На правах рукописи

РУБЦОВ
Антон Геннадьевич

Восстановление отсутствующих данных в символьных последовательностях

05.13.18 – математическое моделирование, численные методы и комплексы программ

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Красноярск 2010

Работа выполнена в Институте вычислительного моделирования Сибирского отделения РАН, г. Красноярск

Научный руководитель: кандидат физико-математических наук Сенашова Мария Юрьевна

Официальные оппоненты: доктор технических наук Миркес Евгений Моисеевич

доктор физико-математических наук, профессор Смирнова Елена Валентиновна

Ведущая организация: ГОУ ВПО «Сибирский государственный аэрокосмический университет им. М.Ф. Решетнева», г. Красноярск

Защита состоится 19 февраля 2010 года на заседании диссертационного совета ДМ 212.099.06 при Сибирском федеральном университете по адресу: 660074, г. Красноярск, ул. Киренского, 26.

С диссертацией можно ознакомиться в научной библиотеке Сибирского федерального университета по адресу: г. Красноярск, ул. Киренского, 26, Г 274.

Автореферат разослан «19» января 2010 года.

Ученый секретарь
диссертационного совета Р.Ю. Царев

Общая характеристика работы

Актуальность. Подавляющее большинство методов обработки и анализа данных работают только с полными данными. На практике очень часто встречаются ситуации, когда какая-либо часть данных отсутствует. Для того чтобы иметь возможность обрабатывать такие данные, необходимо отсутствующие данные восстановить. Существуют различные подходы к этой задаче. Большинство методов позволяет получить значения для пропущенных данных, исходя из принципа максимальной похожести восстановленных данных на полные имеющиеся данные.

При проведении статистического анализа на практике ограничиваются анализом не всей генеральной совокупности в целом, а лишь некоторого выборочного числа наблюдений. Анализируемая выборка должна отвечать критериям качественности и полноты. Но в ситуациях, когда некоторые свойства у исследуемых объектов отсутствуют, происходит смещение основных статистических характеристик. Например, смещения математического ожидания и дисперсии возрастают прямо пропорционально числу пропусков. То есть ошибка напрямую зависит от количества отсутствующих данных. Причиной таких пропусков могут послужить, например, отсутствие значений вследствие каких-то мелких поломок оборудования, не связанных с экспериментальным процессом, или нежелание респондента при проведении статистического опроса отвечать на вопросы о своих доходах.

Знание механизма, приводящего к отсутствию значений, является ключевым при выборе методов анализа и интерпретации результатов. Неполные данные несут в себе новую информацию для исследования, важность которой может быть велика. Поэтому ее следует включать в анализ.

Практически все методы восстановления данных используют аппарат теории вероятности и математической статистики. Как правило, подобные методы восстанавливают пропущенные данные, представимые в какой-нибудь специальной форме, например, в виде таблиц. К тому же, как говорилось ранее, эти данные должны удовлетворят критериям качественности и полноты. Это достаточно жесткое ограничение для практического применения.

Работы, посвященные восстановлению пропущенных данных, в основной своей массе посвящены многомерным данным. В этих работах объекты (данные) представляются точкой в многомерном пространстве, а параметры объекта являются координатами этой точки. При этом для восстановления пропущенных координат зачастую требуется некоторая априорная информация.

Символьные последовательности – это классические объекты математики и встречаются как предмет изучения во многих прикладных задачах, от теоретического программирования и теории управления до биологии и

лингвистики. Изучение символьных последовательностей позволяет ответить на множество вопросов из различных областей науки, связанных с чистой или прикладной математикой.

Как объект прикладного исследования, символьные последовательности возникают во всех областях, где рассматриваются те или иные объекты, состоящие из большого числа одинаковых фрагментов. При этом схожесть или подобие могут носить искусственный характер. Исследователь вправе по своему усмотрению рассматривать некоторые фрагменты исследуемого объекта, например, нуклеотиды в молекуле нуклеиновой кислоты или символы в текстах того или иного естественного языка, записанные в алфавитной системе записи как тождественные друг другу, не отличающиеся ничем, кроме своего положения в рассматриваемом объекте – символьной последовательности.

Биологические макромолекулы играют ключевую роль во многих процессах, протекающих в живых организмах. Изучение структуры, а также функциональных, химических, физико-химических и прочих свойств нуклеиновых кислот активно ведется специалистами в различных областях, и одним из важных направлений является изучение нуклеиновых кислот как символьных последовательностей. Существующие в настоящий момент методы выделения нуклеиновых кислот не позволяют получить высокую степень извлечения. Соответственно, получаемые последовательности зачастую являются неполными, что существенно затрудняет их исследования для биологов и генетиков. Поэтому возникает задача восстановления недостающих данных в символьных последовательностях.

Основная идея диссертации. Разработан метод восстановления отсутствующих данных в символьных последовательностях, использующий только ту информацию, которая содержится в самой последовательности (частотные словари). Критерием отбора заполнений является их максимальная похожесть с имеющимися частями символьной последовательности (критерий максимума условной энтропии).

Методы исследований. Для решения поставленных задач использовались методы эволюционной оптимизации, мелкозернистый параллельный вычислитель (КМК), методы матричной алгебры.

Основные результаты.

– разработан подход к восстановлению отсутствующих данных с помощью кинетической машины Кирдина. Модифицирован и реализован имитатор КМК применительно к задаче восстановления данных.

– получен способ представления опорного частотного словаря в матричной форме. Даны определения матрицы заполнений и индикаторной матрицы. Сформулировано и доказано утверждение о том, что возведение мат-

рицы заполнений в определенную степень эквивалентно построению всех возможных заполнений из заданного опорного частотного словаря для всех возможных опор. Получен алгоритм построения заполнений, основанный на матричном представлении частотного словаря. Получен алгоритм, с помощью которого можно за приемлемое время ответить на вопрос о существовании заполнений из опорного частотного словаря, а также определить число таких заполнений.

– разработан и реализован генетический алгоритм, применительно к задаче восстановления отсутствующих данных.

Научная новизна. Результаты диссертации являются новыми, в частности, разработан метод восстановления данных, который работает с символьными последовательностями, при этом утерянная часть символьной последовательности восстанавливается с использованием только той информации, которая содержится в самой символьной последовательности (частотные словари). Оценка сверху числа заполнений дается выражением $|N|^L$ (N – мощность алфавита, L – длина отсутствующей части), что для алфавитов и характерных размеров пропусков, встречающихся в различных приложениях, представляет собой достаточно большую величину (порядка 10^{12}). Таким образом, задача построения заполнения простым перебором вариантов представляется весьма ресурсоемкой. Необходим метод, снижающий вычислительные затраты. Кроме того, построение каждого из вариантов заполнений не зависит от построения других заполнений, что сделало возможным использование подходов и методов параллельных вычислений.

