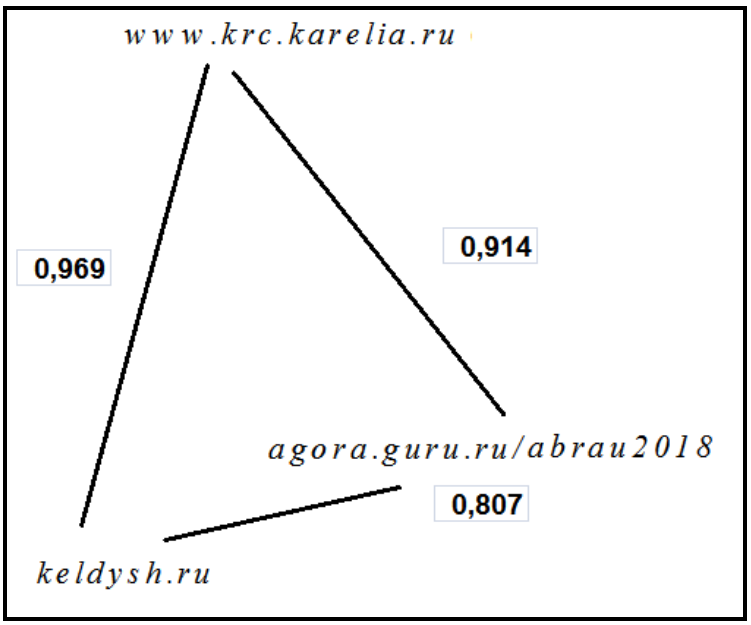

Рис. 1. Расстояния *NCD* между тремя страницами.

Вопросы о том, сколько страниц сайта надо скачать для получения в некотором смысле адекватных результатов, а также чем представление txt предпочтительнее (или хуже) html, пока оставим открытыми.

4. Об определении сходства сайтов оболочки и сайтов целевого множества

Рассмотрим следующую процедуру определения сходства сайтов оболочки и целевого множества. Построим проверочное множество сайтов *AD* как объединение двух подмножеств *A* и *D*, причем в подмножество *A* входят некоторые (возможно, все) сайты целевого множества, а в подмножество *D* некоторые сайты оболочки, очевидно не схожие с сайтами целевого множества.

Для сайтов, входящих в *AD*, можно построить матрицу $\{NCD(i,j)\}$, $i,j \in AD$. Основным свойством проверочного множества должно быть его разбиение одним из методов иерархической кластеризации по вычисленной матрице расстояний $\{NCD(i,j)\}$ не менее чем на 2 кластера, причем в один кластер должны попасть все сайты из *A*, а во все остальные кластеры – сайты из *D*.

Теперь добавим к проверочному множеству испытуемый сайт *x* из множества-оболочки, не входящий в подмножество *B* и вычислим матрицу $\{NCD(i,j)\}^+$ для множества $AB \cup x$.

Решающее правило сходства сформулируем следующим образом:

- если в результате кластеризации по матрице $\{NCD(i,j)\}^+$ сайт *x* попадает в один кластер с сайтами из *A* (даже если они разобьются более чем на один кластер), то он схож с сайтами целевого множества,
- если же сайт *x* попадет в кластер с сайтами из *D*, то он несхож с сайтами целевого множества.

5. Программная реализация

Рабочая версия программы, реализующей описанные выше способы построения матрицы расстояний и иерархической кластеризации, находится в открытом доступе по адресу http://kolmogorov.boincfast.ru. Приводится пример файла исходных данных со списком сайтов в формате *.xls. Можно указать, в каком виде скачиваются и обрабатываются страницы сайта (*.txt или *.html) и количество обрабатываемых страниц.

После загрузки файла исходных данных строится матрица нормализованного расстоянии сжатия между сайтами, которую можно загрузить на локальный компьютер для дальнейшей обработки. При этом на экран выводится список сайтов с указанием успешной обработки, либо ошибки, по которой сайт не был обработан (не обработанные сайты не войдут в матрицу расстояний).

Полученная матрица может быть использована здесь же для проведения экспериментов по иерархической кластеризации с указанием количества кластеров и методов их построения.

6. Эксперимент с сайтами фрагмента Веба Крымского федерального университета

Опишем небольшой, но характерный эксперимент с определением сходства сайтов оболочки и целевого множества для модели фрагмента Веба Крымского федерального университета, состоящего в силу исторических причин из достаточно «молодых» веб-сайтов.

Целевое множество Крымского федерального университета (КФУ; на 05 апреля 2018 г.) состоит примерно из 120 сайтов, среди которых официальный сайт КФУ, сайты структурных подразделений и филиалов, тематические сайты, сайты журналов и конференций и др. Количество внутренних гиперссылок, связывающих эти сайты, приближается к 1300. Общее количество гиперссылок, сделанных с сайтов КФУ на 660 сайтов оболочки – около 4000.

