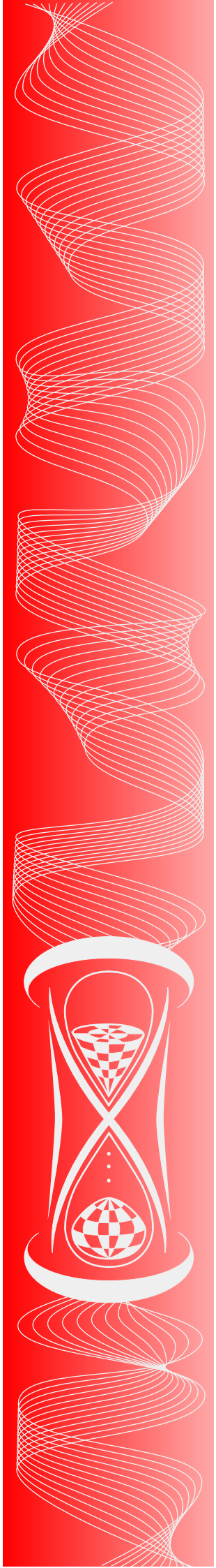
ISSN 2222-8772



Математические Структуры и Моделирование

№ 4(40)
2016

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМ. Ф.М. ДОСТОЕВСКОГО»

МАТЕМАТИЧЕСКИЕ СТРУКТУРЫ И МОДЕЛИРОВАНИЕ

№ 4(40)

Омск
2016

Математические структуры и моделирование. — Омск : Омский государственный университет, 2016. — № 4(40). – 167 с.

ISSN 2222-8772 (print)

ISSN 2222-8799 (online)

Учредитель

Федеральное государственное бюджетное образовательное учреждение высшего образования «Омский государственный университет им. Ф. М. Достоевского».

Свидетельство о регистрации средства массовой информации ПИ № ФС77-57908 от 28 апреля 2014 г.

Адрес научной редакции

Россия, 644077, Омск, пр. Мира 55А, Омский государственный университет им. Ф. М. Достоевского, факультет компьютерных наук.

E-mail: guts@omsu.ru, lavrov@omsu.ru

МАТЕМАТИЧЕСКИЕ СТРУКТУРЫ и МОДЕЛИРОВАНИЕ

Журнал основан в 1998 году.

В журнале публикуются статьи, в которых излагаются результаты исследований по фундаментальной и прикладной математике, теоретической физике, компьютерным наукам, философии и истории математики и информатики, а также размышления, касающиеся окружающей нас природы и общества. Объекты исследования должны быть представлены в форме некоторых математических структур и моделей.

Все статьи журнала проходят обязательное рецензирование.

Журнал является реферируемым. Рефераты статей публикуются в «Реферативном журнале», в журналах «Zentralblat für Mathematik» (Германия), «Mathematical Reviews» (США), индексируется в РИНЦ (elibrary.ru).

Входит в Перечень научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание учёных степеней в соответствии с приказом Минобрнауки России от 25 июля 2014 г. № 793.

Электронная версия журнала представлена в сети Интернет по адресам:

http://msm.univer.omsk.su
http://msm.omsu.ru

Подписной индекс по каталогу «Пресса России»: 94082

Электронная почта главного редактора:

guts@omsu.ru

Электронная почта выпускающего редактора:

lavrov@omsu.ru

СОДЕРЖАНИЕ

Фундаментальная математика и физика

Прикладная математика и моделирование

Продолжение на следующей странице

Наши публикации

Компьютерные науки

Информационная безопасность

Математические структуры и моделирование
2016. № 4(40). С. 5–16

УДК 519.237

МЕРА СХОДСТВА ПОСЛЕДОВАТЕЛЬНОСТЕЙ ОДИНАКОВОЙ РАЗМЕРНОСТИ

Н.А. Гайдамакин
д.т.н., профессор, e-mail: haid2@bk.ru

Уральский федеральный университет им. Б.Н. Ельцина, г. Екатеринбург

Аннотация. Критериями сходства являются количество и качество совпадений элементов в одинаковых позициях с приоритетом качества совпадений. Качество совпадений трактуется как максимизация совпадений элементов в смежных позициях. Приведены примеры расчётов меры сходства для различных видов весовых коэффициентов значимости совпадения элементов в m смежных позициях (совпадения m-грамм).

Ключевые слова: мера сходства, конечные последовательности, m-граммы, размещения.

Рассматривается сходство (similarity) конечных последовательностей $\psi_i = \{a_{i_1}, a_{i_2}, \ldots, a_{i_K}\}$ как объектов, представляющих упорядоченные наборы из K элементов a_{i_m} некоторого множества A произвольной природы, где i — номер последовательности, m — номер позиции.

Меры сходства [1] являются разновидностями мер близости [2] и выражаются функциями s от элементов скалярного произведения множества X на себя $(X \times X)$. При этом значения функции $s \in R$ должны удовлетворять требованиям неотрицательности $(s \geqslant 0)$ и симметричности $(s(x, y) = s(y, x))$.

На практике в качестве мер сходства используются величины из диапазона $[0, 1]$[1].

Одной из универсальных мер близости многомерных объектов одинаковой размерности является расстояние Хэмминга [1] $d_H(A_1, A_2)$:

$$d_H(A_1, A_2) = \sum_{m=1}^{K} |a_{1_m} - a_{2_m}|, \tag{1}$$

где A_1 и A_2 — многомерные объекты, характеризующиеся K числовыми параметрами.

Смысл меры Хэмминга заключается в том, что в многомерном пространстве признаков два объекта тем ближе, чем по меньшему количеству признаков (параметров) они различаются. В случае двоичного характера признаков в качестве меры сходства можно использовать величину μ_H, определяемую на основе

[1] 1 — полное сходство, 0 — полное несходство.

расстояния Хэмминга по следующему соотношению:

$$\mu_H = 1 - \frac{d_H(A_1, A_2)}{K}. \tag{2}$$

Конечные последовательности одинаковой размерности являются объектами, элементы которых в соответствующих позициях можно рассматривать как признаки в общем случае нечисловой природы. Для применения меры Хэмминга к таким объектам в выражении (1) числовую операцию $|a_{i_m} - a_{j_m}|$ необходимо заменить двоичной функцией сравнения:

$$\delta(a_{i_m}, a_{j_m}) = \begin{cases} 0, & \text{если } a_{i_m} = a_{j_m} \\ 1, & \text{если } a_{i_m} \neq a_{j_m} \end{cases}. \tag{3}$$

В результате мера сходства μ_H конечных последовательностей $\psi_i = \{a_{i_1}, a_{i_2}, \ldots, a_{i_K}\}$ и $\psi_j = \{a_{j_1}, a_{j_2}, \ldots, a_{j_K}\}$ по Хэммингу вычисляется на основе следующего выражения:

$$\mu_H = 1 - \frac{1}{K} \sum_{m=1}^{K} \delta(a_{i_m}, a_{j_m}). \tag{4}$$

Вместе с тем, очевидно сходство последовательностей определяется не только количеством совпадений элементов в одинаковых позициях, но и порядком следования совпадений, в частности, в смежных или разрозненных позициях.

Известны [1, 3] и широко применяются меры сходства текстовых строк (слов) как последовательностей символов на основе техники m-грамм[2]. m-граммами являются сочетания из m смежных символов[3]. В качестве меры сходства строк X и Y используется величина μ_m:

$$\mu_m = \frac{2m(X, Y)}{m(X) + m(Y)}, \tag{5}$$

где $m(X)$, $m(Y)$ — количество m-грамм в строках X и Y, соответственно; $m(X, Y)$ — количество m-грамм, одновременно входящих в строку X и в строку Y, *независимо от позиции расположения*.

Смысл меры на основе m-грамм заключается в том, что сходство строк символов (слов) рассматривается в контексте близости их лексического значения, обусловленного некоторым ядром (ядрами) в форме подпоследовательности из m-символов, которая:

а) может сдвигаться по позициям ввиду префиксов, суффиксов, окончаний;

б) «размываться» с учётом определённых особенностей произношения, правописания и т.п.

[2] В некоторых источниках «N-грамм», «q-грамм».

[3] Например, при $m = 2$ (двуграммы) в строке «слово» 6 двуграмм с учётом пустого (отсутствующего) символа, обозначаемого «_» («_с», «сл», «ло», «ов», «во», «о_»).

В результате на основе техники m-грамм можно анализировать сходство строк как последовательностей символов не обязательно одинаковой размерности. Кроме того, сходство строк, как отмечалось, определяется не по совпадению m-грамм, начинающихся в одинаковых позициях, а как одновременное вхождение m-грамм в объекты сравнения *независимо от их позиций в строках сравнения*.

Выбор значений m, являющийся центральным при формировании меры μ_m, осуществляется на основе лексических или семантических (языковых) соображений. Лексическое значение слов определяется их нормализованной основой, выражающейся корнем, который, как правило, состоит из двух-четырёх символов. Семантические значения текстов, точнее предложений, определяются отдельными словами и их сочетаниями. Поэтому при анализе сходства слов как последовательности символов или текстов как последовательности слов на практике ограничиваются анализом только по одному значению m, в большинстве случаев по двуграммам ($m = 2$) или реже по триграммам ($m = 3$).

В общем же случае, когда природа элементов последовательностей является произвольной, выбор значений m является неопределённым.

Далее ограничимся подходом, при котором конечные последовательности одинаковой размерности рассматриваются как случайные реализации некоторой исходной закономерности следования элементов (эталона, типичной последовательности). При этом закономерности следования элементов жёстко связаны с номерами позиций последовательности. В результате какие-либо неискажённые фрагменты исходной закономерности фиксированы по месту в последовательности (т.е. не могут «сдвигаться» по номерам позиций последовательностей) и анализ сходства необходимо вести только по совпадению элементов или их совокупностей (m-грамм) в одинаковых позициях.

Из общих соображений можно определить следующие критерии сходства конечных последовательностей при произвольной природе элементов.

1. *Количественный критерий сходства*:

– сходство последовательностей тем больше, чем в большем количестве позиций совпадают их элементы.

2. *Качественный критерий сходства*:

– сходство последовательностей тем больше, чем больше совпадений элементов в смежных позициях.

Следуя терминологии в технике m-грамм, будем называть одиночные совпадения элементов совпадениями одограмм, совпадения в двух смежных позициях совпадениями двуграмм, совпадения в трёх смежных позициях совпадениями триграмм и т.д. Заметим, что по смыслу совпадения совокупности смежных элементов одна совпадающая m-грамма должна быть отделена от другой совпадающей m-граммы минимум одной позицией, в которой элементы сравниваемых последовательностей не совпадают. В противном случае имеет место совпадение одной m-граммы, в которой число m определяется количеством подряд следующих позиций совпавших элементов (см. рис. 1).

Обозначим количество «одограммных» совпадений n_1, «двуграммных» n_2, . . . , количество «m-граммных» совпадений n_m.

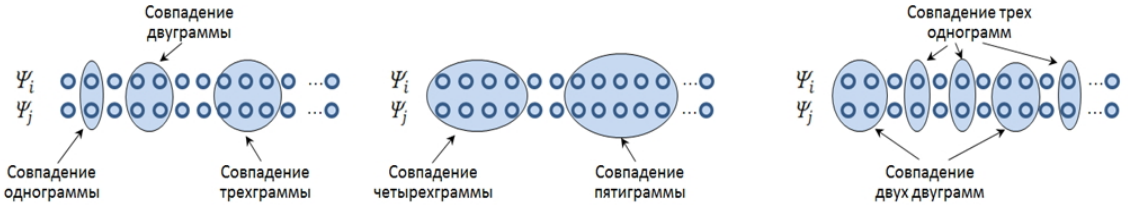


Рис. 1. Примеры сходства последовательностей по совпадению однограмм, двуграмм и т.д.

Будем считать, что с точки зрения сходства, номера позиций, с которых начинаются совпадающие m-граммы как неискажённые фрагменты исходной закономерности (в начале последовательности, в середине, в конце), не имеют значения. Тогда совпадение m-граммы, начинающейся, скажем, с 1-й позиции или со 2-й позиции, или с 5-й позиции и т.д., рассматриваются как один и тот же случай сходства (см. рис. 2).

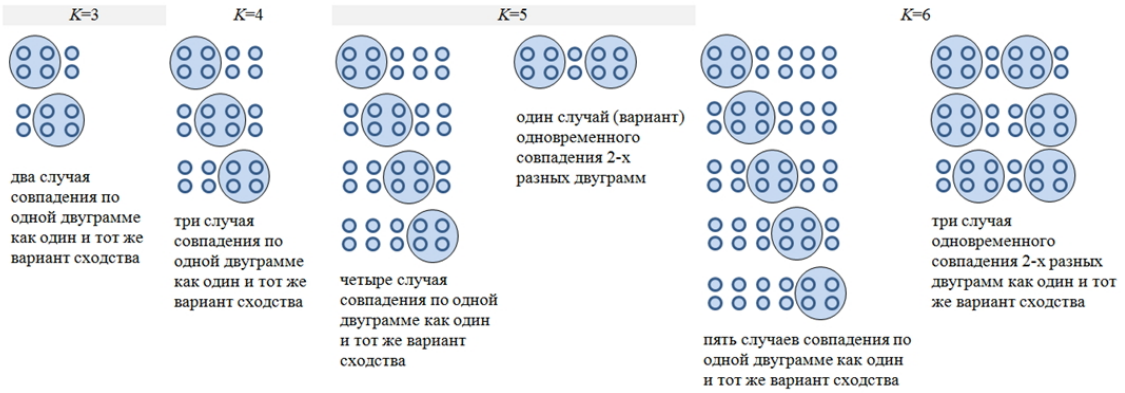


Рис. 2. Варианты совпадения двуграмм при $K = 3$, $K = 4$, $K = 5$ и $K = 6$

В результате для определённого значения размерности сравниваемых последовательностей можно построить ряд вариантов совпадений элементов, при которых сходство последовательностей должно возрастать в соответствии с ростом количества и качества совпадений. На рис. 3 представлены варианты сходства последовательностей при $K = 10$, расположенные слева направо по увеличению количества совпадений, а при одинаковом количестве по увеличению качества совпадений. При этом увеличение качества совпадений трактуется как появление хотя бы одной совпавшей m-граммы, размерность которой (m) на единицу выше самой старшей m-граммы в предыдущем варианте. Будем называть такой подход к определению сходства последовательностей количественно-качественным с приоритетом количества совпадений.

Отметим, что такой подход к трактовке повышения сходства последовательностей может быть не единственным. В частности, возникают сомнения в его «справедливости» при переходе от варианта с совпадением m элементов в виде совпадения одной m-граммы к варианту с совпадением $(m + 1)$ элементов,

Рис. 3. Варианты сходства последовательностей при $K = 10$

который реализован совпадением $(m + 1)$ «разбросанных» по разным позициям одногramм (варианты 4, 7, 12 и аналогичные по смыслу варианты 19, 29, 40, 50 на рис. 3). Также не является очевидным увеличение сходства при переходе в парах вариантов 16-17, 24-25, 31-32, 35-36, 41-42, 45-46.

Количество совпадений элементов определяется суммой произведений значений n_m (количество совпавших m-грамм) на число m (количество элементов в m-грамме) — $\sum_{m=1}^{K} mn_m$. Каждое слагаемое mn_m даёт вклад в общее количество совпадений элементов, реализованный совпадением соответствующих m-грамм. Как следует из выше приведённых критериев сходства, сходство последовательностей должно быть тем выше, чем более «старшими» m-граммами (с большими значениями m) оно реализовано. Тогда одним из подходов к установлению меры сходства последовательностей может быть «взвешивание» слагаемых mn_m в зависимости от размерности m-грамм.

В итоге приходим к следующей мере сходства μ, зависящей от количества и качества совпадений элементов конечных последовательностей:

$$\mu = \frac{1}{K} \sum_{m=1}^{K} mn_m c_m, \tag{6}$$

где n_m — количество совпадений m-грамм; c_m — вес значимости совпадения m-граммы в сходстве последовательностей, $c_m \leqslant 1$.