Одним из вариантов такого (идеального) вычислителя рассматривалось устройство, основанное на идее молекулярных вычислений – кинетическая машина Кирдина (КМК). КМК является математическим аппаратом, обеспечивающим высокий уровень распараллеливания вычислений. Тем не менее, задача физического построения такого устройства далека от разрешения, поэтому был построен имитатор КМК на обычной последовательной машине фон Неймановского типа, который необходим для решения нашей конкретной задачи, а не всех алгоритмов, которые могут быть представимы в КМК. Также, для повышения эффективности построения заполнений, последовательный имитатор КМК был модифицирован.

Однако использование КМК не решает вопрос о существовании заполнения из опорного словаря в силу своей стохастичности. В связи с этим был предложен принципиально новый подход – особое матричное представление опорного частотного словаря.

Для всякого частотного словаря возможны три (различных) матричных представления. Первое полностью эквивалентно самому частотному словарю, второе соответствует Марковскому процессу порядка $q-1$, который

реализует гипотезу о наиболее вероятном продолжении слов этой длины, третье представление эквивалентно задаче определения числа маршрутов на графе, соответствующем матрице. Доказано утверждение о том, что возведение данной матрицы в соответствующую степень эквивалентно построению всех возможных заполнений из заданного опорного частотного словаря. Используя это утверждение, можно точно определить количество всех заполнений и построить их.

Следующий подход к построению заполнений также основан на идее сокращения ресурсоемкости вычислений и нахождении оптимального, с точки зрения выбранного нами критерия, заполнения. Сократить перебор можно только за счет выбора оптимальных и квазиоптимальных направлений на графе заполнений. Выбор таких направлений осуществлялся с помощью эволюционных алгоритмов оптимизации, а именно – генетических алгоритмов.

Значение для теории. Для восстановления отсутствующих данных разработан метод, применимый для последовательностей с различными параметрами. В качестве параметров рассматривались: длина отсутствующей части последовательности, мощность алфавита, составляющего последовательность, мощность частотного словаря, построенного по последовательности.

Решен вопрос о существовании и построении заполнения из слов заданного частотного словаря. Кроме вопроса о существовании заполнения решен вопрос о числе таких заполнений, удовлетворяющих граничным условиям.

Значение для практики. Разработанный метод может применяться во всех задачах, которые требуют восстановления пропусков в символьных последовательностях – от теории передачи данных до молекулярной биологии. С помощью этих методов можно решить задачу восстановления недостающих частей в нуклеотидных последовательностях. Заполнение таких пробелов осмысленной информацией может в значительной степени облегчить и продвинуть работу генетиков.

Также эти методы могут применяться для восстановления пробелов в текстах естественного языка при передаче сообщений, когда сообщение доходит не полностью.

Личный вклад автора. Все результаты, выносимые на защиту, получены лично автором.

Рекомендации по использованию результатов диссертации. Результаты диссертации могут быть использованы в научно-исследовательских организациях, занимающихся обработкой и анализом генетических последовательностей, в частности в Институте цитологии и генетики СО РАН, Институт биологии гена РАН, Институт молекулярной генетики РАН. Кроме

того, результаты диссертации могут быть использованы как основа для спецкурсов в профильных высших учебных заведениях.

Апробация результатов диссертации. Результаты работы были представлены на IX и X Всероссийском семинаре «Моделирование неравновесных систем», XI международной конференции «Информационные и математические технологии в научных исследованиях», XIV и XV Всероссийском семинаре «Нейроинформатика и ее приложения», пятой школе-семинаре «Распределенные и кластерные вычисления», международной конференции «Компьютерное моделирование и интеллектуальные системы», XII Байкальской Всероссийской конференции «Информационные и математические технологии в науке и управлении», VI и VII Всероссийских ФАМ конференциях.

Публикации. По теме диссертации опубликовано 15 научных работ, в том числе: 1 статья в периодических изданиях по списку ВАК, 4 статьи в научных периодических изданиях, 3 статьи в трудах международных научных конференций, 5 статей в трудах Всероссийских научных конференций.

Общая характеристика диссертации. Диссертация состоит из 4 разделов, содержит основной текст на 109 с., 11 иллюстраций, 17 таблиц, список использованных источников из 158 наименований.

Содержание работы

Введение содержит актуальность работы, ее цель, научную новизну, практическую значимость, а также структуру диссертации.

В первой главе дан обзор литературы по существующим методам восстановления утерянных данных.

Во второй главе рассматривается общая постановка задачи восстановления данных и описывается критерий качества восстановления.

В качестве данных рассматриваем конечные символьные последовательности. Полагаем, что алфавит, в котором записаны изучаемые последовательности, заранее известен и конечен. Отсутствие части такой последовательности будем рассматривать как потерю данных. При этом считаем, что длина утерянной части известна, а сама отсутствующая часть является связным диапазоном. Каков бы ни был метод построения цепочки, заполняющий лакуну (пробел), придерживаемся общего подхода, состоящего в том, что заполнения следует выполнять из «маленьких фрагментов» тех частей последовательности, которые доступны.

Рассматривается символьная последовательность, состоящая из символов алфавита Ω и L – длина участка, который необходимо восстановить (будем называть его лакуной). Полагаем, что длина лакуны L много меньше длины самой последовательности.

Определение. Словом длины q будем называть любую связную последовательность этой длины, составленную из символов алфавита Ω.

Определение. Опорным частотным словарём $W(q)$, толщины q будем называть список всех слов этой длины, встречающихся в доступных частях символьной последовательности, с указанием частот этих слов f_ω.

Определение. Под частотой f_ω слова ω будем понимать отношение

$$f_\omega = \frac{N_\omega}{N},$$

где N_ω – число копий слова ω в исходной последовательности, N – общее число слов длины q в исходной последовательности, с учетом кратности всех слов.

Определение. Словарь $W(q)$ называется полным, если он содержит слова, состоящие из всех возможных сочетаний символов алфавита Ω. В противном случае словарь неполный.

Определение. Пополненным частотным словарём $\overline{W}(q)$ будем называть частотный словарь (толщины q), составленный по той последовательности, которая возникает в результате заполнения пробела.

Определение. Левой опорой длины t, ($0 \le t \le q-1$) будем называть слово этой длины, расположенное сразу слева от лакуны.

Определение. Правой опорой длины t, ($0 \le t \le q-1$) будем называть слово этой длины, расположенное сразу справа от лакуны.