Мы сформировали подмножество A проверочного множества из 18 сайтов целевого множества, входящих в наиболее крупный кластер в случае иерархического разбиения всего целевого множества: КФУ (cfuv.ru), Таврическая академия (ta.cfuv.ru), Приемная комиссия КФУ (abitur.cfuv.ru), Научный портал (science.cfuv.ru), Медицинский колледж (medcoll.cfuv.ru), Таврический колледж (college.cfuv.ru) и т.д.

В подмножество D были включены 4 сайта, занимающие ведущие позиции по числу ссылок, сделанных на них с сайтов целевого множества КФУ и известные по их адресам (www.google.com, connect.mail.ru, share.yandex.ru, www.livejournal.com).

Для тестового множества были произвольным образом отобраны 50 сайтов из списка сайтов оболочки: на этапе отбора известны только доменные имена сайтов, а не их названия. Затем сайты «прямым просмотром» идентифицировались по названиям и сходству с сайтами целевого множества.

Само установление сходства, конечно, носит субъективный характер, сайт вуза идентифицируется как очевидно сходный, сайт библиотеки видимо тоже, а вот сайт олимпиады по программированию не всегда. Тем не менее, как есть, так есть, и построенное тестовое множество содержит 19 несхожих и 25 схожих сайтов (6 сайтов из 50 оказалось неработающими).

Для случая скачивания по 1 странице с сайтов проверочного и тестового множеств получены следующие результаты. Из 44 сайтов правильно были идентифицирован 31 сайт, при этом все 25 несхожих сайтов были определены как несхожие, а из 19 схожих сайтов 6 были идентифицированы как схожие и 13 как несхожие.

Для случая скачивания по 10 страниц с каждого сайта все 25 несхожих сайтов были определены как несхожие, а из 19 схожих сайтов 10 были идентифицированы как схожие и 9 как несхожие.

Отметим, что результаты неплохие и наблюдается некоторая положительная динамика результатов в случае увеличения глубины скачивания.

7. Глубина скачивания

Здесь мы уже переходим к ***первому*** очевидно возникающему вопросу. Ответа на него пока нет, но есть некоторые соображения и результаты экспериментов.

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. Построим ряд *NCD(s(1),S(2))*, *NCD(s(2),S(3))*, …, *NCD(s(M),S(M+1))*, … и посмотрим его график для Института прикладных математических исследований Карнц РАН (см. рис. 2).

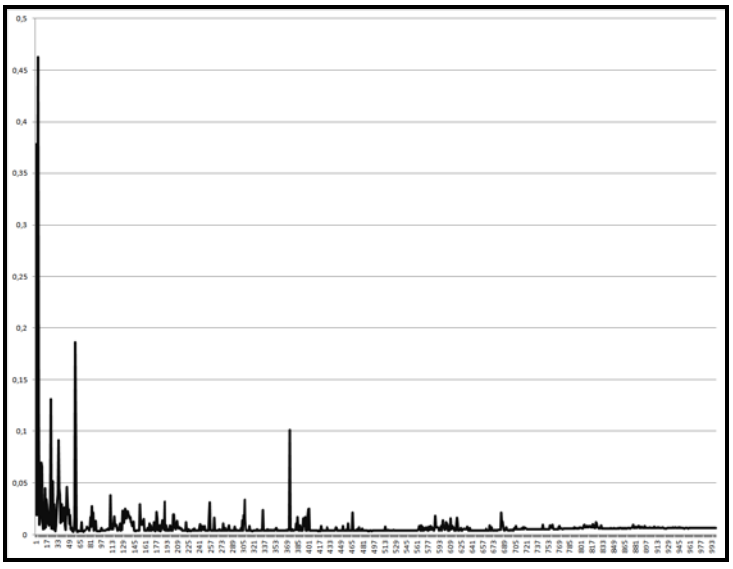

Рис. 2. Ряд *NCD* для 1000 страниц сайта.

Интересно, что «выброс» на 375 странице сайта (http://mathem.krc.karelia.ru/info/gost_7.32-2001.html) имеет очень простое объяснение – это ГОСТ, размещенный на сайте в формате html.

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 («мало скачивать – неадекватно, много скачивать – затратно»).

8. Другие вопросы

Второй 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 сходстве сайтов? Стандартные браузеры и специализированные программы скачивания 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 ***третий*** вопрос: каким образом папку с файлом *.htm и связанную с ним папку изображений и стилей интерпретировать как двоичное слово?

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: что произойдет, если подмножество A, сформировать из сайтов оболочки, очевидно схожих с сайтами целевого множества? – Возможно, результаты улучшатся?

И это далеко не все вопросы.

Заключение

Описан подход к исследованию проблемы сходства веб-сайтов, основанный на Колмогоровской сложности и нормализованном расстоянии сжатия. Показана актуальность подхода для вебометрических исследований, в частности, для построения более точной и концептуальной модели фрагмента Веба.