Прежде всего докажем, что величина μ удовлетворяет требованиям, предъявляемым к мерам сходства:

Лемма 1. *Величина μ, определяемая по соотношению (6), является неотрицательной в диапазоне $[0, 1]$ и обладает свойством симметричности.*

Доказательство. Неотрицательность очевидна, поскольку величина μ формируется суммой положительных величин $mn_m c_m$ $(m = 1, 2, \ldots, K)$.

При отсутствии совпадений элементов все $n_m = 0$ $(m = 1, 2, \ldots, K)$, величина $\mu = 0$ (полное несходство).

При полном сходстве, т.е. когда все K элементов двух последовательностей совпадают, $n_K = 1$, а все остальные $n_m = 0$ $(m = 1, 2, \ldots, K - 1)$. При этом величина $\mu = 1$.

В случае, когда сходство неполное, т.е. совпадения элементов есть и некоторые $n_m \neq 0$, величина $\mu \leqslant 1$, поскольку количество совпадений $\sum_{m=1}^{K} mn_m \leqslant K$, а весовые коэффициенты $c_m \leqslant 1$.

При определении факта совпадения элементов не важен порядок сравнения – совпал ли элемент в какой-либо позиции из первой последовательности с элементом в той же позиции из второй последовательности, или наоборот, совпал ли элемент из второй последовательности с элементом из первой последовательности. В таком случае количества совпавших m-грамм n_m, которые являются переменными величинами в соотношении (6), одинаковы как при сравнении сходства первой последовательности со второй, так и в случае сравнения второй последовательности с первой. В результате μ с точки зрения порядка сравнения последовательностей симметрична. ■

Нетрудно увидеть, что частным случаем меры сходства μ является мера сходства по Хэммингу μ_H. Докажем следующее утверждение.

Лемма 2. *Мера сходства μ, определяемая по соотношению* (6)*, при $c_m = 1$ $(m = 1, 2, \ldots, K)$ эквивалентна мере сходства по Хэммингу μ_H.*

Доказательство. В соответствии с соотношением (4) мера сходства по Хэммингу μ_H определяется количеством совпавших в одинаковых позициях элементов, отнесённых к общему числу позиций (размерности сравниваемых последовательностей) — K.

Обозначим количество совпадений элементов как η. Тогда $\mu_H = \frac{\eta}{K}$.

Мера сходства μ при $c_m = 1$ $(m = 1, 2, \ldots, K)$ складывается из суммы произведений размерностей m-грамм (чисел m) на количество соответствующих m-грамм (числа n_m), отнесённых к величине K. m-грамма представляет собой совпадение элементов в m смежных позициях, отделённых от других совпадений слева и справа, по крайней мере, одной позицией, в которой элементы не совпали (см. рис. 1). Отсюда следует, что каждое слагаемое mn_m, как отмечалось, является вкладом в общее количество совпадений η за счёт совпадений соответствующих m-грамм. В результате

$$\sum_{m=1}^{K} mn_m = \eta.$$

В итоге $\mu = \frac{\eta}{K} = \mu_H$. ■

Одним из логичных подходов к определению коэффициентов c_m может быть их возрастание от соотношения $\frac{m}{K}$. Действительно, чем большую часть последовательностей составляет совпадающая m-грамма, тем более значимым должен быть ее вклад в их сходство.

Отсюда коэффициенты c_m должны определяться некоторой функцией от $\frac{m}{K}$.

На рис. 4 представлены расчёты коэффициента сходства последовательностей из 10 элементов $(K = 10)$ для различных по количеству и качеству совпадений и при различных видах функции $c_m = f\left(\frac{m}{K}\right)$.

При $c_m = 1$, что, как уже отмечалось, соответствует мере Хэмминга, мера сходства $\mu = \frac{1}{K}\sum_{m=1}^{K} mn_m c_m$ ступенчато возрастает пропорционально общему количеству совпадений, не различая их разное качество по одногра́ммам, двугра́ммам, тригра́ммам и т.д.

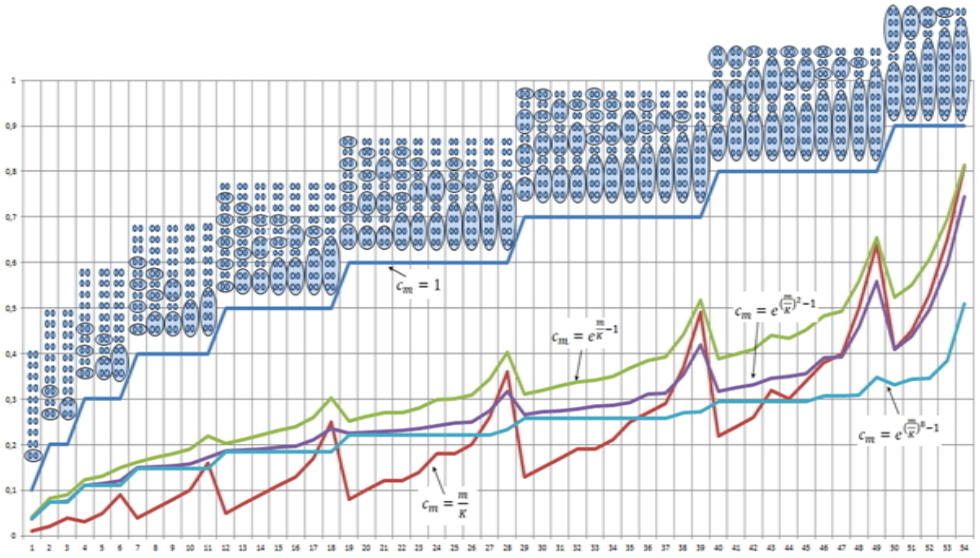


Рис. 4. Зависимость коэффициента сходства последовательностей при $K = 10$ от количества и качества совпадений при различных видах функции $c_m = f\left(\frac{m}{K}\right)$

Простая пропорциональная зависимость весов значимости c_m от $\frac{m}{K}$ даёт по ряду вариантов, приведённому на рис. 3, нелинейно и немонотонно возрастающую картину повышения сходства последовательностей в зависимости от количества и качества совпадений. Так, в отмеченных выше при анализе рис. 3 «критичных» переходов количества-качества совпадениях (варианты 3–4, 6–7, 11–12, 18–19, 28–29, 39–40) сходство последовательностей уменьшается вопреки росту количества совпавших элементов.

Такое поведение меры μ в соответствующих случаях отражает приоритет качества совпадений.

На рис. 4 приведены расчёты и по другим видам функций $c_m = f\left(\frac{m}{K}\right)$, которые также демонстрируют приоритет качества совпадений элементов в сходстве последовательностей.

Другим подходом к установлению весов значимости c_m может быть учёт максимального количества $max_m(K)$ возможных совпадений по конкретной m-грамме (скажем, по двуграммам) в рамках определённой размерности сравниваемых последовательностей. На рис. 5 представлены случаи максимального количества совпадений двуграмм[4] при различных значениях K.

Очевидно, вес m-граммы должен быть тем больше, чем меньше совпадений m-грамм может реализоваться в пределах K-элементов последовательности, т.е. чем меньше $max_m(K)$. Так, вес совпадения двуграммы $c_2(K)$ в последовательности из 3-х или 4-х элементов должен быть выше, чем вес совпадения двуграммы $c_2(K)$ при $K = 5$, $K = 6$, $K = 7$, поскольку при $K = 3$ и $K = 4$ совпадение одной двуграммы реализует весь набор случаев сходства последовательностей по совпадению двуграмм, а при $K = 5$, $K = 6$, $K = 7$ только один

[4]Напомним, что совпадение двуграммы, начинающейся с 1-й или с 2-й или с 3-й позиции, рассматривается как один и тот же случай сходства (см. рис. 2).

Рис. 5. Максимальное количество возможных совпадений по двуграммам при $K = 3$, $K = 4$, $K = 5$, $K = 6$, $K = 7$ и $K = 8$

из 2-х возможных случаев сходства по совпадениям двуграмм.

Докажем следующее утверждение.

Лемма 3. *Максимальное значение количества m-грамм, по которым могут быть совпадения элементов в двух конечных последовательностях размерности K, определяется целой частью отношения величин $(K + 1)$ и $(m + 1)$:*

$$max_m(K) = \left[\frac{K+1}{m+1}\right], \tag{7}$$

где $[\cdot]$ означает целую часть числа.

Доказательство. Величины $max_m(K)$ определяются максимально возможным количеством совпадающих и не сливающихся m-грамм по двум последовательностям из K-элементов.

Совершенно очевидно, что максимальное количество совпадающих m-грамм реализуется их выстраиванием подряд через одну позицию, в которой элементы последовательностей не совпадают.

Отсюда максимальное количество совпадающих и несливающихся m-грамм в последовательностях из K позиций (из K элементов) определяется количеством совокупностей из $m + 1$ идущих подряд до завершения последовательностей позиций — см. рис. 6.

При этом для «последней» совпадающей m-граммы, если она заканчивается в конечной позиции последовательностей, не требуется далее позиция, отделяющая её от следующей m-граммы, которой нет и быть не может — см. на рис. 6 ситуацию по совпадению двуграмм при $K = 5$ и $K = 8$. Отсюда следует, что в расчёт количества позиций, на которых разместится целое число подряд идущих групп из $(m + 1)$ позиций, следует к размерности последовательностей K добавить ещё одну позицию.

Таким образом,

$$max_m(K) = \left[\frac{K+1}{m+1}\right].$$

■

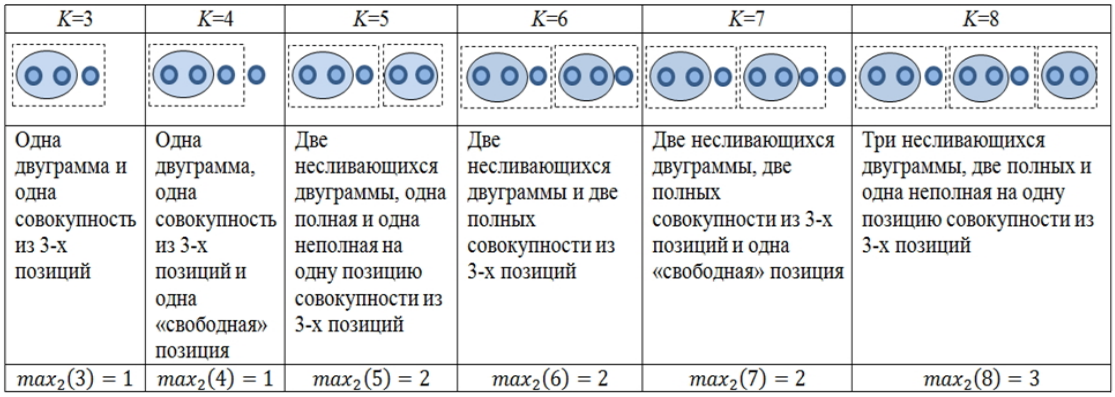


Рис. 6. Максимальное количество не сливающихся двуграмм, трёхэлементных смежных позиций и соответствующие величины $max_2(K)$ при $K = 3$, $K = 4$, $K = 5$, $K = 6$, $K = 7$ и $K = 8$

В таблице 1 представлены значения максимального количества совпадений m-грамм $max_m(K)$ при различных значениях размерности последовательностей ψ_i.

Таблица 1. Значения величины $max_m(K)$ при $m = 1, 2, \ldots, 6$ и $K = 2, 3, \ldots, 12$

	K=2	*K=3*	*K=4*	*K=5*	*K=6*	*K=7*	*K=8*	*K=9*	*K=10*	*K=11*	*K=12*
$max_1(K)$	1	2	2	3	3	4	4	5	5	6	6
$max_2(K)$	1	1	1	2	2	2	3	3	3	4	4
$max_3(K)$	0	1	1	1	1	2	2	2	2	3	3
$max_4(K)$	0	0	1	1	1	1	1	2	2	2	2
$max_5(K)$	0	0	0	1	1	1	1	1	1	2	2
$max_6(K)$	0	0	0	0	1	1	1	1	1	1	1

Как уже отмечалось, можно предполагать, что значимость в сходстве последовательностей совпадения m-граммы тем больше, чем меньше величина $max_m(K)$. На рис. 7 представлены расчёты коэффициента сходства последовательностей из 10 элементов ($K = 10$) при различных видах функции $c_m = f\left(\frac{1}{max_m(K)}\right)$.

Как видно из приведённых на рис. 7 расчётов, использование величин $max_m(K)$ также реализует принцип приоритетности качества совпадений, но со своей спецификой «бросков» меры сходства при изменениях количества и качества совпадений. Так, в частности, при использовании для c_m функций с аргументом $\frac{1}{max_m(K)}$ существенно увеличиваются «броски» меры сходства $\mu = \frac{1}{K}\sum_{m=1}^{K} m n_m c_m$ в «критичных» точках — при переходе от варианта совпадения по одной m-грамме к варианту с $(m + 1)$ совпадений одногрмм.

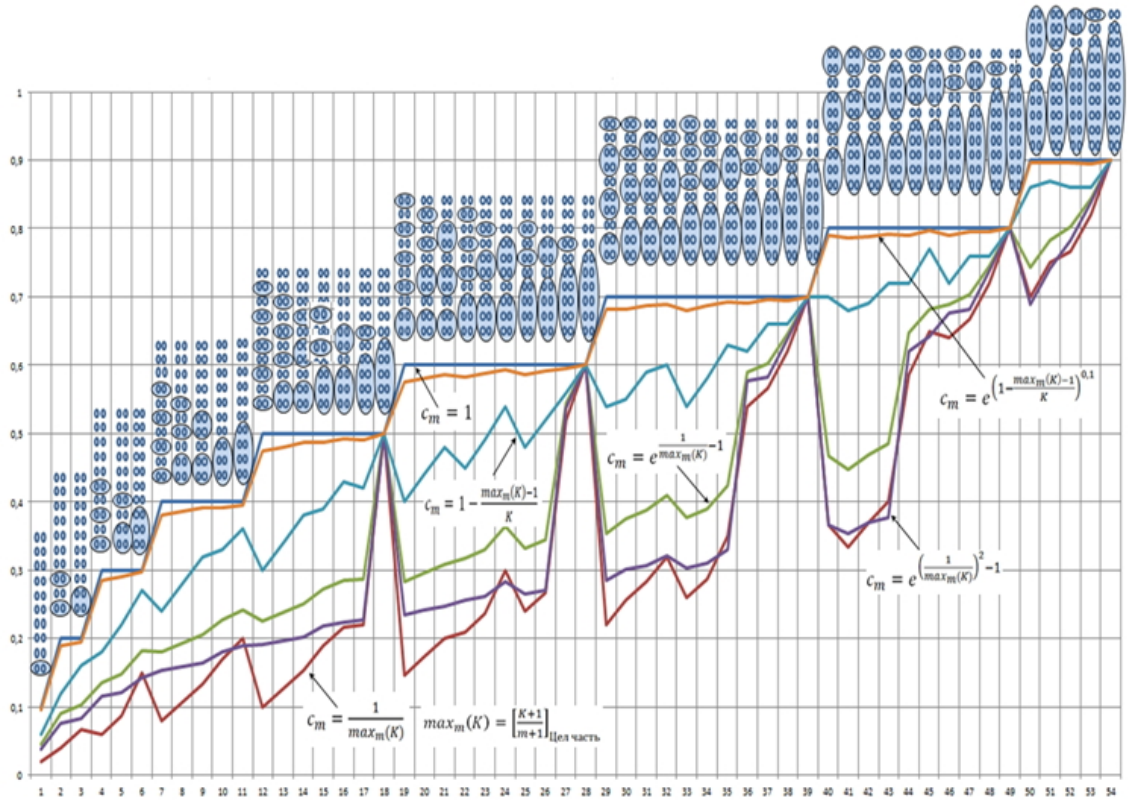


Рис. 7. Зависимость коэффициента сходства последовательностей при $K = 10$ от количества и качества совпадений при различных видах функции $c_m = f\left(\frac{1}{max_m(K)}\right)$

Для наглядности анализа специфики действия приоритета качества совпадений на рис. 8 приведены возрастающие ряды вариантов сходства последовательностей при различных видах функции $c_m = f\left(\frac{m}{K}\right)$ в сравнении с исходным рядом при приоритете количества совпадений (см. рис. 3).