Тем самым, восстанавливаемая часть имеет длину $L+2t$, при условии, что первые и последние t символов являются фиксированными.

Пусть имеется некоторая символьная последовательность Ψ конечной длины N. Построим частотный словарь соответствующий данной последовательности. Зафиксируем длину слова q. Теоретически длина слова может быть любой, не превышающей длины самой последовательности. В частности, частотный словарь может состоять из одного слова, соответствующего самой последовательности. Длина слова q может выбираться исходя из специфики самой последовательности.

Возьмем первые q символов последовательности Ψ, эта подпоследовательность будет представлять первое слово в частотном словаре. Далее произведем сдвиг рамки считывания длиной q символов на один символ вправо и выпишем следующие q символов последовательности Ψ, получим следующее слово и так далее. Процедура повторяется до тех пор, пока не дойдем до конца последовательности. При выписывании таких

подпоследовательностей длины q учитывается кратность каждого слова и запоминается число его копий. Далее для каждого слова определяется частота, с которой оно появляется в исходной последовательности.

Пример. Пусть Ω = {a,c,g}, а Ψ = accaaacaag. Фиксируем $q = 3$. Выписываем из Ψ все подряд цепочки символов длины 3, получим список слов:

acc; cca; caa; aaa; aac; aca; caa; aag.

Учтем число копий:

C(acc) = 1;
C(cca) = 1;
C(caa) = 2;
C(aaa) = 1;
C(aac) = 1;
C(aca) = 1;
C(aag) = 1;

Подсчитаем частоты:

$N = 8$, f(acc) = 0.125;
f(cca) = 0.125;
f(caa) = 0.25;
f(aaa) = 0.125;
f(aac) = 0.125;
f(aca) = 0.125;
f(aag) = 0.125;

В результате получаем опорный частотный словарь толщины 3 для последовательности Ψ = accaaacaag.

Лакуну будем заполнять исходя из той информации, которая имеется в распоряжении. Единственная информация, которая доступна – это знание числа копий отдельных «маленьких фрагментов» имеющихся частей последовательности. Данную информацию содержит опорный частотный словарь.

Так как никакой другой информации не используется, то построение заполнения лакуны означает построение из слов длины q цепочки вида:

$$\omega_1, \omega_2, \omega_3, \ldots, \omega_{L+2t-q}, \omega_{L+2t-q+1} \qquad (1)$$

длиной $L + 2t$, у которой первые t и последние t символов заданы, а для каждой пары соседних слов выполняется условие:

$$\omega_j = i_l \overline{\omega};\ \overline{\omega} i_q = \omega_{j+1}; \qquad (2)$$

т.е. два соседних слова пересекаются по общему подслову длины $q-1$, первое слово в этой цепочке начинается левой опорой α_l, а последнее

заканчивается правой опорой α_r.

Если такая цепочка существует и она единственна, то задача построения заполнения решена. В общем случае число таких заполнений будет больше одного. Число всех возможных заполнений для полного словаря $W(q)$ составляет $|\Omega|^L$. Это число является верхней границей количества возможных заполнений. Для неполного словаря число возможных заполнений меньше. Если процедура восстановления отсутствующих данных порождает не одно заполнение, а несколько, возникает задача выбора лучшего из них.

В нестрогом изложении принцип заполнения заключается в следующем: заполнять лакуны надо таким образом, чтобы последовательность, получающаяся после заполнения (восстановленная), была наиболее похожа на те части последовательности, которые имеются в наличии, однако это требуется сделать так, чтобы восстановленная последовательность несла в себе минимум дополнительной информации.

Таким образом, будем сравнивать восстановленную последовательность с имеющимися в распоряжении исследователя частями последовательности. Основная трудность здесь заключается в том, что в пространстве символьных последовательностей сложно ввести метрику. Формально метрика в таком пространстве существует – это метрика Хемминга. Однако такая метрика малопродуктивна. Она позволяет различать лишь полностью совпадающие последовательности и все остальные. Более распространенным в настоящее время методом сравнения символьных последовательностей является метод выравнивания. Метод заключается в подгонке одной последовательности под другую с помощью разрешенных операций: вставка пробела, замена либо удаление символов так, чтобы эти две последовательности совпали. Каждой такой операции назначается определенный штраф. Ближайшей последовательностью считается та, для которой подобные преобразования дают наименьшее значение суммарного штрафа. Однако у этого метода есть два принципиальных недостатка. Метод выравнивания требует выбора системы штрафных функций и выбора опорной последовательности, относительно которой проводится выравнивание. И то, и другое выбирается исходя из соображений, лежащих за пределами собственно метода выравнивания. К тому же метод выравнивания чувствителен к длине сравниваемых последовательностей. Точность сравнения падает экспоненциально с ростом их длины (если только они не имеют совпадающих участков, сопоставимых по длине со всей последовательностью).

Фактически невозможно выровнять две последовательности, сильно различающиеся по длине. Выравнивание также имеет свои ограничения и по числу сравниваемых последовательностей. Несмотря на то, что увеличение числа сравниваемых последовательностей ведет, как правило, к росту точности выравнивания, общее число выравниваемых последовательностей едва ли

может превышать 10^2 ; содержательно выровнять тысячу (и более) последовательностей едва ли возможно.

Кроме того, рассмотренные выше критерии не учитывают количественную сторону внесенной дополнительной информации. Учесть это можно, работая с частотными словарями. Подобие (или близость) двух либо нескольких последовательностей можно определить путём вычисления энтропии их частотных словарей. Вычисление энтропии позволяет оценить количество информации, необходимое для преобразования одной последовательности в другую. Соответственно, будем выбирать такие заполнения, для которых рост энтропии является наименьшим среди всех возможных. Строго эти соображения формулируются в виде двух экстремальных принципов: принцип максимума энтропии восстановленного частотного словаря и принцип максимума условной энтропии опорного частотного словаря относительно пополненного. Опишем их подробнее.

Напомним, что если существует цепочка вида (1), составленная из слов опорного словаря и она единственна, то задача построения заполнения решена. Если же существует несколько цепочек вида (1), составленных из слов опорного словаря, то среди всех возможных следует выбрать ту цепочку, которая обеспечивает максимум энтропии

$$\tilde{S} = -\sum_{\omega} \left\{ \tilde{f}_{\omega} \cdot \ln \tilde{f}_{\omega} \right\} \qquad (3)$$

пополненного частотного словаря $\overline{W}(q)$, где $\tilde{f}_{\omega}$ – частота слов, вычисленная по тексту, полученному в результате заполнения лакуны. Подчеркнём, что задача построения заполнения словами из опорного частотного словаря может не иметь решения. Однако вне зависимости от того, существует или нет заполнение лакуны словами из опорного словаря, можно строить заполнение лакуны всеми возможными в данном алфавите словами. Очевидно, что такое заполнение существует всегда и оно не единственно.