Можно наметить и другие направления использования результатов проводимых исследований. Одно из них навеяно «выбросом», приведенным в разделе 7: возможно, таким методом можно отыскивать взломанные или умышленно искаженные страницы сайтов (автор может вспомнить случаи десятилетней давности, связанные со страницами на официальных сайтах вузов, содержащие «фермы» ссылок на проносайты).

Разработана рабочая версия программы, представленная в открытом доступе, и реализующая основные моменты экспериментальных исследований (https://kolmogorov.boincfast.ru).

Эксперименты говорят об определенном потенциале такого подхода, но при этом обязывают поставить ряд задач, сформулированных выше в виде вопросов и попыток ответов на некоторые из них, представляющих очевидные направления дальнейших исследований.

Финансовое обеспечение исследований осуществлялось из средств федерального бюджета на выполнение государственного задания КарНЦ РАН (Институт прикладных математических исследований КарНЦ РАН) и при финансовой поддержке РФФИ (проект № 18-07-00628 А).

Литература

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. Гухман А. А. Введение в теорию подобия. – М.: Высшая школа, 1973. 296 с.
5. Cilibrasi R., Vitanyi P. Clustering by compression // IEEE Transactions on Information Theory. 2005. Vol. 51, Iss.4, P. 1523-1545.
6. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. 1965. т. 1, вып. 1. С. 3-11.
7. 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.
8. 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.
9. Суркова А.С. Анализ и моделирование текстовых данных в задачах обеспечения кибербезопасности // Системы управления и информационные технологии. №3.1(61). 2015. С. 178-182.

10. Печников А.А. О схожести сайтов и колмогоровской сложности // Norwegian Journal of development of the International Science. 2018. № 14. Vol.1. С. 25-29.
11. Печников А.А. Концептуальная модель фрагмента Веба и примеры её реализации // Информационная среда вуза XXI века: материалы IV Всероссийской научно-практической конференции (20-24 сентября 2010 г.). Петрозаводск, 2010. С. 172-173.
12. Маслов М. Ю., Пяллинг А.А. КС-классификатор и дорожка классификации веб-сайтов РОМИП'2010 // Труды РОМИП 2010. Под ред. И.С. Некрестьянова. — Казань, 2010. С. 80-97.
13. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. – М.: МЦНМО. – 2013. – 576 с.
14. Ming Li, Vitanyi P. An Introduction to Kolmogorov Complexity and Its Applications. – 3rd ed. New York: Springer-Verlag, 2008. – 809 p.

References

1. Maslov M.Yu., Pjalling A.A., Trifonov S.I. Avtomaticheskaja klassifikacija web-saitov // Trudy 10-i Vserossiiskoi nauchnoi konferencii "Elektronnye biblioteki: perspektivnye metody i tehnologii, elektronnye kollekcii" – RCDL'2008. Dubna. Rossija. S. 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. Guhman A.A. Vvedenie v teoriju podobija. – M.: Vysshaja shkola, 1973. 296 s.
5. Cilibrasi R., Vitanyi P. Clustering by compression // IEEE Transactions on Information Theory. 2005. Vol. 51, Iss.4, P. 1523-1545.
6. Kolmogorov A.N. Tri podhoda k opredeleniju ponjatija "kolichestvo informacii" // Problemy peredachi informacii. 1965. t. 1., vyp. 1., S. 3-11.
7. 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.
8. 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.
9. Surkova A.S. Analiz I modelirovanie tekstovyh dannyh v zadachah obespechenija kiberbezopasnosti // Sistemy upravlenija I informacionnye tehnologii. №3.1(61). 2015. S. 178-182.

10. Pechnikov A.A. O shojesti saitov i kolmogorovskoi slojnosti // Norwegian Journal of development of the International Science. 2018. № 14. Vol.1. C. 25-29.
11. Pechnikov A.A. Konceptual’naja model’ fragmenta Weba i primery ee realizacii // Informacionnaja sreda vuza XXI veka: materialy IV Vserossiiskoi nauchno-prakticheskoi konferencii (20-24 sentjabrja 2010 g.). Petrozavodsk, 2010. S. 172-173.
12. Maslov M.Yu., Pjalling A.A. KS-klassifikator I dorojka klassifikacii web-saitov ROMIP’2010 // Trudy ROMIP 2010. Pod. red. I.S. Nekrestjanova – Kazan’, 2010. S. 80-91.
13. Vereschagin N.K., Uspenskii V.A., Shen’ A. Kolmogorovskaja slojnost’ i algoritmicheskaja sluchainost’. – M., MCNMO. – 2013. – 576 s.
14. Ming Li, Vitanyi P. An Introduction to Kolmogorov Complexity and Its Applications. – 3rd ed. New York: Springer-Verlag, 2008. – 809 p.