Таким образом, используя различные виды функций, устанавливающих коэффициенты c_m, можно обеспечивать разный характер поведения меры сходства конечных последовательностей в общей идеологии количества и качества совпадений их элементов. Выбор тех или иных функций значимости c_m может определяться спецификой природы анализируемых последовательностей, их элементов и особенностями исследовательских задач.

Следует также отметить, что в контексте рассмотренной задачи, близкими по смыслу объектами к конечным последовательностям являются размещения [4]. Размещения представляют собой упорядоченные наборы из K элементов из некоторого множества N элементов. При этом размещения могут быть без повторов, когда какой-либо элемент может входить в размещение только один раз или с повторами в противном случае. Очевидно сходство размещений определяется, как и в случае конечных последовательностей, тем, насколько совпадают их элементы в соответствующих позициях. В этом смысле представленная мера сходства может быть применена также и к анализу сходства

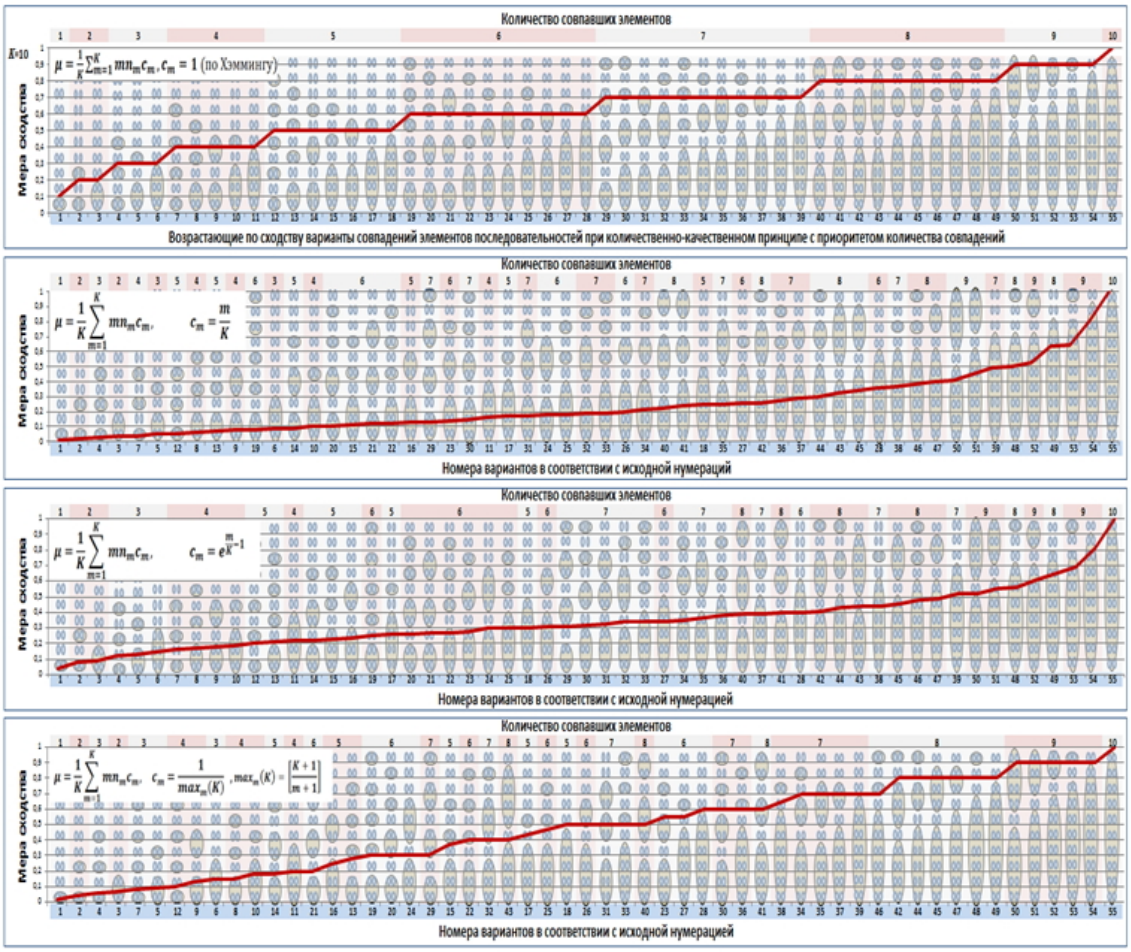


Рис. 8. Возрастающие по сходству ряды вариантов совпадений элементов последовательностей при различных видах функции $c_m = f\left(\frac{m}{K}\right)$

размещений.

Литература

1. Deza M., Deza E. Encyclopedia of Distances. 2014. URL: `http://www.liga.ens.fr/~deza/1-59.pdf` (дата обращения: 03.07.2016).
2. Сёмкин Б.И., Двойченков В.И. Об эквивалентности мер сходства и различия // Исследование систем. Т. 1. Анализ сложных систем. Владивосток: ДВНЦ АН СССР, 1973. С. 95–104.
3. Ukkonen E. Approximate string-matching with q-grams and maximal matches // Theoretical Computer Science. 1992. Vol. 92(1). P. 191–211.
4. Андерсон Д. Дискретная математика и комбинаторика. М. : Вильямс, 2006. 960 с.

MEASURES OF SIMILARITY AMONG FINITE SEQUENCES

N.A. Gaydamakin
Dr.Sc. (Eng.), Professor, e-mail: haid2@bk.ru

Ural Federal University n.a. B.N. Yeltsin, c. Ekaterinburg

Abstract. This article presents a measure of similarity among finite sequences of arbitrary nature elements. The measure uses number and quality of matches between elements at the same positions as similarity criteria with priority to quality of the match. Quality of the match is treated as maximization of matches between adjacent elements. Examples of similarity measure calculations are adduced for a number of match significance weight coefficients in m-adjacent element positions (m-grams matches).

Keywords: measure of similarity, finite sequences, m-grams, variations.

Дата поступления в редакцию: 23.08.2016

Математические структуры и моделирование
2016. № 4(40). С. 17–23

УДК 519.214

О МОМЕНТАХ СИММЕТРИЧЕСКИХ ФУНКЦИЙ ОТ ЗАВИСИМЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

А.Г. Гринь
д.ф.-м.н., профессор, e-mail: griniran@gmail.com

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В работе получены оценки для моментов и равномерная интегрируемость определённого класса функций от случайных величин, которые образуют стационарную последовательность с равномерно сильным перемешиванием.

Ключевые слова: симметрические функции от случайных величин, равномерно сильное перемешивание, равномерная интегрируемость.

В работе [1] на основе некоторого аналога неравенства М. Пелиград получены оценки моментов так называемых обобщённых сумм слабо зависимых случайных величин. Для «обычных» сумм такие оценки впервые получены И. А. Ибрагимовым (см., например, [3, лемма 18.5.1]); на этих оценках базировалось доказательство центральной предельной теоремы для последовательностей с φ-перемешиванием. В настоящей работе показывается, как разработанную в [1] технику модифицировать для более общей ситуации — для функций от случайных величин, не являющихся, вообще говоря, результатом последовательного применения бинарных операций (обобщённых сумм).

Пусть при каждом $n \in \mathbb{N}$ определена вещественнозначная функция $f(\mathbf{x}) = f(x_1, x_2, ..., x_n),\ x_1, ..., x_n \in \mathbb{D} \subseteq \mathbb{R}$ (то есть, определена последовательность функций, но, чтобы не загромождать рассуждений, мы не будем подчёркивать зависимость f от n какими-либо индексами и называть f последовательностью).

Будем предполагать, что функция f при любых $x_1, ..., x_n \in \mathbb{D},\ y_1, ..., y_n \in \mathbb{D}$ удовлетворяет следующим условиям (условия A):

A_1. Симметричность: $f(x_{i_1}, ..., x_{i_n}) = f(x_1, x_2, ..., x_n)$ для любой перестановки $\{i_1, ..., i_n\}$ множества $\{1, ..., n\}$;

A_2. $f(x_1, x_2, ..., x_{n-1}, 0) = f(x_1, x_2, ..., x_{n-1})$;

A_3. $|f(\mathbf{x} \pm \mathbf{y})| \leqslant |f(\mathbf{x})| + |f(\mathbf{y})|$.

Ясно, что из условия A_3 следует утверждение:

A'_3. $||f(x_1, x_2, ..., x_n)| - |f(x_1, x_2, ..., x_k)|| \leqslant |f(x_{k+1}, , ..., x_n)|$ для любого $1 \leqslant k \leqslant n$. (Согласно сказанному выше $f(x_1, x_2, ..., x_k) = f(x_1, ..., x_k, 0, ..., 0)$.)

Более того, из A_3 вытекает $|f(x_k)| \leqslant |f(x_1, x_2, ..., x_k)| + |f(x_1, x_2, ..., x_{k-1})|$, $k = 2, ..., n$, так что

$$\max_{1 \leqslant k \leqslant n} |f(x_k)| \leqslant 2 \max_{1 \leqslant k \leqslant n} |f(x_1, x_2, ..., x_k)| \leqslant 2(|f(x_1)| + ... + |f(x_n)|). \qquad (1)$$

В [2] приводятся многочисленные примеры функций, удовлетворяющих условиям A.

Пусть $\{\xi_n\} = \{\xi_n,\ n = 1, 2, ...\}$ — стационарная в узком смысле последовательность и пусть $\mathcal{F}_{\leqslant n}$ и $\mathcal{F}_{\geqslant n} - \sigma$ -алгебры, порождённые семействами $\{\xi_i : i \leqslant n\}$ и $\{\xi_i : i \geqslant n\}$. Говорят, что последовательность $\{\xi_n\}$ удовлетворяет *условию равномерно сильного перемешивания (φ-перемешивания)* с коэффициентом перемешивания $\varphi(n)$, если

$$\varphi(n) = \sup\left\{\frac{|\mathbf{P}(AB) - \mathbf{P}(A)\mathbf{P}(B)|}{\mathbf{P}(A)} :\ A \in \mathcal{F}_{\leqslant 0},\ B \in \mathcal{F}_{\geqslant n}\right\} \to 0, \quad n \to \infty.$$

Будем обозначать

$$X_{k,m}(b) = f\left(\frac{\xi_k}{b}, ..., \frac{\xi_m}{b}\right), \quad X_n(b) = X_{1,n}(b), \quad X_n = X_n(1),$$

$$\overline{X}_n(b) = \max_{1\leqslant k\leqslant n} |X_k(b)|, \quad Y_k(b) = f\left(\frac{\xi_k}{b}\right), \quad k, m, n \in \mathbb{N},\ b > 0.$$

Лемма 1. *Пусть $\varepsilon > 0,\ x > 0$ и $k \leqslant n$, а функция f удовлетворяет условиям A. Если последовательность $\{c_n\}$ такова, что*

$$\max_{1\leqslant j\leqslant n} \mathbf{P}\{|X_j(c_n)| \geqslant \varepsilon\} + \varphi(m) \leqslant \gamma < 1,$$

то при любых $a > 0$

$$\mathbf{P}\{\overline{X}_k(c_n) \geqslant 2x + \varepsilon\} \leqslant \frac{1}{1-\gamma}\left(\mathbf{P}\{|X_k(c_n)| \geqslant x\} + \mathbf{P}\left\{\max_{1\leqslant j\leqslant n} |Y_j(c_n)| \geqslant \frac{x}{m}\right\}\right).$$

Доказательство. Пусть $E_i = \{\overline{X}_{i-1}(c_n) < 2x + \varepsilon \leqslant |X_i(c_n)|\},\ i = 1, ..., n$. Тогда $E_iE_j = \varnothing,\ i \neq j,\ \bigcup\limits_{i=1}^{k} E_i = \{\overline{X}_k(c_n) \geqslant 2x + \varepsilon\}$.

В силу свойства A'_3 при $i + m \leqslant k$

$$\{|X_i(c_n)| \geqslant 2x + \varepsilon,\ \max_{1\leqslant j\leqslant n} |Y_j(c_n)| < \frac{x}{m},\ |X_{i+m,k}(c_n)| < \varepsilon\} \subseteq \{|X_k(c_n)| \geqslant x\},$$

то есть при $1 \leqslant k \leqslant n - 1$

$$\{|X_k(c_n)| < x\} \subseteq \{|X_i(c_n)| < 2x + \varepsilon\} \cup \{|X_{i+m,k}(c_n)| \geqslant \varepsilon\} \cup \left\{\max_{1\leqslant j\leqslant n} |Y_j(c_n)| \geqslant \frac{x}{m}\right\},$$

откуда

$$\{|X_k(c_n)| < x, E_i\} \subseteq \{|X_{i+m,k}(c_n)| \geqslant \varepsilon, E_i\} \cup \left\{\max_{1\leqslant j\leqslant n} |Y_j(c_n)| \geqslant \frac{x}{m}, E_i\right\}. \qquad (2)$$

С помощью (2) и условия φ-перемешивания получаем

$$\mathbf{P}\{\overline{X}_k(c_n) \geqslant x\} \leqslant \mathbf{P}\{|X_k(c_n)| \geqslant a\} + \sum_{i=1}^{k} \mathbf{P}\{|X_k(c_n)| < x,\ E_i\} \leqslant$$

$$\leqslant \mathbf{P}\{|X_k(c_n)| \geqslant x\} + \sum_{i=1}^{k} \mathbf{P}\{|X_{i+m,k}(c_n)| \geqslant \varepsilon, E_i\} + \mathbf{P}\left\{\max_{1\leqslant j\leqslant n} |Y_j(c_n)| \geqslant \frac{x}{m}\right\} \leqslant$$

$$\leqslant \mathbf{P}\{|X_k(c_n)| \geqslant x\} + \left(\max_{1\leqslant i\leqslant k} \mathbf{P}\{|X_i(c_n)| \geqslant \varepsilon\} + \varphi(m)\right) \sum_{i=1}^{k} \mathbf{P}\{E_i\} +$$

$$+\mathbf{P}\left\{\max_{1\leqslant j\leqslant n} |Y_j(c_n)| \geqslant \frac{x}{m}\right\} \leqslant \mathbf{P}\{|X_k(c_n)| \geqslant x\} + \gamma \mathbf{P}\{\overline{X}_n(c_n) \geqslant 2x + \varepsilon\} +$$

$$+\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} |Y_j(c_n)| \geqslant \frac{x}{m}\right\},$$

откуда следует утверждение леммы. ■

Следующее предложение — это аналог неравенства М. Пелиград (леммы 3.1 из [4]).