Возможен иной подход к выбору наилучшего заполнения (1). Он реализует принцип максимального подобия построенного заполнения имеющимся частям последовательности. Здесь следует выбрать такое заполнение (1), для которого условная энтропия

$$\tilde{S} = -\sum_{\omega} \left\{ f_{\omega} \cdot \ln \frac{f_{\omega}}{\tilde{f}_{\omega}} \right\} \qquad (4)$$

опорного частотного словаря $W(q)$ относительно пополненного $\overline{W}(q)$ достигнет максимума. Здесь f_{ω} – частота слов в опорном частотном словаре, а

$\tilde{f}_\omega$ – частота слов в словаре, построенном по всей последовательности, полученной в результате заполнения лакуны (пополненного); понятно, что $f_\omega = 0$ для некоторых ω', в то время, как $\tilde{f}_{\omega'} > 0$.

Эти два принципа выбора наилучшего заполнения не являются взаимоисключающими либо конкурирующими. Каждый из них может быть применён независимо в одной и той же ситуации, каждый из них позволит выбрать то или иное заполнение. Более того, каждый из них обеспечивает наилучшее заполнение, понимаемое в разных смыслах.

Используя максимум энтропии пополненного частотного словаря, получим частотный словарь, который будет максимально похож (частоты будут близки к равновесным) на равновесный словарь. Это означает, что само по себе заполнение выбрано таким, чтобы получившаяся в результате последовательность была менее всего «нагружена» дополнительной информацией. Такого рода дополнительная информация может быть привнесена исследователем непроизвольно, если он будет выбирать слова для построения заполнения.

Максимум условной энтропии выберет такую последовательность, которая даст частотный словарь, наименее отличающийся (частоты слов изменятся незначительно) от опорного словаря. Тем самым, максимум условной энтропии реализует принцип максимального подобия и минимума дополнительной информации.

Следует особо подчеркнуть, что предлагаемый метод восстановления утерянных данных и критерии выбора наилучшего восстановления не могут быть применимы для построения «настоящей» исходной последовательности. Это связано с тем, что, во-первых, сам по себе метод частотных словарей никогда «не знает» исходной последовательности и в лучшем случае можно вести речь о сравнении словарей. Заметим, что совпадение частотных словарей двух последовательностей отнюдь не гарантирует совпадения самих последовательностей. Соответственно, восстановление понимается как некоторое моделирование данных, с минимальным изменением этих модельных данных по сравнению с имеющимися образцами (минимальное изменение частот).

В третьей главе рассматриваются подходы к восстановлению данных с помощью имитации кинетики химических реакций – кинетической машины Кирдина **(**КМК**)**, матричного представления частотного словаря и эволюционных алгоритмов оптимизации.

Неформально КМК можно представить себе как аналог химического реактора, в котором происходят реакции. Имеется химический реактор идеального смешения, в котором плавают слова. В реактор добавляются правила – катализаторы; одни из них, взаимодействуя со словами, способствуют их

распаду, другие, встречая пару подходящих слов, способствуют их синтезу, и, наконец, третьи заменяют в словах некоторые подцепочки.

КМК — абстрактная модель параллельных вычислений, предложенная А.Н. Кириным в октябре 1997 года на конференции «Нейроинформатика и ее приложения».

Идея КМК состоит в следующем. Пусть Ω – алфавит символов. Ω^* – множество всех конечных слов или цепочек в алфавите Ω. Обрабатываемой единицей является ансамбль слов M из алфавита Ω, который отождествляется с функцией F_M с конечным носителем на Ω^*, принимающей неотрицательные целые значения $F_M : \Omega^* \to N \cup \{0\}$. Значение $F_M(\omega)$ интерпретируется как число экземпляров слова ω в ансамбле M.

Обработка ансамблей в КМК состоит в совокупности элементарных событий, происходящих недетерминировано и параллельно. Элементарное событие $S: M \to M'$ состоит в том, что из ансамбля M изымается ансамбль K^- и добавляется ансамбль K^+, т.е. $F_{M'} = F_M - F_{K^-} + F_{K^+}$. Ансамбли K^- и K^+ однозначно задаются правилами или командами, которые объединяются в программу. Команды могут быть только трёх видов:

- Распад $uvw \to uf + gw$;
- Синтез $uk + qw \to usw$;
- Прямая замена $uvw \to usw$.

Применительно к задаче восстановления отсутствующих данных команды КМК выглядели следующим образом. Для построения частотного словаря W_q из текста T использовалась команда распада, работающая по следующей формуле:

$$uf^1v^{q-1}g^1w \to uf^1v^{q-1} + v^{q-1}g^1w, \qquad (5)$$

где, в качестве M берется ансамбль, состоящий из одного слова T. Верхним индексом обозначено количество символов в слове. После того, как машина остановится, ансамбль M будет содержать все слова длины q, встречающиеся в исходном тексте, с учетом их кратности.

Программа, реализующая процесс заполнения лакуны в терминах КМК выглядит следующим образом:

$$\begin{aligned} &\alpha_l + \alpha_l v^{q-t} \to \alpha_l v^{q-t} *, \qquad \text{(а)} \\ &v^{q-t}\alpha_r + \alpha_r \to *v^{q-t}\alpha_r \end{aligned} \qquad (6)$$

$$uv^{q-1} * + v^{q-1}v^1 \longrightarrow uv^{q-1}v^1 * \qquad \text{(б)}$$
$$v^1v^{q-1} + * v^{q-1}w \longrightarrow * v^1v^{q-1}w$$

$$u * + * v \longrightarrow uv \qquad (7)$$

Первые две строчки программы осуществляют инициализацию заполнений, т.е. обеспечивают взаимодействие правой (левой, соответственно) опоры длины t с подходящим словом длины q. Третья и четвертая строчки осуществляют рост заполнений. И, наконец, последняя строка склеивает левые и правые части. Символ $<*>$ не принадлежит алфавиту Ω и используется в программе, чтобы помечать те слова, которые успешно прошли стадию инициации.

Был реализован последовательный имитатор КМК. Для повышения эффективности построения заполнений лакун в символьной последовательности последовательный имитатор КМК был модифицирован; всего было внесено три модификации:

– все заполнения росли только в одном направлении – слева направо, для определённости;

– модификации подвергся словарь, по которому строилось заполнение;

– периодически проводилась селекция всех слов, являющихся продолжениями опор.