Лемма 2. *Если последовательность* $\{c_n\}$ *и* $m > 0$ *таковы,*

$$\max_{1\leqslant k\leqslant n} \mathbf{P}\{|X_k(c_n)| \geqslant \varepsilon\} + \varphi(m) \leqslant \gamma < 1,$$

то при любом $x > 0$

$$\mathbf{P}\{|X_n(c_n)| \geqslant 3x + 2\varepsilon\} \leqslant \frac{\gamma}{1-\gamma}\mathbf{P}\{|X_n(c_n)| \geqslant x\} + \frac{1}{1-\gamma}\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} |Y_k(c_n)| \geqslant \frac{x}{m}\right\}.$$

Доказательство. Пусть $E_k = \{\overline{X}_{k-1}(c_n) < 2x+\varepsilon \leqslant |X_k(c_n)|\},\ k = 1, ..., n$. Тогда $E_iE_j = \varnothing,\ i \neq j,\ \bigcup\limits_{k=1}^{n} E_k = \{\overline{X}_n(c_n) \geqslant 2x + \varepsilon\}$. В силу свойства A'_3

$$|X_n(c_n)| \leqslant |X_{k-1}(c_n)| + \sum_{j=k}^{k+m} |Y_j(c_n)| + |X_{k+m,n}(c_n)|,$$

откуда следует

$$\left\{|X_n(c_n)| \geqslant 3x + 2\varepsilon,\ E_k,\ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < x/m\right\} \subseteq \{E_k,\ |X_{k+m,n}(c_n)| \geqslant \varepsilon\}. \quad (3)$$

Аналогично выводится

$$\left\{X_n(c_n) \geqslant 3x + 2\varepsilon,\ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < x/m\right\} \subseteq$$

$$\subseteq \left\{\overline{X}_{n-m}(c_n) \geqslant 2x + \varepsilon,\ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < x/m\right\},$$

следовательно

$$\left\{X_n(c_n) \geqslant 3x + 2\varepsilon,\ \max_{1\leqslant j\leqslant n} |Y_j(c_n)| < \frac{x}{m}\right\} =$$

$$= \left\{ X_n(c_n) \geqslant 3x + 2\varepsilon,\ \overline{X}_{n-m}(c_n) \geqslant 2x + \varepsilon,\ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < \frac{x}{m} \right\}. \qquad (4)$$

С помощью (3) и (4) получаем $\mathbf{P}\{|X_n(c_n)| \geqslant 3x + \varepsilon\} \leqslant$

$$\leqslant \mathbf{P}\left\{ X_n(c_n) \geqslant 3x + 2\varepsilon, \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < x/m \right\} + \mathbf{P}\left\{ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| \geqslant x/m \right\} =$$

$$= \mathbf{P}\left\{ X_n(c_n) \geqslant 3x + 2\varepsilon, \overline{X}_{n-m}(c_n) \geqslant 2x + \varepsilon, \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < x/m \right\} +$$

$$+\mathbf{P}\left\{ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| \geqslant x/m \right\} = \sum_{k=1}^{n-m} \mathbf{P}\left\{ X_n(c_n) \geqslant 3x + 2\varepsilon, E_k, \max_{1\leqslant j\leqslant n} |Y_j(c_n))| < x/m \right\} +$$

$$+\mathbf{P}\left\{ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| \geqslant x/m \right\}. \qquad (5)$$

Из соотношений (3), (5) и условия φ-перемешивания следует

$$\mathbf{P}\{X_n(c_n) \geqslant 3x+2\varepsilon\} \leqslant \sum_{k=1}^{n-m} \mathbf{P}\{E_k,\ |X_{k+m,n}(c_n)| \geqslant \varepsilon\} + \mathbf{P}\left\{ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| \geqslant x/m \right\} \leqslant$$

$$\leqslant \mathbf{P}\left\{ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| \geqslant x/m \right\} + \left(\max_{1\leqslant k\leqslant n} \mathbf{P}\{|X_k(c_n)| \geqslant \varepsilon\} + \varphi(m) \right) \sum_{k=1}^{n-m} \mathbf{P}\{E_k\} =$$

$$\leqslant \lambda \mathbf{P}\{\overline{X}_n(c_n) \geqslant 2x + \varepsilon\} + \mathbf{P}\left\{ \max_{1\leqslant j\leqslant n} |Y_j(c_n))| \geqslant x/m \right\}.$$

Из этого соотношения с помощью Леммы 1 выводим утверждение леммы. ■

Покажем, как с помощью леммы 2 можно получать оценки для моментов величин $X_n(c_n)$. Пусть $\mathbf{E}|Y_1(c_n))|^p < \infty$. Тогда в силу (1) $\mathbf{E}|X_n(c_n))|^p < \infty$, и если $\gamma > 0$ в формулировке леммы 2 таково, что $\dfrac{\gamma(3+2\varepsilon)^p}{1-\gamma} < 1$, то с помощью леммы 2 выводим

$$(3+2\varepsilon)^{-p}\mathbf{E}|X_n(c_n)|^p = p\int_0^\infty x^{p-1}\mathbf{P}\{|X_n(c_n)| \geqslant (3+2\varepsilon)x\}\,dx \leqslant$$

$$\leqslant 1 + p\int_1^\infty x^{p-1}\mathbf{P}\{|X_n(c_n)| \geqslant 3x + 2\varepsilon\}\,dx \leqslant 1+$$

$$+\frac{p\gamma}{1-\gamma}\int_0^\infty x^{p-1}\mathbf{P}\{|X_n(c_n)| \geqslant x\}\,dx + \frac{p}{1-\gamma}\int_0^\infty x^{p-1}\mathbf{P}\left\{ \max_{1\leqslant k\leqslant n} |Y_k(c_n)| \geqslant \frac{x}{m} \right\} dx =$$

$$= 1 + \frac{\gamma}{1-\gamma}\mathbf{E}|X_n(c_n)|^p + \frac{m^p}{(1-\gamma)}\mathbf{E}\max_{1\leqslant k\leqslant n} |Y_k(c_n)|^p.$$

Отсюда следует, что

$$\mathbf{E}|X_n(c_n)|^p \leqslant A + B\,\mathbf{E}\max_{1\leqslant k\leqslant n}|Y_k(c_n)|^p, \tag{6}$$

где A и B не зависят от n.

Пусть последовательность $\{\xi_n\}$ удовлетворяет условию равномерно сильного перемешивания $S_n = \sum\limits_{i=1}^{n}\xi_i$, $\mathbf{E}\xi_1 = 0$, $\sigma_n^2 = \mathbf{E}S_n^2 \to \infty$, $n \to \infty$. К последовательности $\{\xi_n\}$ применима центральная предельная теорема тогда и только тогда, когда последовательность $\{\sigma_n^{-2}S_n^2\}$ равномерно интегрируема. Существенный прогресс в предельных теоремах для последовательностей с φ-перемешиванием достигнут М. Пелиград в [4] на основе доказательства того, что равномерная интегрируемость $\{\sigma_n^{-2}S_n^2\}$ равносильна равномерной интегрируемости $\{\sigma_n^{-2}\max\limits_{1\leqslant i\leqslant n}\xi_i^2\}$.

Получим здесь аналогичный результат, где вместо сумм S_n участвуют функции $f(\xi_1, ..., \xi_n)$, удовлетворяющие условиям A .

Теорема 1. *Пусть функция f удовлетворяет условиям A, $E|Y_1(c_n))|^p < \infty$, а последовательность $\{c_n\}$ и $m > 0$ таковы,*

$$\max_{1\leqslant k\leqslant n}\mathbf{P}\{|X_k(c_n)| \geqslant \varepsilon\} + \varphi(m) \leqslant \gamma, \quad \frac{\gamma(3+2\varepsilon)^p}{1-\gamma} < 1.$$

Тогда последовательность $\{\overline{X}_n^p(c_n)\}$ равномерно интегрируема тогда и только тогда, когда равномерно интегрируема $\left\{\max\limits_{1\leqslant k\leqslant n}|Y_k(c_n))|^p\right\}$.

Доказательство. Пусть

$$\mathbf{E}\{|\xi|^p, |\xi| \geqslant N\} = -\int_N^\infty x^p\, d\mathbf{P}\{|\xi| \geqslant x\} = N^p\mathbf{P}\{|\xi| \geqslant N\} + p\int_N^\infty x^{p-1}\mathbf{P}\{|\xi| \geqslant x\}\, dx.$$

В силу леммы 2 при $N \geqslant 1$

$$\mathbf{E}\{|X_n(c_n)|^p, |X_n(c_n)| \geqslant (3+2\varepsilon)N\} \leqslant (3+2\varepsilon)^p N^p\mathbf{P}\{|X_n(c_n)| \geqslant N(3+2\varepsilon)\} +$$

$$+p(3+2\varepsilon)^p\int_N^\infty x^{p-1}\mathbf{P}\{|X_n(c_n)| \geqslant 3x+2\varepsilon\}\, dx \leqslant$$

$$\leqslant \frac{\gamma(3+2\varepsilon)^p}{1-\gamma}\mathbf{E}\{|X_n(c_n)|^p, |X_n(c_n)| \geqslant N\} +$$

$$+\frac{m^p(3+2\varepsilon)^p}{(1-\gamma)}\mathbf{E}\left\{\max_{1\leqslant k\leqslant n}|Y_k(c_n)|^p, |\max_{1\leqslant k\leqslant n}|Y_k(c_n)| \geqslant N/m\right\}. \tag{7}$$

Пусть последовательность $\left\{\max\limits_{1\leqslant k\leqslant n}|Y_k(c_n))|^p\right\}$ равномерно интегрируема, то есть

$$\lim_{N\to\infty}\sup_{n\geqslant 1}\mathbf{E}\left\{\max_{1\leqslant k\leqslant n}|Y_k(c_n)|^p, \max_{1\leqslant k\leqslant n}|Y_k(c_n)| \geqslant N\right\} = 0$$

и пусть

$$R = \lim_{N\to\infty} \sup_{n\geqslant 1} \mathbf{E}\left\{|X_n(c_n)|^p, |X_k(c_n)| \geqslant N\right\}.$$

Из равномерной интегрируемости $\left\{\max\limits_{1\leqslant k\leqslant n} |Y_k(c_n))|^p\right\}$ и (6) следует

$$\sup_{n\geqslant 1} \mathbf{E}|X_n(c_n)|^p \leqslant A + B \sup_{n\geqslant 1} \mathbf{E} \max_{1\leqslant k\leqslant n} |Y_k(c_n)|^p < \infty,$$

так что $0 \leqslant R < \infty$, а из (7) вытекает $R \leqslant \tau R$, где $\tau = \dfrac{\gamma(3+2\varepsilon)^p}{1-\gamma} < 1$, следовательно, $R = 0$ и последовательность $\{|X_n(c_n)|^p\}$ равномерно интегрируема.

Далее, из леммы 1 при $N \geqslant 1$ аналогично (7) выводим

$$\mathbf{E}\{\overline{X}_n^p(c_n), \overline{X}_n(c_n) \geqslant (2+\varepsilon)N\} \leqslant \frac{\gamma(2+\varepsilon)^p}{1-\gamma}\mathbf{E}\{|X_n(c_n)|^p, |X_n(c_n)| \geqslant N\} +$$

$$+\frac{\gamma(2+\varepsilon)^p m^p}{1-\gamma}\mathbf{E}\left\{\max_{1\leqslant k\leqslant n} |Y_k(c_n)|^p, |\max_{1\leqslant k\leqslant n} |Y_k(c_n)| \geqslant N/m\right\},$$

следовательно, из равномерной интегрируемости $\left\{\max\limits_{1\leqslant k\leqslant n} |Y_k(c_n))|^p\right\}$ и $\{|X_n(c_n)|^p\}$ следует равномерная интегрируемость $\{\overline{X}_n^p(c_n)\}$.

Наоборот, в силу (1) $\max\limits_{1\leqslant k\leqslant n} |Y_k(c_n))|^p \leqslant 2^p \overline{X}_n^p(c_n)$ и равномерная интегрируемость $\{\overline{X}_n^p(c_n)\}$ влечёт равномерную интегрируемость $\left\{\max\limits_{1\leqslant k\leqslant n} |Y_k(c_n))|^p\right\}$ очевидным образом. ■

Будем говорить, что для функции f выполнено условие A_4, если при любом $\lambda > 0$

$$f(\lambda x_1, ..., \lambda x_n) = \lambda f(x_1, ..., x_n).$$

В большинстве примеров, приводимых в [2], функции удовлетворяют условию A_4 (в том числе так называемые симметрические калибровочные функции).

Если функция f удовлетворяет условию A_4, то из оценок (6) получаем следующий результат, обобщающий теорему 1 из [2] и лемму 18.5.1 из [3].

Теорема 2. *Пусть функция f удовлетворяет условиям A_1–A_4 и пусть $0 < q < p$, $\mathbf{E}|\xi_1|^p < \infty$. Тогда*

$$\max_{1\leqslant k\leqslant n} \mathbf{E}|X_k|^p \leqslant A\left(\max_{1\leqslant k\leqslant n} \mathbf{E}|X_k|^q\right)^{p/q} + B\,\mathbf{E}\max_{1\leqslant k\leqslant n} |\xi_k|^p,$$

где A и B не зависят от n.

Доказательство. Пусть $0 < q < p$, а $\varepsilon > 0$, $m > 0$ и $\gamma > \varphi(m)$ удовлетворяют условиям теоремы 1. Обозначим $c_n^q = N \max\limits_{1\leqslant k\leqslant n} \mathbf{E}|X_k|^q$, где $N > 0$ таково, что

$$\varphi(m) + \max_{1\leqslant k\leqslant n} \mathbf{P}\{|X_k| \geqslant \varepsilon c_n\} \leqslant \varphi(m) + \frac{\max\limits_{1\leqslant k\leqslant n} |X_k|^q}{\varepsilon^q c_n^q} \leqslant \gamma, \quad \frac{\gamma(3+2\varepsilon)^p}{1-\gamma} < 1.$$

Так как $\{c_n\}$ неубывающая последовательность, то при $k \leqslant n$ и при любом $x > 0$ из леммы 2 следует

$$\mathbf{P}\{|X_k| \geqslant (3x + 2\varepsilon)c_n\} \leqslant \frac{\gamma}{1-\gamma}\mathbf{P}\{|X_k| \geqslant xc_n\} + \frac{1}{1-\gamma}\mathbf{P}\left\{\max_{1\leqslant j\leqslant n}|Y_k| \geqslant \frac{xc_n}{m}\right\},$$

откуда аналогично (6) выводим

$$\mathbf{E}|X_k|^p \leqslant Ac_n^p + B\mathbf{E}\max_{1\leqslant j\leqslant n}|\xi_j|^p,\ k \leqslant n, \tag{8}$$

где $A > 0$ и $B > 0$ не зависят от n. Из последнего соотношения следует утверждение теоремы. ■

Пусть $X_n = \sum\limits_{i=1}^{n}\xi_i,\ \mathbf{E}\xi_1 = 0,\ \sigma_n^2 = \mathbf{E}S_n^2 \to \infty,\ n \to \infty$. Тогда σ_n является правильно меняющейся последовательностью порядка $1/2$ [3, теорема 18.2.3], которая при $n \to \infty$ эквивалентна некоторой неубывающей последовательности [5, с. 26], так что $\max\limits_{1\leqslant k\leqslant n}\sigma_k \sim \sigma_n$. Далее, при $p > 2$

$$\mathbf{E}\max_{1\leqslant k\leqslant n}|\xi_k|^p \leqslant n\mathbf{E}|\xi_1|^p = o(\sigma_n^p),$$

и из теоремы 2 следует неравенство И. А. Ибрагимова (лемма 18.5.1 из [3]): $\mathbf{E}|S_n|^p \leqslant C\sigma_n^p$, где $C > 0$ не зависит от n.

Литература

1. Гринь А.Г. О моментах обобщённых сумм // Математические структуры и моделирование. 2015. №. 4(36). С. 23-28.
2. Гринь А.Г. О предельных теоремах для функций от независимых случайных величин // Математические структуры и моделирование. 2016. № 29. С. 4-12.
3. Ибрагимов И.А., Линник Ю.В. Независимые и стационарно связанные величины. М. : Наука, 1965. 524 с.
4. Peligrad M. An invariance principle for φ-mixing sequences // Ann. Probab. 1985. V. 13, N. 4. P. 1304–1313.
5. Сенета Е. Правильно меняющиеся функции. М. : Наука, 1985, 141 с.

ON THE MOMENTS OF SYMMETRIC FUNCTIONS OF DEPENDENT RANDOM VARIABLES

A.G. Grin'

Dr.Sc. (Phys.-Math.), Professor, e-mail: griniran@gmail.com

Dostoevsky Omsk State University

Abstract. Estimates for the moments and uniform integrability of a certain class of functions of random variables with uniformly strong mixing is obtained in this article.

Keywords: symmetric functions of random variables, uniformly strong mixing condition, uniform integrability.

Дата поступления в редакцию: 09.10.2016

Математические структуры и моделирование
2016. № 4(40). С. 24–38

УДК 530.12+524.8

АНАЛИЗ В КОСМИЧЕСКИХ РАССЛОЕНИЯХ НА ОСНОВЕ ГРУППЫ U(1,1): ОСНОВНЫЕ ТАБЛИЦЫ ИНФИНИТЕЗИМАЛЬНОГО SU(2,2)-ДЕЙСТВИЯ

А.В. Левичев[1]
профессор, д.ф.-м.н., с.н.с., e-mail: alevichev@gmail.com
А.Ю. Пальянов[2,3]
к.ф.-м.н., с.н.с. , e-mail: palyanov@iis.nsk.su

[1]Институт математики СО РАН им. С.Л. Соболева
[2]Институт систем информатики СО РАН им. А.П. Ершова
[3]Новосибирский государственный университет

Аннотация. Хронометрическая теория Сигала исходит из пространства-времени $\mathbf{D}$, которое может быть представлено как группа Ли с причинной структурой, задаваемой инвариантной лоренцевой формой на алгебре Ли $u(2)$. Аналогично пространство-время $\mathbf{F}$ представлено группой Ли с причинной структурой, задаваемой инвариантной лоренцевой формой на алгебре Ли $u(1,1)$. Группы Ли G, G_F вводятся как представления $SU(2,2)$, связанные сопряжением конкретной матрицей W из $Gl(4)$. Дробно-линейное действие G на $\mathbf{D}$ глобально и конформно; оно играет важную роль в анализе пространственно-временных расслоений, основанном на параллелизующей группе $U(2)$: этот анализ проведён Панейтцем и Сигалом в 1980-х гг. Дробно-линейное действие G_F на $\mathbf{F}$ (введённое в 2000-х гг. первым автором) тоже конформно. В статье показано, что (несмотря на имеющиеся сингулярности этого действия) группа $U(1,1)$ может быть выбрана в качестве параллелизующей. Приводятся методы, применением которых нами получены таблицы (аналогичные «таблицам Панейтца-Сигала»), необходимые для (предстоящего) анализа пространственно-временных расслоений на основе параллелизующей группы $U(1,1)$.