Таким образом, алгоритм работы имитатора КМК можно представить следующим образом:

- выбор параметров имитатора – толщины словаря, количества обрабатываемых слов, количества контрольных точек (в которых проводится селекция слов);
- инициализация левых опор. Формирование опорного частотного словаря;
- формирование модифицированного словаря;
- обработка слов;
- выбор заполнений, которые удовлетворяют граничным условиям;
- выбор заполнения в соответствии с определённым критерием.

Следует отметить, что существует много различных способов организации вычислений, в том числе, используя языки программирования, машину Тьюринга и т.д. Однако КМК является высоко параллельным вычислительным устройством, физическая реализация которого отсутствует и в ближайшее время вряд ли появится. Поэтому в данной работе была предпринята попытка имитации работы специального вычислителя, такого как КМК применительно к задаче восстановления данных.

Далее рассматривается подход к восстановлению данных с помощью матричного представления частотного словаря. Всякий частотный словарь представляет собой (упорядоченный) список слов длины q. Такой список можно однозначно преобразовать в матрицу A порядка $N^{q-1} \times N^{q-1}$ (N – мощность алфавита), в которой строки и столбцы помечены словами ω' и ω'' длины q каждое. Тогда любое слово ω из словаря $W(q)$, начинающееся последними $q-1$ символами слова ω' и заканчивающееся первыми $q-1$ символами слова ω'', соответствует элементу матрицы, находящемуся на пересечении данных строки и столбца; а сам этот элемент является частотой слова ω.

Заменив в матрице A элементы в каждой строке таким образом, чтобы их сумма по строке стала равной единице, а соотношение между элементами сохранилось, получаем матричное представление $\overline{A}$ модифицированного частотного словаря. Наконец, введём ещё одно представление частотного словаря. Заменив в матрице все ненулевые элементы на единицу, а нулевые оставив неизменными, получаем матрицу I, которую будем называть индикаторной матрицей.

Таким образом, для всякого частотного словаря возможны три (различных) матричных представления. Первое полностью эквивалентно самому частотному словарю, второе соответствует Марковскому процессу порядка $q-1$, который реализует гипотезу о наиболее вероятном продолжении слов этой длины, третье представление эквивалентно задаче определения числа маршрутов на графе, соответствующем матрице.

Пусть имеется некоторый текст, состоящий из символов алфавита $\Omega = \{0,1\}$. Пусть этому тексту соответствует опорный частотный словарь $W(q) = \{00,01,10,11\}$ и, соответственно, частоты слов равны f_1, f_2, f_3, f_4. Тогда матричное представление модифицированного словаря будет следующим:

$$A = \begin{pmatrix} & 00 & 01 & 10 & 11 \\ 00 & f_1 / (f_1 + f_2) & f_2 / (f_1 + f_2) & - & - \\ 01 & - & - & f_3 / (f_3 + f_4) & f_4 / (f_3 + f_4) \\ 10 & f_1 / (f_1 + f_2) & f_2 / (f_1 + f_2) & - & - \\ 11 & - & - & f_3 / (f_3 + f_4) & f_4 / (f_3 + f_4) \end{pmatrix}$$

а индикаторная матрица будет иметь такой вид:

$$I = \begin{pmatrix} & 00 & 01 & 10 & 11 \\ 00 & 1 & 1 & 0 & 0 \\ 01 & 0 & 0 & 1 & 1 \\ 10 & 1 & 1 & 0 & 0 \\ 11 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Процесс построения заполнений можно представить в виде роста дерева. В корне этого дерева находится слово, совпадающее с левой опорой. Каждый узел – это слово, которое получается присоединением одного символа к слову из родительского узла. Из каждого узла выходит столько ветвей, сколько можно присоединить символов к текущему слову. Глубина такого дерева будет составлять $L+t+1$, где L – длина лакуны, t – длина правой опоры. Так как при построении заполнений часто возникают ситуации, когда на некотором шаге к слову нельзя присоединить ни один символ, то, соответственно, дерево может быть неполным.

Определение. Дерево, включающее в себя как ветки, которые могут дорасти до глубины $L+t+1$, так и ветки, которые обрываются на некотором уровне, будем называть *полным* деревом.

Определение. Дерево, включающее в себя только те ветки, которые достигают глубины $L+t+1$, будем называть *неполным* деревом.

Составим таблицу, соответствующую частотному словарю длины q. В первом столбце и первой строке таблицы расположим все слова из частотного словаря. Если последние $q-1$ символов i – ой строки совпадают с первыми $q-1$ символами j – го столбца, то в ячейке, стоящей на пересечении этой строки и этого столбца пишем последний символ слова, расположенного в j – ом столбце. Если совпадения нет, то соответствующая ячейка будет пустой.

Таблица 1. Матрица заполнений.

	$\overline{\omega_1}\nu_1$	$\overline{\omega_1}\nu_2$	$\overline{\omega_1}\nu_3$	…	$\overline{\varphi_1}\nu_1$	$\overline{\varphi_1}\nu_2$	$\overline{\varphi_1}\nu_3$	…
$\nu_1\overline{\omega_1}$	ν_1	ν_2	ν_3	…	0	0	0	…
$\nu_1\overline{\omega_2}$	0	0	0	…	0	0	0	…
$\nu_1\overline{\omega_3}$	0	0	0	…	0	0	0	…
…	…	…	…	…	…	…	…	…
$\nu_1\overline{\varphi_1}$	0	0	0	…	ν_1	ν_2	ν_3	…
$\nu_1\overline{\varphi_2}$	0	0	0	…	0	0	0	…
$\nu_1\overline{\varphi_3}$	0	0	0	…	0	0	0	…
…	…	…	…	…	…	…	…	…

Полученную матрицу будем называть матрицей заполнений. Элементами матрицы являются последовательности символов. Определим операции сложения и умножения для элементов матрицы.

Определение. Под сложением двух элементов матрицы будем понимать конкатенацию строк при помощи служебного символа, не входящего в алфавит Ω. Суммой двух строк «00000» и «11111» будет строка следующего вида «00000+11111», где '+' – служебный символ.

Определение. Под произведением двух элементов матрицы будем понимать элемент, полученный по следующему правилу: к каждой строке первого множителя, находящейся между служебными символами, приписывается каждая строка второго множителя, находящаяся между служебными символами.

Например, первый множитель – строка «001+000», второй – строка «011+100». Тогда произведением этих строк является строка «001011+000011+001100+000100».

Определение. Под нулевым элементом матрицы будем понимать элемент, содержащий пустую строку. Прибавление пустого элемента исходного не меняет. Умножение на пустой элемент дает пустой элемент.