Ключевые слова: параллелизации расслоений над пространством-временем, космос Сигала, действия конформной группы $SU(2,2)$ на $U(2)$ и на $U(1,1)$, DLF-теория.

1. Введение, мотивация и основные обозначения

Группы Ли $U(2)$ и $U(1,1)$ являются одними из основных объектов, рассматриваемых в работе.

В статье установлено, что $U(1,1)$ может быть выбрана в качестве параллелизующей группы. Ниже (в Секциях 2, 3) приводятся методы, применением которых нами получены таблицы (аналогичные «таблицам Панейтца-Сигала»),

необходимые для (предстоящего) анализа пространственно-временных расслоений на основе параллелизующей группы $U(1,1)$. Одним из результатов такого анализа должна стать классификация частиц, «живущих» в пространстве-времени $\mathbf{F}$.

Под $U(2)$ понимается совокупность всех два на два матриц Z (с комплексными, вообще говоря, элементами), удовлетворяющих соотношению

$$ZZ^* = \mathbf{1}.$$

Здесь и далее $\mathbf{1}$ — единичная матрица, а * означает комплексное сопряжение и транспонирование. Напомним, что группа $U(2)$ не является прямым произведением своей центральной подгруппы с подгруппой $SU(2)$. Двукратное накрытие $\mathbf{D}^{(2)}$ для $U(2)$, состоящее из всех пар вида $(p, \mathbf{u})$, является прямым произведением (групп) S^1 и S^3 (здесь S^3 представлена группой $SU(2)$, т.е. $\mathbf{u}$ — это соответствующая матрица, а модуль комплексого числа p равен 1). Накрывающее отображение переводит $(p, \mathbf{u})$ в матрицу $p\mathbf{u}$ из $U(2)$. Подразумевается, что на $\mathbf{D}^{(2)}$ введена лоренцева метрика

$$(dt)^2 - (du)^2. \tag{1.1}$$

Здесь t — параметр на S^1, а $(du)^2$ — стандартная риманова метрика на S^3.

Аналогично, под $U(1,1)$ понимается совокупность всех два на два матриц U, удовлетворяющих соотношению

$$UsU^* = s.$$

Здесь s — диагональная матрица с элементами $1, -1$ по главной диагонали.

Двукратное накрытие $\mathbf{F}^{(2)}$ для $U(1,1)$, состоящее из всех пар вида (q, V), является прямым произведением (групп) S^1 и $SU(1,1)$; т.е. V — это матрица из $SU(1,1)$, а модуль комплексого числа q равен 1. Накрывающее отображение переводит (q, V) в матрицу qV из $U(1,1)$. Дальнейшие детали о группах $\mathbf{D}^{(2)}, U(2), \mathbf{F}^{(2)}, U(1,1)$ и метриках на них приведены в Приложении А.

Нередко две из этих групп (в частности, когда они снабжены двусторонне-инвариантными метриками лоренцевой сигнатуры — см. [6]) обозначаются $\mathbf{D} = U(2)$ и $\mathbf{F} = U(1,1)$.

Отметим, что хронометрическая теория Сигала (см. [9]) основана на пространстве-времени $\mathbf{D}$. Так как DLF–теория исходит сразу из трёх миров ($\mathbf{D}$, $\mathbf{L}$ и $\mathbf{F}$), то её можно считать обобщением теории Сигала (DLF–теория представлена в [6], в то время как некоторые её исходные положения были намечены уже в [2] (сс. 1302-1303).

Прежде чем приступить к формулировке результатов статьи, напомним, что параллелизация (расслоений над пространством-временем — см. определения и теоремы существования в [9], Секция IV) является важным математическим методом современной теоретической физики. Именно каждому «объекту» сопоставляется его состояние (часто называемое волновой функцией, но этот последний термин целесообразнее употреблять в более специализированной ситуации, а именно — ПОСЛЕ параллелизации). Если в качестве объекта рассматривается элементарная частица («живущая» в некотором мире событий $\mathbf{W}$), то совокупность её возможных состояний является вполне определённым подпространством множества сечений (бесконечно дифференцируемых, суммируемых

с квадратом и т.д. — в данном случае нет необходимости уточнять эти детали) некоторого векторного расслоения с базой **W**. На этой стадии состояния ещё не принимают числовых (для скалярной частицы) или C^k-значений ($k > 1$, для частиц ненулевого спина). Необходим переход от (абстрактных) сечений к параллелизованным сечениям (т.е. к волновым функциям). Затем вводится структура гильбертова пространства и т.д. (нет необходимости детализировать эти этапы в данной статье). Процедура параллелизации определяется выбором параллелизующей (четырехмерной) подгруппы N в группе G, где G — это группа симметрий мира **W** (в контексте данной статьи G — это (конформная) группа $SU(2,2)$, см. ниже). Начиная с такого этапа, N как бы заменяет исходный мир событий **W** (типичная ситуация состоит в том, что группа N является конечно-листным накрытием мира **W**).

Элементарные частицы и их взаимодействия моделируются в терминах индуцированных представлений группы G. В рамках стандартной теоретической физики G — это десятимерная группа Пуанкаре, а в качестве параллелизующей подгруппы практически всегда (зачастую — «по умолчанию») выбиралась векторная группа мира Минковского M. Проводилось индуцирование по подгруппе Лоренца (такой подход был заявлен знаменитой статьёй Вигнера [15]). Проблемы выбора параллелизации не возникало ещё и потому, что, фактически, рассмотрение начиналось с параллелизованных сечений (т.е. с волновых функций).

В работах школы Сигала (см. [13]) чаще всего использовались параллелизации, основанные на группе $U(2)$. Иногда они сравнивались с плоской параллелизацией (определяемой векторной группой мира M). На с. 170 известной монографии [1] роль выбора параллелизации обсуждается с точки зрения вопросов, возникающих в квантовой теории поля.

В [6] было предложено рассмотреть другие (кроме **D** и M) параллелизующие группы. В связи с этим важен результат статьи [3] (см. Теорему 1, ниже), сформулированный в терминах коммутативной $\mathbf{D} - \mathbf{F}$ диаграммы. На его основе делается вывод, что (несмотря на наличие сингулярностей) осуществима как сама **F**-параллелизация, так и её (каноническое) сравнение с **D**-параллелизацией. Термин 'сравнение параллелизаций' был введён в [9], там же были рассмотрены и некоторые примеры. Дело в том, что действие (той или иной) подгруппы группы G может быть реализовано сложно или просто — в зависимости от выбора параллелизации. Одним из основных следствий применения **D**-параллелизации явилась классификация хронометрических частиц спина 1/2. Таковых оказалось четыре: протон, электрон и два вида нейтрино (см. [12] и [5]). Поэтому вопрос отыскания соответствующей классификации на основе **F**-параллелизации представляется весьма интересным. Так как пространства-времена **F** и **D** связаны конформным (не сводящимся к изометрии) преобразованием, то заранее неизвестно набор каких именно частиц будет получен при использовании **F**–параллелизации. Необходимость рассмотрения **F**-параллелизации обеспечена ещё и тем фактом, что среди всех вещественных 4-мерных алгебр Ли лишь $u(2)$ и $u(1,1)$ являются редуктивными.

Каждая из групп Ли G, G_F, изоморфна $SU(2,2)$. Именно G состоит из всех

4 на 4 матриц g (с определителем 1), для которых выполняется

$$g^* S g = S. \tag{1.2}$$

Здесь $S = diag\{1, 1, -1, -1\}$ диагональная матрица.

Через W обозначается 4 на 4 матрица

$$W = \begin{bmatrix} P & Q \\ Q & P \end{bmatrix}, \tag{1.3}$$

образованная 2 на 2 блоками

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, Q = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{1.4}$$

Ясно, что

$$detW = -1, W^2 = \mathbf{1}, P^2 = P, Q^2 = Q, PQ = QP = 0. \tag{1.5}$$

Сопрягая матрицу S матрицей W, получаем

$$\tilde{S} = diag\{1, -1, -1, 1\},$$

которая задаёт другую 'копию' (обозначаем её G_F) группы $SU(2,2)$. Именно G_F состоит из всех 4 на 4 матриц $\tilde{g}$ (с определителем 1), для которых выполняется

$$\tilde{g}^* \tilde{S} \tilde{g} = \tilde{S}. \tag{1.6}$$

Соответствие

$$\tilde{g} = W g W \tag{1.7}$$

является изоморфизмом групп Ли G, G_F.

Каждая g из G является 4 на 4 матрицей, задаваемой 2 на 2 блоками A, B, C, D:

$$g = \begin{bmatrix} A & B \\ C & D \end{bmatrix}. \tag{1.8a}$$

Аналогично каждая

$$\tilde{g} = \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix}. \tag{1.8b}$$

Известно, что дробно-линейное действие

$$g(Z) = (AZ + B)(CZ + D)^{-1} \tag{1.9}$$

группы G (см. [10], с.35) определено на всей $\mathbf{D} = U(2)$. Дробно-линейное (определённое лишь локально) действие

$$\tilde{g}(U) = (\tilde{A}U + \tilde{B})(\tilde{C}U + \tilde{D})^{-1} \tag{1.10}$$

группы G_F на $\mathbf{F} = U(1,1)$ было введено в [6].

Замечание 1. Так как в Секции 3 мы рассматриваем лишь действие (1.10), то (для упрощения обозначений) знак тильды там будет опущен. Тем самым общий элемент группы G_F будет обозначаться через g (с блоками A, B, C, D).

2. Коммутативная *D–F* диаграмма и смежная проблематика

Пусть дана 2 на 2 матрица Y. Через $W(Y)$ обозначаем матрицу $(PY+Q)(QY+P)^{-1}$, если она определена. Задаём вложение (многообразия) $\mathbf{F}$ в $\mathbf{D}$ формулой

$$Z = W(U) = (PU+Q)(QU+P)^{-1}. \tag{2.1}$$

Нетрудно проверить, что (2.1) определено для любой U из $\mathbf{F}$. Отображение W конформно, но в данной статье это свойство не используется.

Формула (2.1) — это частный случай (см. [8], с. 32) *формулы Свидерского*. Легко проверяется, что обратное отображение

$$U = W(Z) = (PZ+Q)(QZ+P)^{-1} \tag{2.2}$$

определено для тех (и только тех) Z, которые не принадлежат тору $\mathbf{T}$, состоящему из всех матриц K в $\mathbf{D}$ вида

$$K = \begin{bmatrix} 0 & p \\ q & 0 \end{bmatrix}. \tag{2.3}$$

Здесь p, q могут быть произвольными комплексными числами с модулем 1.

Следующее важное утверждение (в нём используются обозначения (1.9), (1.10)) доказано в [3]:

Теорема 1 ($\mathbf{D}-\mathbf{F}$ диаграмма). Если $\tilde{g}(U)$ определено, то

$$g(W(U)) = W(\tilde{g}(U)). \tag{2.4}$$

Замечание 2. В [3] не было исследовано, когда (т.е. при каких U из $\mathbf{F}$) правая часть (2.4) определена. Конечно же (см. нашу формулу 1.10), она определена тогда и только тогда, когда определитель матрицы $\tilde{C}U + \tilde{D}$ не равен нулю. Однако это условие оказывается менее удобным для применения, нежели приводимое ниже (в Теореме 2).

В [7] доказано следующее утверждение:

Теорема 2. Пусть $\tilde{g}$ принадлежит G_F, а U — матрица в $\mathbf{F}$. Матрица $\tilde{g}(U)$ определена тогда и только тогда, когда $g(W(U))$ не принадлежит тору $\mathbf{T}$.

Замечание 3. На основании Теоремы 2 можно сказать, что сингулярности $SU(2,2)$-действия в $\mathbf{F}$ являются *ручными*.

Теоремы 1, 2, с учётом полученных ранее результатов (см. [9], [6], [8]), дают основание утверждать, что анализ пространственно-временных расслоений на основе параллелизующей группы $U(1,1)$ математически возможен. Его

осуществление (и сравнение с результатами, основанными на параллелизующей группе $U(2)$) представляет несомненный интерес. По аналогии с [9], Глава V, этот (новый) анализ следует начать с рассмотрения скалярных расслоений. Вместо группы изометрий $\mathbf{K}$ (с алгеброй Ли $R \oplus su(2) \oplus su(2)$) пространства-времени $\mathbf{D}$ нужно будет взять группу изометрий $\mathbf{K}_F$ (с алгеброй Ли $R \oplus su(1,1) \oplus su(1,1)$) мира $\mathbf{F}$. При построении базиса скалярных представлений вместо 'левой' и 'правой' алгебр Ли $su(2)$ (см. [9], Секция 5.4) будут выбраны 'левая' и 'правая' алгебры Ли $su(1.1)$. Так как речь идёт о представлениях над полем комплексных чисел, то сравнение двух 'картин' ('компактной' — на основе $U(2)$, и 'некомпактной' — на основе $U(1,1)$) будет вполне осуществимо. Здесь имеется в виду хорошо известный 'унитарный трюк'. В целом, упомянутые вопросы интересны как математически (ковариантность волновых операторов, инвариантные формы в пространствах индуцированных представлений, классы специальных функций и др.), так и с точки зрения приложений в физике: см., например, [5, сс. 88-89], где предлагается отождествить инвариантное подпространство т.н. *спэннорного* [12] представления с совокупностью состояний протона (что объяснило бы стабильность протона).

3. Основные результаты: отыскание Таблиц I, III и IV

Последовательность, в которой мы излагаем необходимые сведения/результаты (для предстоящего использования $\mathbf{F} = U(1,1)$ в качестве параллелизующей группы), аналогична той, которая имеется в работах [11] и [9] для случая $\mathbf{D} = U(2)$. В частности, сохраняется нумерация таблиц (в упомянутых двух статьях приведено десять таблиц): каждой $\mathbf{D}$-таблице соответствует её $\mathbf{F}$-аналог. Важно иметь в виду, что все как чисто математические результаты школы Сигала, так и их многочисленные приложения в теоретической физике были получены на основе применения информации, приведённой в этих $\mathbf{D}$-таблицах. Поэтому получение $\mathbf{F}$-аналогов этих таблиц представляется важным (и совершенно необходимым для реализации нашей программы, изложенной в конце предыдущей секции). На данном этапе нами получены Таблицы I, III и IV (см. Приложение Б). Отметим, что Таблицу II (мы её не приводим) легко составить на основе Таблицы I (см. её второй столбец) и Таблицы III. Мы не приводим $\mathbf{D}$-таблицы, так как они доступны в сети [11], [9]. Кроме того, они имеются в [4]. Матрицы $\mathbf{L_{ij}}$ (в их блочной форме) приведены в третьем столбце нашей Таблицы I, они являются базисными элементами алгебры Ли $su(2,2)$. В приведённых там блоках через s, b_0, b_1, b_2, b_3 обозначены следующие 2 на 2 матрицы:

$$s = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, b_1 = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, b_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, b_3 = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, b_0 = \begin{bmatrix} i & 0 \\ 0 & i \end{bmatrix}.$$

Отметим, что каждая $\mathbf{L_{ij}}$ получена сопряжением (с помощью матрицы W — см. наше (1.2) выше) из соответствующего элемента $\mathbf{D}$-таблицы I. Их коммутационные соотношения (для любого из случаев $\mathbf{D}$, $\mathbf{F}$) таковы:

$$[\mathbf{L_{im}}, \mathbf{L_{mk}}] = -e_m \mathbf{L_{ik}},$$

где под $(e_{-1}, e_0, e_1, e_2, e_3, e_4)$ понимается набор $(1, 1, -1, -1, -1, -1)$. Индексы i, j принимают значения $-1, 0, 1, 2, 3, 4$ с $i < j$. Через L_{ij} будут обозначаться соответствующие векторные поля на $U(1,1)$. Они задаются действием (1.10). Считаем, что всегда $\mathbf{L_{ij}} = -\mathbf{L_{ji}}$, откуда следуют соотношения $L_{ij} = -L_{ji}$.