Утверждение. Возведение матрицы заполнений в степень $L+t$ эквивалентно построению всех возможных заполнений из заданного опорного частотного словаря для всех возможных опор. Здесь L – длина лакуны, t – длина правой опоры.

Следствие. Если вместо матрицы заполнений использовать индикаторную матрицу I, то у матрицы I^{L+t} в столбце, в котором левые $q-1$ символов слова совпадают с правой опорой и строке, в которой правые $q-1$ символов слова совпадают с левой опорой, будет стоять число всех возможных заполнений для выбранных опор, которые только можно построить из заданного опорного частотного словаря. Существование такого числа отличного от нуля, отвечает на вопрос о существовании заполнения из опорного частотного словаря как такового и говорит о числе таких заполнений.

Далее рассматривается подход к восстановлению данных с помощью генетических алгоритмов – ГА. Преимущества данного подхода перед матричным представлением и КМК заключается в том, что генетические алгоритмы существенно сужают пространство поиска и с помощью генетических операторов рекомбинации, мутации и селекции выбирают перспективные области за короткое время. Здесь речь идет именно об алгоритмах, так как, меняя хотя бы один генетический оператор, мы получаем совершенно другой алгоритм.

Генетический алгоритм (ГА) представляет собой метод оптимизации, основанный на концепциях естественного отбора и генетики. В этом подходе переменные, характеризующие решение, представлены в виде ген в хромосоме. ГА оперирует конечным множеством решений (популяцией) – генерирует

новые решения как различные комбинации частей решений популяции, используя, такие операторы, как селекция, рекомбинация и мутация. Новые решения располагаются в популяции в соответствии со стратегией замещения. При использовании ГА задача нахождения оптимального заполнения сводится к задаче нахождения оптимального пути на графе заполнений. При этом максимизируется значение функции условной энтропии, которая является функцией многих переменных.

Представляет интерес найти те значения параметров, при которых достигается наилучшее точное значение функции (которое возможно по данному словарю).

Разработан и реализован ГА применительно к задаче восстановления отсутствующих данных. Общий алгоритм, применительно к задаче восстановления данных будет следующим:

а) Инициализация начальной популяции. Генерирование символьных последовательностей длиной, равной длине лакуны L. Последовательности заполняются случайным образом символами алфавита, из которого состоит исследуемый текст;

б) Оценка начальной популяции. Считается энтропия для всех полученных текстов. Лучший индивид запоминается;

в) Селекция. Одним из методов селекции отбираются строки, которые станут родителями;

г) Скрещивание. Одним из методов скрещивания получаются два потомка. Затем они оцениваются, и выбирается лучший индивид;

д) Мутация. Все индивиды с заданной вероятностью подвергаются мутации. Затем, мутировавшие индивиды переоцениваются;

е) Формирование нового поколения. Лучший индивид запоминается;

ж) Если не выполняется критерий останова, то повторять шаги в), г), д), е).

В четвертой главе приводятся результаты вычислительных экспериментов. Для апробации представленных алгоритмов были реализованы последовательный имитатор КМК, матричный способ построения заполнений и генетический алгоритм. В качестве среды разработки использовались C++ Builder 6.0 и Microsoft Visual Studio 2005.

В качестве тест-объекта был взят текст генетической последовательности с кодом AB012132 – геном вируса парагриппа человека. Длина текста составляла 14573 символа. Искусственно создавались лакуны длиной 6, 10, 20 символов. Затем с помощью вышеописанных алгоритмов строились заполнения данных лакун.

Сначала оценивалось количество возможных вариантов заполнения лакун, удовлетворяющих граничным условиям. Для выбранного текста была построена индикаторная матрица для различных толщин словаря. Для каждой лакуны и толщины словаря было определено количество заполнений, которые возможно построить по заданному опорному частотному словарю. Коли-

чество возможных заполнений для лакун длиной 10 и 20 символов представляет достаточно большую величину порядка 10^{12}. Построить и обработать и такое количество заполнений очень сложно. Поэтому с помощью матричного представления частотного словаря были построены заполнения только для лакуны длиной 6 символов.

При использовании имитатора КМК строился опорный частотный словарь $W(q)$ для различных значений q и по нему строились заполнения. Рассматривались заполнения с толщиной словаря от 3 до 8 символов. Число опор для каждой толщины словаря бралось равным 100000. КМК реализует гипотезу о наиболее вероятном продолжении заполнения. Именно такие заполнения доставляют меньшее значение энтропии.

При заполнении лакун с помощью генетических алгоритмов (ГА) также строился опорный частотный словарь $W(q)$ и по нему строились заполнения. Основная сложность при использовании генетических алгоритмов – это оптимально подобрать параметры алгоритма. Построение заполнений производилось при различных значениях параметров ГА. В качестве общих параметров использовались следующие: размер популяции – 200 индивидов, число поколений 200, использование штрафной функции – да, размер турнира – 3 (для турнирной селекции). Так как алгоритм стохастический, то производилось не менее 5 запусков, затем выбирался лучший результат.

Сравнение результатов, полученных разными алгоритмами представлены в таблице 2.

Таблица 2. Сравнение результатов экспериментов при длине лакуны 6 символов и различных толщинах словаря

q	Матричное представление	КМК	ГА
3	aaatat -5,575631539025E-7	aaatat -5,575631539025E-7	aaatat -5,575631539025E-7
4	aaatat -2,849668656467E-6	aaagat -2,874278297468E-6	aaatat -2,849668656467E-6
5	aaaaat -1,003320483111E-5	aaaaat -1,003320483111E-5	aaaaat -1,003320483111E-5
6	aaaaca -4,180962370172E-5	agatat -4,24869269827E-5	aaaaca -4,180962370172E-5

Исходная часть и значение условной энтропии представлены в таблице 3.

Таблица 3.Утерянная часть длиной 6 символов и ее значения энтропии при различных толщинах словаря

q	Исходная часть	Энтропия
3	agtgca	-2,158534029496E-6
4	agtgca	-9,749696922197E-6
5	agtgca	-3,332945613707E-5
6	agtgca	-1,486498956082 E-4

Таблица 4. Сравнение результатов экспериментов при длине лакуны 10 и 20 символов и толщине словаря 3

	КМК	ГА
10	agataattat -6,304889570291E-7	aatatcagaa, gaaatatcaa, agaaatatca. -6,280759463479E-07
20	acatcagataatatgttgaa -1,031549682494E-6	tatttaacaatgaaaagatc -7,892298411952E-07

При этом для лакуны 10 символов исходная часть была aagcagtgca с энтропией -2,385809980663E-6. Для лакуны 20 символов исходная часть была aagcagtgcagtcagttcag с энтропией -5,148549560011E-6.

Матричное представление частотного словаря позволяет построить оптимальное заполнение. Как видно из результатов, таблица 2, КМК и ГА работают эффективно и построили оптимальные и близкие к оптимальным заполнения. При толщине словаря 4, КМК сработал хуже, чем ГА. Если же обратиться к таблице 4, то видно, что ГА строил заполнения эффективней, чем КМК. При этом ГА затратил на построение гораздо меньше времени – 5 мин. против 50 мин. у КМК.

Все полученные алгоритмы являются эффективными способами построения заполнений. Матричное представление используется для определения существования заполнения из опорного частотного словаря и построения заполнений для небольших лакун (до 10 символов) и словарей (до 65 слов). ГА и КМК также являются эффективными процедурами построения заполнений. КМК показал себя несколько хуже, чем ГА с точки зрения качества построения заполнений (значение энтропии) и временных затрат. Однако КМК может работать там, где не работает ГА. Для ГА затраты на вычисление целевой функции являются существенными. И там, где длина лакуны превосходит 20 символов, а мощность словаря превосходит 3000 слов, ГА не работает. Для КМК – это не критично и он может работать на лакунах длиной более 300 символов.

Настоящие алгоритмы разрабатывались для задачи восстановления отсутствующих данных в генетических последовательностях. Однако в этих алгоритмах есть одна важная особенность – они не привязаны к конкретному тексту, поскольку используют только числовые характеристики самого текста. Поэтому данные алгоритмы могут применяться не только для восстановления генетических последовательностей, но и любых других текстов тоже.

В качестве альтернативной последовательности был взят текст Всемирной декларации прав человека. Из оригинального текста декларации были удалены пробелы и знаки препинания. Для удобства рассматривали связную последовательность, что не повлияло на общность результатов. Получившаяся символьная последовательность составила около 7500 символов, длина лакуны – 100 символов. Восстановление производилось с помощью КМК. Использовались словари толщиной от 3 до 13 символов. Число левых опор для каждой толщины словаря бралось равным 100000. Число заполнений, совпавших с правой опорой составляло для всех словарей толщины $q \le 11$ около 0,0001 %. Для словарей толщины $q > 11$, заполнений, совпавших с правой опорой, получено не было. В таблице 5 представлены значения условной энтропии для полученных заполнений при различных толщинах словаря.

Таблица 5. Значения условной энтропии для наилучшего заполнения.

	Мощность алфавита
q	32
3	-0.002501334
4	-0.004114464
5	-0.005274570
6	-0.006469274
7	-0.007102911
8	-0.006994414
9	-0.007878056
10	-0.00755565
11	-0.00835729

Ниже приведены варианты заполнений лакуны для словарей толщины $q = 4$, $q = 6$, $q = 8$ и собственно текст, который был удален из Декларации для получения лакуны; правая, а также левая опоры не показаны. После удаления знаков препинания и пробелов исходный (удаленный) текст выглядел следующим образом:

егоимуществакаждыйчеловекимеетправонасвободумыслисовестиир елигииэтоправовключаетсвободуменятьсвоюрел.

При восстановлении удаленного текста по словарям $W(4)$, $W(6)$, $W(8)$, соответственно, были получены следующие цепочки:

– имладатьсяниисовявляющеедолжностиправонасоциальности
идругиминациюкаждыйчеловекимеетправонасторойирел

– егограждансствоиработающийимеетправонаравнуюзащитой
каждыйчеловекимеетправоприниматьучаствоватьсвоюрел

– егограждансстваилитакихпосягательствкаждыйчеловекимеетправо
наобеспечивающихсвободуисповедоватьсвоюрел

Построение заполнений при толщине словаря шесть и более символов позволяет получить заполнения, в которых есть понятные фразы.

$W(8)$:

его гражданства или таких посягательств каждый человек имеет право
на обеспечивающих свободу исповедовать свою рел

И интуитивно становится понятно, что каждый человек имеет право исповедовать любую религию. В результате, хоть мы и не восстановили в точности текст, но смогли распознать его смысловую часть

его имущества каждый человек имеет право на свободу мысли совести религии это право включает свободу менять свою рел

Понятно, что никакой разумный алгоритм не гарантирует нам точного восстановления. Например, если в пробеле присутствовали символы, которые не встречаются в исследуемых частях текста, то этот фрагмент безвозвратно утерян. Однако возможно построить некоторую модель утерянных данных, учитывая числовые характеристики самого текста. Более того, если части утерянного фрагмента встречаются в доступных исследователю частях, а условная энтропия опорного частотного словаря относительно пополненного (для данного фрагмента) минимальна, то возможно точно восстановить утерянный фрагмент. Результаты показали, что с помощью разработанных методов можно решать задачи по восстановлению данных. Эффективность каждого из методов зависит от длины лакуны и характера самих данных.

Заключение

В работе решается задача восстановления утерянных (или пропущенных) данных в символьных последовательностях, в частности, генетических последовательностях. Для восстановления таких данных используется только та информация, которая содержится в исходном тексте и доступна исследователю – частотные словари. В качестве критерия качества восстановления используется условная энтропия (мера схожести).

В работе предложено три алгоритма восстановления утерянных данных в символьных последовательностях.

Первый – с помощью кинетической машины Кирдина. Построен имитатор КМК применительно к задаче восстановления данных. Для повышения эффективности построения заполнений в символьной последовательности имитатор КМК был модифицирован. Всего было внесено три модификации:

– все заполнения росли только в одном направлении — слева направо, для определённости;

– модификации подвергся словарь, по которому строились заполнения;

– периодически проводилась селекция всех слов, являющихся продолжениями опор;

Второй подход заключается в представлении опорного частотного словаря в виде матрицы A. Рассмотрено специальное матричное представление частотного словаря. Даны определения матрицы заполнений и индикаторной матрицы. Сформулировано и доказано утверждение о том, что возведение матрицы заполнений в степень $L+t$ эквивалентно построению всех возможных заполнений из заданного опорного частотного словаря для всех возможных опор, где L – длина лакуны, t – длина правой опоры. Получен алгоритм построения заполнений, основанный на матричном представлении частотного словаря. Получен алгоритм, с помощью которого можно за приемлемое время ответить на вопрос о существовании заполнений из опорного частотного словаря, а также определить число таких заполнений.

Третий подход – использование генетических алгоритмов. Представлена структура ГА применительно к задаче восстановления утерянных данных.