Лоренцева метрика на $\mathbf{D} = U(2)$ задаётся следующим условием: лево-инвариантные векторные поля

$$X_0 = L_{-10}, X_1 = L_{14} - L_{23}, X_2 = L_{24} - L_{31}, X_3 = L_{34} - L_{12} \tag{3.1}$$

образуют ортонормированный репер. Именно скалярный квадрат вектора X_0 равен 1, а каждого из оставшихся трёх векторов — минус единице. Известно, что это условие согласовано с выбором (1.1) метрики на $\mathbf{D}^{(2)}$, т.е. накрывающее отображение является изометрией. Векторные поля (3.1) использованы в качестве базисных во втором столбце $\mathbf{D}$-таблицы I. Их аналогом для случая $\mathbf{F}$ являются поля

$$H_0 = L_{10} - L_{12}, H_1 = L_{-11} - L_{02}, H_2 = L_{-12} + L_{01}, H_3 = L_{34}, \tag{3.2}$$

фигурирующие во втором столбце нашей Таблицы I.

Лоренцева метрика (как на $\mathbf{F}^{(2)}$, так и на $\mathbf{F}$) задаётся условием ортонормированности полей (3.2). Именно скалярный квадрат вектора H_0 равен 1, а каждого из оставшихся трёх векторов — минус единице. Как известно, лево-инвариантные векторные поля генерируют правые сдвиги. Именно полям (3.2) отвечают сдвиги на

$$\begin{bmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{bmatrix}, \begin{bmatrix} C & iS \\ -iS & C \end{bmatrix}, \begin{bmatrix} C & S \\ S & C \end{bmatrix}, \begin{bmatrix} e^{it} & 0 \\ 0 & e^{it} \end{bmatrix} \tag{3.3}$$

соответственно. В (3.3) под C, S понимаются гиперболические косинус и синус вещественного аргумента t.

Базис право-инвариантных векторных полей на $\mathbf{F}$ вводится так:

$$J_0 = L_{-10} + L_{12}, J_1 = L_{-11} + L_{02}, J_2 = L_{-12} - L_{01}, J_3 = L_{34}. \tag{3.4}$$

Он фигурирует в нашей Таблице IV, в которой приведены действия рассматриваемых векторных полей на (введённые в Приложении А) переменные $v_{-1}, v_0, v_1, v_2, v_3, v_4$.

Вообще, Таблицу IV (с учётом данных столбца 3 Таблицы I) можно считать исходной (для дальнейшего нахождения инфинитезимального действия группы G_F). В оставшейся части данной секции мы приводим соответствующие сведения и аргументацию (см. Теоремы 3 и 4, Леммы 1– 5 и сопутствующие им обозначения). Эта аргументация основана на дробно-линейном действии, поэтому она применима как в $\mathbf{D}$-, так и в $\mathbf{F}$-случае. Отметим, что ни в [11], ни в [9] подобной аргументации не было приведено.

Теорема 3. Векторные поля L_{ij}, H_m, J_k задаются Таблицей IV.

Справедливость этой теоремы будет установлена на основе нескольких вспомогательных утверждений.

В дальнейшем, если m – зависящая от параметра t величина на $\mathbf{F}^{(2)}$ (или на $\mathbf{F}$), то под $\dot{m}$ понимается её производная по t, вычисленная при t=0. Если дифференцируется выражение, заключённое в скобки, то результат записывается в виде: $(...)\dot{}$. Под L (см. нашу Таблицу IV) понимается векторное поле (т.е. линейный дифференциальный оператор первого порядка), соответствующее матрице $\mathbf{L}$ (как правило, $\mathbf{L}$ — это одна из матриц, приведённых в последнем столбце Таблицы I).

Замечание 4. Через A, B, C, D обозначаются блоки матрицы $g = exp(t\mathbf{L})$. Тем самым мы упрощаем обозначения Секции 1 (см. Замечание 1 в её конце): в формуле (1.10) опускаем тильду. Такое упрощение уместно, так как действие на $\mathbf{D}$ в данной секции не рассматривается.

Лемма 1. Действие оператора L на переменные v_{-1}, v_0 может быть найдено из уравнения

$$2(\dot{v}_{-1} + i\dot{v}_0)(v_{-1} + iv_0) + (v_{-1} + iv_0)^2(det(CU + D))\dot{} = (det(AU + B))\dot{}. \qquad (3.5)$$

Доказательство. Если $\hat{U} = g(U)$ в (1.10) определено, то это соотношение эквивалентно

$$\hat{U}(CU + D) = AU + B. \qquad (3.6)$$

Из (3.6) следует

$$det(\hat{U})det(CU + D) = det(AU + B). \qquad (3.7)$$

Напомним (см. Приложение А), что $det(\hat{U}) = (\hat{v}_{-1} + i\hat{v}_0)^2$ и $detU = (v_{-1} + iv_0)^2$. Для того чтобы найти действие оператора L на переменные v_{-1}, v_0, достаточно продифференцировать обе части соотношения (3.7) по t и положить $t = 0$. Получаем искомое (3.5). Лемма 1 доказана.

Лемма 2. Действие оператора L на переменные v_1, v_2, v_3, v_4 может быть найдено из уравнения

$$(\dot{v}_{-1} + i\dot{v}_0)V + (v_{-1} + iv_0)\dot{V} + U(CU + D)\dot{} = (AU + B)\dot{}. \qquad (3.8)$$

Доказательство. Напомним (см. Приложение А), что через V обозначена матрица из $SU(1,1)$ в разложении $U = (v_{-1} + iv_0)V$. Аналогично, $\hat{U} = (\hat{v}_{-1} + i\hat{v}_0)\hat{V}$. Чтобы получить соотношение (3.8), достаточно продифференцировать (3.6) по t и положить $t = 0$. Не забываем, что при $t = 0$ выполнены соотношения $A = D = \mathbf{1}, B = C = \mathbf{0}$. Согласно (3.5) множитель $(\dot{v}_{-1} + i\dot{v}_0)$ в (3.8) уже подсчитан. Лемма 2 доказана.

Обозначим через a, b, c, d производные от матриц A, B, C, D по t, вычисленные при $t = 0$. Матрицы a, b, c, d приведены в третьем столбце Таблицы I для каждой $\mathbf{L} = \mathbf{L_{ij}}$. Через trM обозначается след матрицы M, а элементы 2 на 2 матрицы M нумеруются так:

$$M = \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix}.$$

Лемма 3. Выполняются (3.9) и (3.10):

$$(det(CU+D))\dot{} = trd + tr(cU), \tag{3.9}$$

$$(det(AU+B))\dot{} = (tra)detU + tr(b\tilde{U}), \tag{3.10}$$

где через $\tilde{U}$ обозначена матрица

$$\tilde{U} = \begin{bmatrix} U_4 & -U_3 \\ -U_2 & U_1 \end{bmatrix}.$$

Доказательство: непосредственный подсчёт.

Так как в столбце 3 таблицы I два блока всегда равны нулю, то при составлении таблицы IV уместно использовать следующие два утверждения (Леммы 4 и 5).

Лемма 4. Если $a = d = 0$, то (3.5) упрощается до

$$2(\dot{v}_{-1} + i\dot{v}_0) + (v_{-1} + iv_0)^2 tr(cV) = tr(b\tilde{V}). \tag{3.11}$$

Если же $b = c = 0$, то (3.5) эквивалентно

$$2(\dot{v}_{-1} + i\dot{v}_0)(v_{-1} - iv_0) + trd = tra. \tag{3.12}$$

Доказательство: применение формул (3.9), (3.10) в (3.5).

Лемма 5. Если $a = d = 0$, то (3.8) упрощается до

$$(\dot{v}_{-1} + i\dot{v}_0)(v_{-1} - iv_0)V + \dot{V} + (v_{-1} + iv_0)VcV = (v_{-1} - iv_0)b. \tag{3.13}$$

Если же $b = c = 0$, то (3.8) эквивалентно

$$(\dot{v}_{-1} + i\dot{v}_0)(v_{-1} - iv_0)V + \dot{V} + Vd = aV. \tag{3.14}$$

Доказательство: применение формул (3.9), (3.10) в (3.8).

Замечание 5. Применение формул (3.11)–(3.14) завершает доказательство Теоремы 3.

Теорема 4. Разложения векторных полей L_{ij} по базисным полям H_m задаются (вторым) столбцом Таблицы I.

Доказательство: проверка правильности этих разложений легко осуществима на основе Таблицы IV.

Оставшаяся часть основного текста статьи посвящена получению Таблицы III. В ней используются координаты p, q, r, t: они введены соотношениями (A8) Приложения A.

Ограничимся доказательством справедливости разложения $J_0 = \partial_r + \partial_q$, остальные разложения отыскиваются аналогично.

Так как $J_0 = L_{-10} + L_{12}$, то (см. нашу Таблицу IV и Приложение A) значение векторного поля J_0 в точке (q, V) многообразия $\mathbf{F}^{(2)}$ можно отождествить с элементом $(0, M)$ линейного пространства $E^2 \oplus E^4$. Здесь

$$M = \begin{bmatrix} -v_3 + iv_4 & -v_2 + iv_1 \\ -v_2 - iv_4 & -v_3 - iv_4 \end{bmatrix}.$$

Применяя соотношения (A8) Приложения A, получаем разложение $J_0 = \partial_r + \partial_q$.

4. Приложение A: Соглашения о группах U(2), U(1,1) и их 2-накрытиях

Помимо введённого в Секции 1 2-накрытия $\mathbf{F}^{(2)}$ группы $U(1,1)$ определим соответствующее разложение $U = qV$ для матрицы U из $\mathbf{F}$. Рассмотрим прямую сумму $E^6 = E^2 \oplus E^4$ двух евклидовых пространств: E^2 с прямоугольными координатами v_{-1}, v_0; E^4 с прямоугольными координатами v_1, v_2, v_3, v_4. Каждая точка в $\mathbf{F}^{(2)}$ — это такой набор $(v_{-1}, v_0, v_1, v_2, v_3, v_4)$, что выполнены условия A1, A2:

$$v_{-1}{}^2 + v_0{}^2 = 1, \tag{A1}$$

$$v_3{}^2 + v_4{}^2 - v_1{}^2 - v_2{}^2 = 1. \tag{A2}$$

Ясно, что $\mathbf{F}^{(2)}$ есть прямое произведение окружности S^1 с элементами $q = v_{-1} + iv_0$ и подгруппы $SU(1,1)$. Известно, что $SU(1,1)$ имеет топологию прямого произведения S^1 и R^2.

Матрица V из $SU(1,1)$ задаётся так:

$$V = \begin{bmatrix} v_4 + iv_3 & v_1 + iv_2 \\ v_1 - iv_2 & v_4 - iv_3 \end{bmatrix}. \tag{A3}$$

Накрывающее отображение из $\mathbf{F}^{(2)}$ в $U(1,1)$ переводит пару $(v_{-1} + iv_0, V)$ в матрицу $(v_{-1} + iv_0)V$, элемент группы $U(1,1)$:

$$U = (v_{-1} + iv_0)V. \tag{A4}$$

Если дана матрица U в $U(1,1)$, то сомножители $(v_{-1}+iv_0)$ и V определяются с точностью до знака, так как каждому U соответствуют два накрывающих элемента в $\mathbf{F}^{(2)}$: $(v_{-1} + iv_0, V)$ и $(-v_{-1} - iv_0, -V)$. В этом смысле координаты $v_{-1}, v_0, v_1, v_2, v_3, v_4$ можно использовать для параметризации как точек в $\mathbf{F}^{(2)}$, так и в $U(1,1)$.

В [14] была доказана единственность $SU(2,2)$-действия в $\mathbf{D}^{(2)}$, накрывающего дробно-линейное $SU(2,2)$-действие (1.9) в $\mathbf{D}$. $SU(2,2)$-действие в $\mathbf{F}^{(2)}$, накрывающее дробно-линейное $SU(2,2)$-действие (1.10) в $\mathbf{F}$, введём следующим образом. Сначала построим аналог $W^{(2)}$ отображения (2.1). Каждой паре (q, V) из $\mathbf{F}^{(2)}$ сопоставляем такую пару $(p, \mathbf{u})$ из $\mathbf{D}^{(2)}$, что $p = (v_4 + iv_3)/(v_3{}^2 + v_4{}^2)^{1/2}$. Отметим, что положительность квадратного корня в знаменателе обеспечена принадлежностью матрицы V группе $SU(1,1)$. Требуем коммутативности диаграммы

$$\begin{array}{ccc} (p, \mathbf{u}) & \longrightarrow & p\mathbf{u} \\ \uparrow & & \downarrow \\ (q, V) & \longrightarrow & qV = W(p\mathbf{u}) \end{array} \tag{A5}$$

Тем самым, матрица $\mathbf{u}$ из $SU(2)$ определяется однозначно. Напомним, что вторая вертикальная стрелка в (A5) — это биекция (2.2) между (частью) $\mathbf{D}$ и (всем) $\mathbf{F}$.

Теперь задаём $SU(2,2)$-действие в $\mathbf{F}^{(2)}$ требованием коммутативности диаграммы

$$\begin{array}{ccc} \mathbf{F}^{(2)} & \longrightarrow & \mathbf{F}^{(2)} \\ \downarrow & & \downarrow \\ \mathbf{D}^{(2)} & \longrightarrow & \mathbf{D}^{(2)} \end{array} \tag{A6}$$

В (A6) нижний уровень — это действие в $\mathbf{D}^{(2)}$, а каждая из вертикальных стрелок — это только что введённое отображение $W^{(2)}$.

Нетрудно проверить, что для таким образом введённого $SU(2,2)$-действия в $\mathbf{F}^{(2)}$ следующая диаграмма коммутативна:

$$\begin{array}{ccc} \mathbf{F}^{(2)} & \longrightarrow & \mathbf{F}^{(2)} \\ \downarrow & & \downarrow \\ \mathbf{F} & \longrightarrow & \mathbf{F} \end{array} \tag{A7}$$

Другими словами, $SU(2,2)$-действие в $\mathbf{F}^{(2)}$ накрывает таковое в $\mathbf{F}$. Отметим, что действие в $\mathbf{F}^{(2)}$ не является глобально определённым (так как действие в $\mathbf{F}$ не является таковым).

$\mathbf{F}$-аналогом полярных $\mathbf{D}$-координат (фигурирующих в $\mathbf{D}$-таблицах II, III) являются t, p, q, r:

$v_{-1} = \cos t, v_0 = \sin t, v_1 = C_q \sinh p,$

$v_2 = S_q \sinh p, v_3 = S_r \cosh p, v_4 = C_r \cosh p.$ (A8)

Под C_q, S_q, S_r, C_r в (A8) понимаются косинус и синус соответствующего аргумента. Обозначения $\partial_t, \partial_p, \partial_q, \partial_r$ соответствующих векторных полей на $\mathbf{F}$ используются в Таблице III.

5. Приложение Б: Таблицы I, III и IV

Таблица I		
Символ генератора	Векторное поле на F как линейная комбинация H_j	Матрица генератора
L_{-10}	$({v_3}^2+{v_4}^2)H_0+(v_1v_4+v_2v_3)H_1+(v_1v_3-v_2v_4)H_2$	$0.5\cdot\begin{pmatrix} b_3 & 0 \\ 0 & -b_3 \end{pmatrix}$
L_{-11}	$(v_1v_4-v_2v_3)H_0+({v_4}^2-{v_2}^2)H_1+(v_3v_4-v_1v_2)H_2$	$0.5\cdot\begin{pmatrix} sb_1 & 0 \\ 0 & -sb_1 \end{pmatrix}$
L_{-12}	$-(v_1v_3+v_2v_4)H_0-(v_1v_2+v_3v_4)H_1+({v_4}^2-{v_1}^2)H_2$	$0.5\cdot\begin{pmatrix} sb_2 & 0 \\ 0 & -sb_2 \end{pmatrix}$
L_{-13}	$-v_0v_3H_0-v_0v_2H_1-v_0v_1H_2+v_{-1}v_4H_3$	$0.5\cdot\begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}$
L_{-14}	$-v_{-1}v_3H_0-v_{-1}v_2H_1-v_{-1}v_1H_2-v_0v_4H_3$	$0.5\cdot\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
L_{01}	$(v_1v_3+v_2v_4)H_0+(v_1v_2+v_3v_4)H_1+({v_3}^2-{v_2}^2)H_2$	$0.5\cdot\begin{pmatrix} -sb_2 & 0 \\ 0 & -sb_2 \end{pmatrix}$
L_{02}	$(v_1v_4H_0+({v_1}^2-{v_3}^2)H_1)+(v_3v_4-v_1v_2)H_2$	$0.5\cdot\begin{pmatrix} sb_1 & 0 \\ 0 & sb_1 \end{pmatrix}$
L_{03}	$v_0v_4H_0+v_0v_1H_1-v_0v_2H_2+v_{-1}v_3H_3$	$0.5\cdot\begin{pmatrix} 0 & -s \\ -s & 0 \end{pmatrix}$
L_{12}	$({v_1}^2+{v_2}^2)H_0+(v_1v_4+v_2v_3)H_1+(v_1v_3-v_2v_4)H_2$	$0.5\cdot\begin{pmatrix} b_3 & 0 \\ 0 & b_3 \end{pmatrix}$
L_{23}	$-v_0v_2H_0-v_0v_3H_1+v_0v_4H_2-v_{-1}v_1H_3$	$0.5\cdot\begin{pmatrix} 0 & b_1 \\ b_1 & 0 \end{pmatrix}$
L_{31}	$-v_0v_1H_0-v_0v_4H_1-v_0v_3H_2+v_{-1}v_2H_3$	$0.5\cdot\begin{pmatrix} 0 & b_2 \\ b_2 & 0 \end{pmatrix}$
L_{04}	$v_{-1}v_4H_0+v_{-1}v_1H_1-v_{-1}v_2H_2-v_0v_3H_3$	$0.5\cdot\begin{pmatrix} 0 & b_3 \\ -b_3 & 0 \end{pmatrix}$
L_{14}	$v_{-1}v_1H_0+v_{-1}v_4H_1+v_{-1}v_3H_2+v_0v_2H_3$	$0.5\cdot\begin{pmatrix} 0 & sb_1 \\ -sb_1 & 0 \end{pmatrix}$
L_{24}	$v_{-1}v_2H_0-v_{-1}v_3H_1+v_{-1}v_4H_2+v_0v_1H_3$	$0.5\cdot\begin{pmatrix} 0 & sb_2 \\ -sb_2 & 0 \end{pmatrix}$
L_{34}	H_3	$0.5\cdot\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$

Таблица III: коэффициенты разложений по $\partial_t, \partial_p, \partial_q, \partial_r$				
Векторное поле	∂_t	∂_p	∂_q	∂_r
$H_0 = L_{-10} - L_{12}$	0	0	-1	1
$H_1 = L_{-11} - L_{02}$	0	sin(q-r)	(coth p)cos(q-r)	-(tanh p)cos(q-r)
$H_2 = L_{-12} + L_{01}$	0	cos(q-r)	(coth p)sin(r-q)	(tanh p)sin(q-r)
$J_0 = L_{-10} + L_{12}$	0	0	1	1
$J_1 = L_{-11} + L_{02}$	0	sin(q+r)	(coth p)cos(q+r)	(tanh p)cos(q+r)
$J_2 = L_{-12} - L_{01}$	0	cos(q+r)	-(coth p)sin(q+r)	-(tanh p)sin(q+r)
$H_3 = J_3 = L_{34}$	1	0	0	0

Таблица IV						
L	Lv_{-1}	Lv_0	Lv_1	Lv_2	Lv_3	Lv_4
L_{-10}	0	0	0	0	v_4	$-v_3$
L_{-11}	0	0	0	v_4	0	v_2
L_{-12}	0	0	v_4	0	0	v_1
L_{-13}	$v_{-1}v_0v_4$	${v_{-1}}^2v_4$	$-v_0v_1v_4$	$-v_0v_2v_4$	$-v_0v_3v_4$	$v_0(1-{v_4}^2)$
L_{-14}	${v_0}^2v_4$	$-v_{-1}v_0v_4$	$-v_{-1}v_1v_4$	$-v_{-1}v_2v_4$	$-v_{-1}v_3v_4$	$v_{-1}(1-{v_4}^2)$
L_{01}	0	0	0	v_3	v_2	0
L_{02}	0	0	v_3	0	v_1	0
L_{03}	$-v_{-1}v_0v_3$	${v_{-1}}^2v_3$	$-v_0v_1v_3$	$-v_0v_2v_3$	$v_0(1-{v_3}^2)$	$-v_0v_3v_4$
L_{04}	$v_3{v_0}^2$	$-v_{-1}v_0v_3$	$-v_{-1}v_1v_3$	$-v_{-1}v_2v_3$	$v_{-1}(1-{v_3}^2)$	$-v_{-1}v_3v_4$
L_{12}	0	0	$-v_2$	v_1	0	0
L_{23}	$v_{-1}v_0v_1$	$-{v_{-1}}^2v_1$	$v_0(1+{v_1}^2)$	$v_0v_1v_2$	$v_0v_1v_3$	$v_0v_1v_4$
L_{31}	$-v_{-1}v_0v_2$	${v_{-1}}^2v_2$	$-v_0v_1v_2$	$-v_0(1+{v_2}^2)$	$-v_0v_2v_3$	$-v_0v_2v_4$
L_{14}	$-v_2{v_0}^2$	$v_{-1}v_0v_2$	$v_{-1}v_1v_2$	$v_{-1}(1+{v_2}^2)$	$v_{-1}v_2v_3$	$v_{-1}v_2v_4$
L_{24}	$-v_1{v_0}^2$	$v_{-1}v_0v_1$	$v_{-1}(1+{v_1}^2)$	$v_{-1}v_1v_2$	$v_{-1}v_1v_3$	$v_{-1}v_1v_4$
L_{34}	$-v_0$	v_{-1}	0	0	0	0
H_0	0	0	v_2	$-v_1$	v_4	$-v_3$
H_1	0	0	$-v_3$	v_4	$-v_1$	v_2
H_2	0	0	v_4	v_3	v_2	v_1
$H_3 = J_3$	$-v_0$	v_{-1}	0	0	0	0
J_0	0	0	$-v_2$	v_1	v_4	$-v_3$
J_1	0	0	v_3	v_4	v_1	v_2
J_2	0	0	v_4	$-v_3$	$-v_2$	v_1

Литература

1. Baez J.C., Segal I.E., Zhou Z. Introduction to Algebraic and Constructive Quantum Field Theory. Princeton University Press, Princeton, 1992.
2. Гуц А.К., Левичев А.В. К основам теории относительности. Доклады Академии Наук СССР. 1984. № 277. С. 1299–1303.
3. Kon M., Levichev A. Towards Analysis in Space-Time Bundles Based on Pseudo-Hermitian Realization of the Minkowski Space // Journal of Functional Analysis. 2016. (submitted).
4. Левичев А.В., Левичева В.Ю. Анализ в космических расслоениях. Выпуск 1: Основы хронометрии и скалярное расслоение. Новосибирский государственный университет, Новосибирск, 1993.
5. Levichev A.V. Segal's chronometry: emergence of the theory and its application to physics of particles and interactions // The Search for Mathematical Laws of the Universe: Physical Ideas, Approaches and Concepts, eds. MM Lavrentiev and VN Samoilov (Novosibirsk: Academic Publishing House). 2010. C. 69–99.
6. Levichev A.V. Pseudo-Hermitian realization of the Minkowski world through DLF theory // Physica Scripta. 2010. T. 83, N. 1. C. 015101.
7. Levichev A. A Contribution to the DLF-theory: on singularities of the SU(2,2)-action in U(1,1) // Journal of Modern Physics. 2016. (accepted for publication).
8. Levichev A.V., Feng J. More on the Mathematics of the DLF Theory: Embedding of the Oscillator World L into Segal's Compact Cosmos D // AJUR. 2013. V. 11(3–4). P. 29–33.
9. Paneitz S.M., Segal I.E. Analysis in space-time bundles I: General considerations and the scalar bundle // Journal of Functional Analysis. 1982. V. 47. P. 78–142.
10. Segal I.E. Mathematical Cosmology and Extragalactic Astronomy. New York: Academic Press, 1976.
11. Segal I.E., Jakobsen H.P., Ørsted B., Paneitz S.M., Speh B. Covariant chronogeometry and extreme distances: Elementary particles // Proceedings of the National Academy of Sciences. 1981. T. 78, N. 9. C. 5261–5265.
12. Segal I.E. Is the Cygnet the quintessential baryon? // Proc. Natl. Acad. Sci. 1991. V. 88. P. 994–998.
13. Segal Archive, MIT, `http://math.mit.edu/segal-archive/publications_03_09_08.pdf`
14. Werth J.-E. Conformal group actions and Segal's cosmology // Rep. Mathematical Phys. 1986. V. 23(2). P. 257–268.
15. Wigner E.P. On unitary representations of the inhomogeneous Lorentz group // Annals of Mathematics. 1939. V. 40 (1). P. 149-204.

U(1,1)-BASED ANALYSIS IN SPACE-TIME BUNDLES: THE TABLES OF THE INFINITESIMAL SU(2,2)-ACTION

A.V. Levichev[1]
Dr.Sc. (Phys.-Math.), Professor, Senior Researcher, e-mail: alevichev@gmail.com
A.Yu. Palyanov[2,3]
Ph.D. (Phys.-Math.), Senior Researcher, e-mail: palyanov@iis.nsk.su

[1]Sobolev Institute of Mathematics
[2]A.P. Ershov Institute of Informatics Systems
[3]Novosibirsk State University

Abstract. Segal's Chronometric Theory is based on the space-time **D** which can be represented by a Lie group with a causal structure determined by an invariant Lorentzian form on the Lie algebra $u(2)$. Similarly, the space-time **F** is represented by a Lie group with a causal structure determined by an invariant Lorentzian form on the Lie algebra $u(1,1)$. The Lie groups G, G_F are introduced as two representations of $SU(2,2)$ which are conjugate via particular matrix W from $Gl(4)$. Linear-fractional G-action on **D** is global and conformal; it is instrumental in the analysis of space-time bundles which is based on the parallelizing group $U(2)$. The latter analysis was carried out by Paneitz and Segal in 1980s. Linear-fractional G_F-action on **F** (introduced by Levichev in 2000s) is also conformal. Despite singularities of the latter action, the group $U(1,1)$ can be chosen as the parallelizing one. In the paper we obtain tables (similar to the "Paneitz-Segal tables") which are necessary in order to perform the analysis of space-time bundles based on the parallelizing group $U(1,1)$.

Keywords: parallelizations of space-time bundles, Segal's cosmos, conformal group $SU(2,2)$ actions on $U(2)$ and on $U(1,1)$, DLF-theory.

Дата поступления в редакцию: 23.07.2016

Mathematical
Structures and Modeling
2016. N. 4(40). PP. 39–43

UDC 519.85

WHY LOCATING LOCAL OPTIMA IS SOMETIMES MORE COMPLICATED THAN LOCATING GLOBAL ONES

Olga Kosheleva
Ph.D. (Phys.-Math.), Professor, e-mail: olgak@utep.edu
Vladik Kreinovich
Ph.D. (Phys.-Math.), Professor, e-mail: vladik@utep.edu

University of Texas at El Paso, El Paso, Texas 79968, USA

Abstract. In most applications, practitioners are interested in locating global optima. In such applications, local optima that result from some optimization algorithms are an unnecessary side effect. In other words, in such applications, locating global optima is a much more computationally complex problem than locating local optima. In several practical applications, however, local optima themselves are of interest. Somewhat surprisingly, it turned out that in many such applications, locating all local optima is a much more computationally complex problem than locating all global optima. In this paper, we provide a theoretical explanation for this surprising empirical phenomenon.

Keywords: local optima, global optima, computational complexity.

1. Formulation of the Problem

A usual understanding is that global optimization is harder. There are many optimization techniques, starting with the simple gradient descent. A usual problem with these techniques is that when they converge, they often lead to a *local* optimum, not to a global one. It takes a special effort to come up with a global optimum instead of a local one.

From this viewpoint, it looks like global optimization is more difficult than locating local optima; see, e.g., [1–3, 5, 6].

While locating a local optimum may be easier, locating all local optima is difficult. What is indeed relatively easy is locating *a* local optimum. In some practical situations, however, we are actually interested in *all* local optima (see, e.g., [7]); for example:

- in spectral analysis, chemical species are identified by local maxima of the spectrum;
- in radioastronomy, radiosources and their components are identified as local maxima of the brightness distribution; see, e.g., [8];
- elementary particles are identified by locating local maxima of the dependence of scattering intensity on the energy.

It turns out that empirically, the computation problem of finding all local optima is much more computationally complicated than the problem of finding all global optima; see, e.g., [4].

Problem – and what we do in this paper. While empirically, computing local optima is often more complex than computing global ones, there has been, to the best of our knowledge, no convincing theoretical explanation for this complexity.

The main goal of this paper is to provide such a theoretical explanation.

2. Local Optima Are Often More Complex to Locate Than Global Optima: A Possible Theoretical Explanation

Approximating the objective function: a frequent way to solve optimization problems. Often, the computational complexity of an optimization problem is due to the complexity of the objective function. Thus, a reasonable idea is:

- to approximate the original objective function $f(x)$ by a close simpler one $f_\varepsilon(x)$,
- solve the corresponding optimization problem for this simpler objective function $f_\varepsilon(x)$, and
- to use the resulting solution x_ε as a first approximation to the solution of the original optimization problem.

This idea indeed helps in solving global optimization problems, see, e.g., [1–3,5,6].

What we do in this paper. What we will prove is that this simplifying idea cannot be used for locating local optima. This is our first theoretical explanation of why locating local optima is often more computationally complicated than locating global optima.

Definitions and the main result. Let us first explain why the above idea is helpful for locating global maxima: namely, that the above idea helps us dismiss some locations as definitely not containing locations of global optima:

Proposition 1. *Let $f(x)$ and $f_\varepsilon(x)$ be two functions which are ε-close, i.e., for which $|f(x)-f_\varepsilon(x)|\leqslant\varepsilon$ for all x, and let x_ε be a location of the global maximum of the function $f_\varepsilon(x)$, i.e., $f_\varepsilon(x_\varepsilon)=\max\limits_x f_\varepsilon(x)$. Then, for each location $x_{\max}$ of the global maximum of the function $f(x)$, we have $f_\varepsilon(x_{\max})\geqslant f_\varepsilon(x_\varepsilon)-2\varepsilon$.*

Proof. From the fact that $x_{\max}$ is a location of the global maximum of the function $f(x)$, we conclude, in particular, that $f(x_{\max})\geqslant f(x_\varepsilon)$. Here,

$$|f(x)-f_\varepsilon(x)|\leqslant\varepsilon$$

for all x, in particular, $f_\varepsilon(x_{\max})\geqslant f(x_{\max})-\varepsilon$ and $f(x_\varepsilon)\geqslant f_\varepsilon(x_\varepsilon)-\varepsilon$. Thus,

$$f_\varepsilon(x_{\max})\geqslant f(x_{\max})-\varepsilon\geqslant f(x_\varepsilon)-\varepsilon\geqslant(f_\varepsilon(x_\varepsilon)-\varepsilon)-\varepsilon=f_\varepsilon(x_\varepsilon)-2\varepsilon.$$

The proposition is proven.

Proposition 2. *For every $\varepsilon > 0$, for every continuous function $f_\varepsilon(x)$, and for every point x_0, there exists a function $f(x)$ which is ε-close to $f_\varepsilon(x)$ and which attains a local maximum at the point x_0.*

Discussion. Thus, even if we know everything about the approximating function, we cannot dismiss any point x as a possible location of a local maximum – in other words, the above idea indeed does not work for locating local optima.