Проведены вычислительные эксперименты по заполнению лакун в символьных последовательностях. Полученные результаты показали, что каждый из предложенных алгоритмов может применяться для восстановления данных. Качество восстановления проверялось при различных значениях параметров.

Каждый алгоритм имеет свою область применения. Так имитатор КМК «хорошо» заполняет пробелы в текстах разной сложности, алфавитов различной мощности, и применим для больших лакун. Однако он не гарантирует построения оптимального заполнения. В то время как матричное представление позволяет построить все заполнения по данному словарю, но не работает на больших алфавитах и лакунах. Матричное представление лучше всего использовать для небольших (до 10 символов) лакун, когда мощность словаря не превышает 2000 слов.

ГА хорошо зарекомендовали себя при мощности словаря до 64 слов и толщине словаря 3 символа. При этом длина лакуны может достигать 50 символов.

Полученные методы зарекомендовали себя как эффективные способы восстановления утерянных или пропущенных данных в символьных последовательностях.

Публикации автора по теме диссертации

1. Рубцов А.Г., Сенашова М.Ю. Матричное представление частотного словаря для восстановления отсутствующих данных // Журнал Сибирского федерального университета. – Красноярск, серия «Математика и физика», январь 2009, том 2. №1. с. 105-115.

2. Рубцов А.Г. Восстановление отсутствующих данных и принцип максимального подобия / Рубцов А.Г., Садовский М.Г., Сенашова М.Ю. // Вычислительные технологии / Издательство СО РАН. - Новосибирск. 2008. Т. 13., 3, С. 114- 127.

3. Рубцов А.Г. Кинетическая машина Кирдина и задача восстановления утерянных данных / Сенашова М.Ю., Рубцов А.Г., Садовский М.Г. // "Радіоелектроніка, Інформатика, Управління". – 2007. – № 1. – с. 87-93.

4. Рубцов А.Г. Кинетическая машина Кирдина в проблеме восстановления отсутствующих фрагментов символьных последовательностей. / Сенашова М.Ю., Садовский М.Г., Рубцов А.Г. // Ползуновский альманах. – Барнаул, 2006. №4. С. 59-63.

5. Рубцов А.Г. Применение генетических алгоритмов для восстановления отсутствующих данных в символьных последовательностях. / Сенашова М.Ю., Рубцов А.Г. // Ползуновский альманах. – Барнаул, 2007. №3. С. 84-87.

6. Рубцов А.Г. Восстановление отсутствующих данных в символьных последовательностях. / Рубцов А.Г., Садовский М.Г., Сенашова М.Ю. // Сборник научных трудов международной конференции «Компьютерное моделирование и интеллектуальные системы»:. – Запорожье: ЗНТУ, 2007. –206-212.

7. Рубцов А.Г.Применение кинетической машины Кирдина для восстановления утерянных данных в символьных последовательностях / Сенашова М.Ю., Рубцов А.Г., Садовский М.Г. // Информационные и математические технологии в научных исследованиях / Труды XI международной конференции «Информационные и математические технологии в научных исследованиях». Часть II. - Иркутск: ИСЭМ СО РАН, 2006. - С.168-176.

8. Рубцов А.Г. Восстановление отсутствующих данных в символьных последовательностях. Генетические алгоритмы / Сенашова М.Ю., Рубцов А.Г. // Материалы VIII международной научно-методической конферен-

ции “Информатика: проблемы, методология, технологии” (7-8 февраля 2008 г.). – Воронеж: Воронежский государственный университет, 2008. с. 232-237

9. Рубцов А.Г. Восстановление отсутствующих данных в символьных последовательностях, оценка количества заполнений /Рубцов А.Г., Садовский М.Г., Сенашова М.Ю. // Материалы IX Всероссийского семинара "Моделирование неравновесных систем-2006", 13-15 октября 2006 г. / Под ред. В.В. Слабко. Отв. за выпуск М.Ю. Сенашова, ИВМ СО РАН, Красноярск, 2006, с.145-148.

10. Рубцов А.Г. Принцип максимального подобия в проблеме восстановления утерянных данных. / Рубцов А.Г., Сенашова М.Ю., Садовский М.Г. // Нейроинформатика и ее приложения: Материалы XIV Всероссийского семинара, 6-8 октября 2006 г. / Под ред. А.Н. Горбаня, Е.М. Миркеса. Отв. За выпуск Г.М. Садовская, ИВМ СО РАН, Красноярск, 2006, с.88-90.

11. Рубцов А.Г. Оценка количества заполнений при восстановлении отсутствующих данных / Рубцов А.Г., Садовский М.Г., Сенашова М.Ю. // Распределенные и кластерные вычисления. Избранные материалы Пятой школы-семинара. - Красноярск: Институт вычислительного моделирования СО РАН. - 2007. - С. 132-149

12. Рубцов А.Г. Восстановление отсутствующих данных. Оценка вычислительных затрат / Сенашова М.Ю., Рубцов А.Г., Садовский М.Г. // Информационные и математические технологии в науке и управлении / Труды XII Байкальской Всероссийской конференции «Информационные и математические технологии в науке и управлении». Часть II. – Иркутск: ИСЭМ СО РАН, 2007. – С. 99-106.

13. Рубцов А.Г. Генетические алгоритмы в задаче восстановления отсутствующих данных / Рубцов А.Г., Сенашова М.Ю., Садовский М.Г. // Нейроинформатика и ее приложения: Материалы XV Всероссийского семинара, 5-7 октября 2007 г. / Под ред. А.Н. Горбаня, Е.М. Миркеса. Отв. за выпуск Г.М. Садовская, ИВМ СО РАН, Красноярск, 2007, с.108-114.

14. Рубцов А.Г. Восстановление отсутствующих данных в символьных последовательностях при помощи матричного представления частотного словаря / Сенашова М.Ю., Садовский М.Г. и Рубцов А.Г.// Труды Шестой Всероссийской ФАМ'2007 конференции. Часть первая. (Под ред. Олега Воробьёва). Красноярск: ИВМ СО РАН, СФУ, КГТЭИ, СИБУП, Издательство "Гротеск", 2007, - с. 271-278

15. Рубцов А.Г. Матричное представление частотного словаря в задаче восстановления отсутствующих данных / Сенашова М.Ю., Рубцов А.Г. // Материалы V Всесибирского конгресса женщин–математиков, 15-18 января 2008 г. Красноярск: РИО СФУ, 2008, –с. 367-372

Рубцов Антон Геннадьевич
Восстановление отсутствующих данных в символьных последовательностях.
Автореф. дисс. на соискание учёной степени кандидата физ.-мат. наук.
Подписано в печать 15.01.2010. Заказ №_____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.