Proof. Since the function $f_\varepsilon(x)$ is continuous, there exists a $\delta > 0$ for which $d(x_0, x) \leqslant \delta$ implies that $|f_\varepsilon(x) - f_\varepsilon(x_0)| \leqslant \dfrac{\varepsilon}{2}$.

Let us define an auxiliary function $g(x)$ which is equal to:

- $g(x) = f_\varepsilon(x_0)$ when $d(x, x_0) \geqslant \dfrac{\delta}{2}$ and
- $g(x) = f_\varepsilon(x_0) + \dfrac{\varepsilon}{2} - d(x, x_0) \cdot \dfrac{\varepsilon}{\delta}$ for all other x.

One can easily see that this function is continuous, and that it has a local maximum (actually, even global maximum) for $x = x_0$.

For values x for which $d(x, x_0) \leqslant d$, the largest possible difference

$$|g(x) - f_\varepsilon(x_0)|$$

between $g(x)$ and $f_\varepsilon(x_0)$ is attained in the second case at the point x_0, when the distance $d(x, x_0) = 0$. In this case, the difference is equal to $|g(x_0) - f_\varepsilon(x_0)| = \dfrac{\varepsilon}{2}$. Thus, for all x, we have $|f_\varepsilon(x_0) - g(x)| \leqslant \dfrac{\varepsilon}{2}$. So, for all these x, we have

$$|f_\varepsilon(x) - g(x)| \leqslant |f_\varepsilon(x) - f_\varepsilon(x_0)| + |f_\varepsilon(x_0) - g(x)| \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Hence, when x is δ-close to x_0, the values $g(x)$ and $f_\varepsilon(x)$ are ε-close.

Let us now consider the second auxiliary function $w(x)$, which is equal to:

- $w(x) = 1$ when $d(x, x_0) \leqslant \dfrac{\delta}{2}$;
- $w(x) = 1 - \dfrac{d(x, x_0)}{\delta/2}$ when $\dfrac{\delta}{2} \leqslant d(x, x_0) \leqslant \delta$; and
- $w(x) = 0$ when $d(x, x_0) \geqslant \delta$.

One can check that this function $w(x)$ is also continuous, and its values are always between 0 and 1.

Thus, the convex combination $f(x) \stackrel{\text{def}}{=} w(x) \cdot g(x) + (1 - w(x)) \cdot f_\varepsilon(x)$ is continuous and ε-close to the original function $f_\varepsilon(x)$. For points x for which $d(x, x_0) \leqslant \dfrac{\delta}{2}$, we have $f(x) = g(x)$, and thus, the function $f(x)$ indeed attains a local maximum for $x = x_0$. The proposition is proven.

Additional theoretical explanation. An additional theoretical explanation for the empirical computational complexity of locating local optima is that this problem also has a higher logical complexity, i.e., needs more quantifiers to describe.

Indeed, the fact that a function $f(x)$ attains its global maximum at a point x_0 is naturally described by a one-quantifier formula $\forall x\,(f(x) \leqslant f(x_0))$. However, to describe the fact that there is a local maximum at the point x_0, we need two quantifiers: $\exists\delta\,\forall x\,(d(x, x_0) \leqslant \delta \to f(x) \leqslant f(x_0))$.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

The authors are thankful to Walter Murray and to all the participants of the International Conference on Frontiers in Global Optimization (Santorini, Greece, June 8–12, 2003) for valuable discussions.

References

1. Hendrix E.M.T., G.-Tóth B. Introduction to Nonlinear and Global Optimization. Springer Verlag, New York, 2010.
2. Horst R., Pardalos P.M., Thoai N.V. Introduction to Global Optimization. Kluwer Academic Publishers, Dordrecht, 2000.
3. Locatelli M., Schoen F. Global Optimization: Theory, Algorithms, and Applications. SIAM Publishers, Philadelphia, Pennsylvania, 2013.
4. Murray W. Proceedings of the International Conference on Frontiers in Global Optimization. Santorini, Greece, June 8–12, 2003.
5. Nocedal J., Wright S. Numerical Optimization. Springer Verlag, New York, 2006.
6. Pardalos P.M., Romeijn H.E. Handbook of Global Optimization: Volume 2. Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Dordrecht, 2002.
7. Villaverde K., Kreinovich V. A linear-time algorithm that locates local extrema of a function of one variable from interval measurement results // Interval Computations. 1993. N. 4. P. 176–194.
8. Verschuur G.L., Kellermann K.I. Galactic and Extra-Galactic Radio Astronomy. Springer Verlag, Berlin, Heidelberg, New York, 1974.

ПОЧЕМУ ЗАДАЧА ПОИСКА ЛОКАЛЬНОГО ОПТИМУМА ИНОГДА СЛОЖНЕЕ, ЧЕМ ЗАДАЧА ПОИСКА ГЛОБАЛЬНОГО ОПТИМУМА

О. Кошелева
к.ф.-м.н., профессор, e-mail: olgak@utep.edu
В. Крейнович
к.ф.-м.н., профессор, e-mail: vladik@utep.edu

Техасский университет в Эль Пасо, США

Аннотация. В большинстве случаев при решении практических задач исследователи заинтересованы в поиске глобального оптимума. В таких задачах локальные оптимумы, найденные в результате использования некоторых алгоритмов оптимизации, являются ненужным побочным эффектом. Другими словами, в таких задачах поиск глобального оптимума является гораздо более вычислительно сложной проблемой, чем обнаружение локальных оптимумов. Однако, в ряде практических задач локальные оптимумы сами представляют интерес. Несколько удивительно, но, как оказалось, что во многих таких задачах поиск всех локальных оптимумов является гораздо более вычислительно сложной проблемой, чем поиск всех глобальных оптимумов. В этой статье мы приводим теоретическое объяснение этого удивительного эмпирического явления.

Ключевые слова: локальный оптимум, глобальный оптимум, вычислительная сложность.

Дата поступления в редакцию: 11.07.2016

Mathematical
Structures and Modeling
2016. N. 4(40). PP. 44–47

UDC 519.86+519.214

BELL-SHAPED CURVE FOR PRODUCTIVITY GROWTH: AN EXPLANATION

Olga Kosheleva
Ph.D. (Phys.-Math.), Professor, e-mail: olgak@utep.edu
Vladik Kreinovich
Ph.D. (Phys.-Math.), Professor, e-mail: vladik@utep.edu

University of Texas at El Paso, El Paso, Texas 79968, USA

Abstract. A recent analysis of the productivity growth data shows, somewhat surprisingly, that the dependence of the 20-century productivity growth on time can be reasonably well described by a Gaussian formula. In this paper, we provide a possible theoretical explanation for this observation.

Keywords: Gaussian curve, productivity growth, Central Limit theorem.

1. Formulation of the Problem

An empirical fact. A recent book [2] shows that, when averaged over decades, the productivity growth $n(t)$ in the US from 1900 until now follows a bell-shaped Gaussian curve $n(t) = c_0 \cdot \exp(-k_0 \cdot (t - t_0)^2)$, for appropriate values c_0, k_0, and t_0 (see also [1]).

This fact is somewhat surprising. Such curves normally describe how the probability of a certain value x depend on this value. It is somewhat surprising to find a similar curve in the description of how productivity growth depends on time.

What we do in this paper. In this paper, we provide a possible explanation for this surprising phenomenon.

2. Our Explanation

Main idea. Eventually, productivity growth can be traced to new inventions. However, the appearance of a new invention does not immediately boost the productivity growth:

- inventions are usually formulated in somewhat abstract theoretical terms, and therefore
- it takes quite some time and effort to adopt and modify the original invention so that it would start boosting up productivity.

From the main idea to precise formulas: first approximation. Let $c(t)$ be the number of inventions per time unit. An invention made at time t_1 leads to

a productivity boost at some later time $t > t_1$. The corresponding time delay $\Delta t = t - t_1$ is, in general, different for different inventions. The exact value of this delay is unpredictable, so it makes sense to consider this delay as a random variable.

Let $\rho(\Delta t)$ be the probability density that describes the probability distribution of different delays. The boost of productivity at moment t_1 can be caused by inventions made at different past moments of time $t = t_1 - (t_1 - t)$. At each moment t, $c(t)$ inventions were made, and the probability of each of these inventions leading to a productivity boost at moment t_1 – i.e., the fraction of those inventions that lead to a productivity boost at moment t_1 – is proportional to $\rho(t_1 - t)$. Thus, at moment t_1, the increase in productivity caused by these inventions is proportional to the product $\rho(t_1 - t) \cdot c(t)$. The overall productivity increase at moment t_1 can be obtained if we add up all the increases corresponding to all moments t. Thus, the productivity growth $n(t_1)$ at moment t_1 is proportional to the sum $\int \rho(t_1 - t) \cdot c(t)\, dt$.

A more detailed analysis. In the above description, we considered a transition from an invention to productivity growth as a single stage. In reality, this transition is very complex, it contains many stages.

First, we need to transform the original raw ideas into solid science. This may also involve several steps. For example:

- first, we test the idea on a small sample, to provide a proof of concept;
- once this testing confirms the idea, we get the funding to test it on a larger sample, etc.

Often, during this testing, it becomes necessary to modify and update the original idea.

All this constitutes research. Once the research is done, we need to work on development, to think of how the original research ideas can be best implemented in an efficient way. This may also take several steps:

- first we implement it on a small scale (as computers and additive manufacturing were),
- then we find the way to make it more widely spread, etc.

How this informal analysis changes the corresponding mathematical model. According to the above analysis, instead of a *single* large delay Δt, it is more appropriate to consider it as the sum of *several* (much smaller) delays corresponding to different stages of the transition from the original invention to the increase in productivity: $\Delta t = \Delta t_1 + \ldots + \Delta t_m$.

Different stages are reasonably independent. Thus, if we denote by $\rho_i(\Delta t_i)$ the probability distribution corresponding to the i-th stage, then:

- the average number $c_1(t_1)$ of inventions that finished the first stage at moment t_1 will be proportional to $\int \rho_1(t_1 - t) \cdot c(t)\, dt$;

- the average number $c_2(t_2)$ of inventions that finished the second stage at moment t_2 is proportional to $\int \rho_2(t_2 - t_1) \cdot c_1(t_1)\, dt_1$, i.e., to

$$\int \rho_2(t_2 - t_1) \cdot \rho_1(t_1 - t) \cdot c(t)\, dt\, dt_1;$$

- the average number $c_3(t_3)$ of inventions that finished the third stage at moment t_3 is proportional to $\int \rho_3(t_3 - t_2) \cdot c_2(t_2)\, dt_2$, i.e., to

$$\int \rho_3(t_3 - t_2) \cdot \rho_2(t_2 - t_1) \cdot \rho_1(t_1 - t) \cdot c(t)\, dt\, dt_1\, dt_2;$$

- etc.

- finally, the productivity growth $n(t_m)$ at moment t_m is proportional to the average number of inventions that finished all m stages at this moment, i.e., to

$$\int \rho_m(t_m - t_{m-1}) \cdot \ldots \cdot \rho_1(t_1 - t) \cdot c(t)\, dt\, dt_1\, \ldots\, dt_{m-1}.$$

This formula helps explain the Gaussian shape. From the mathematical viewpoint, this formula means that the desired function $n(t)$ is proportional to the convolution of a large number of functions $c(t)$, $\rho_1(\Delta t_1)$, $\rho_2(\Delta t_2)$, $\ldots$, and $\rho_m(\Delta t_m)$. This is exactly the same formula that describes how the probability density function (pdf) of the sum of many independent random variables depends on the pdfs of the components of this sum.

For random variables, there is a Central Limit Theorem, according to which, under some reasonable conditions, the distribution of the sum of many relatively small random variables is close to Gaussian; see, e.g., [3]. In terms of convolution, this means that, under the corresponding conditions, the convolution of a large number of non-negative functions is close to the Gaussian bell-shaped curve.

Since, according to our argument, the productivity growth function $n(t)$ can be described as such a convolution, it then follows that this function is indeed close to Gaussian. In other words, we have the desired explanation.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

References

1. Glaeser E. Those were the days // The Wall Street Journal. January 16–17, 2016. P. C5 and C7.
2. Gordon R.J. The Rise and Fall of American Growth. Princeton University Press, Princeton, New Jersey, 2016.
3. Sheskin D.J. Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, Boca Raton, Florida, 2011.

КОЛОКОЛООБРАЗНАЯ КРИВАЯ РОСТА ПРОИЗВОДИТЕЛЬНОСТИ: ОБЪЯСНЕНИЕ

О. Кошелева
к.ф.-м.н., профессор, e-mail: olgak@utep.edu
В. Крейнович
к.ф.-м.н., профессор, e-mail: vladik@utep.edu

Техасский университет в Эль Паco, США

Аннотация. Недавний анализ данных роста производительности показывает удивительный факт: зависимость роста производительности 20-го века от времени может быть достаточно хорошо описана гауссовой формулой. В этой статье мы приводим возможное теоретическое объяснение этого наблюдения.

Ключевые слова: кривая Гаусса, рост производительности, центральная предельная теорема.

Дата поступления в редакцию: 11.07.2016

Mathematical Structures and Modeling
2016. N. 4(40). PP. 48–54

UDC 625.8+517.9

WHY COMPACTION METER VALUE (CMV) IS A GOOD MEASURE OF PAVEMENT STIFFNESS: TOWARDS A POSSIBLE THEORETICAL EXPLANATION

Andrzej M. Pownuk
Ph.D. (Phys.-Math.), Instructor, e-mail: ampownuk@utep.edu
Pedro Barragan Olague
Student, e-mail: pabarraganolague@miners.utep.edu
Vladik Kreinovich
Ph.D. (Phys.-Math.), Professor, e-mail: vladik@utep.edu

University of Texas at El Paso, El Paso, Texas 79968, USA

Abstract. To measure stiffness of the compacted pavement, practitioners use the Compaction Meter Value (CMV); a ratio between the amplitude for the first harmonic of the compactor's acceleration and the amplitude corresponding to the vibration frequency. Numerous experiments show that CMV is highly correlated with the pavement stiffness, but as of now, there is no convincing theoretical explanation for this correlation. In this paper, we provide a possible theoretical explanation for the empirical correlation. This explanation also explains why, the stiffer the material, the more higher-order harmonics we observe.

Keywords: pavement, compaction, Compaction Meter Value (CMV).

1. Compaction Meter Value (CMV) – An Empirical Measure of Pavement Stiffness

Need to measure pavement stiffness. Road pavement must be stiff: the pavement must remain largely unchanged when heavy vehicles pass over it.

To increase the pavement's stiffness, pavement layers are usually compacted by the rolling compactors. In the cities, only non-vibrating compactors are used, to avoid human discomfort caused by vibration. However, in roads outside the city limits, vibrating compactors are used, to make compaction more efficient. In this paper, we will denote the vibration frequency by f.

Compaction is applied both to the soil and to the stiffer additional pavement material that is usually placed on top of the original soil. To check whether we need another round of compaction and/or another layer of additional material on top, we need to measure the current pavement stiffness.

Ideally, we should measure stiffness as we compact. In principle, we can measure stiffness *after* each compaction cycle, but it would be definitely more

efficient to measure it *during* the compaction – this way we save time and we save additional efforts needed for post-compaction measurements.

What we *can* rather easily measure during compaction is acceleration; it is therefore desirable to estimate the pavement stiffness based on acceleration measurements.

Compaction Meter Value (CMV). It turns out that reasonably good estimates for stiffness can be obtained if we apply Fourier transform to the signal describing the dependence of acceleration on time, and then evaluate *Compaction Meter Value* (CMV), a ratio A_2/A_1 between the amplitudes corresponding to the frequencies $2f$ and f. This measure was first introduced in the late 1970s [3, 10, 11].

Numerous experiments have confirmed that CMV is highly correlated with more direct characteristics of stiffness such as different versions of elasticity modulus; see, e.g., [2, 6, 7, 12, 13].

CMV remains one of the main ways of estimating stiffness; see, e.g., [5].

Can we use other Fourier components? Since the use of the double-frequency component turned out to be so successful, a natural idea is to try to use other Fourier components.

It turns out that when the soil is soft (not yet stiff enough), then even the double-frequency Fourier component is not visible above noise. As the pavement becomes stiffer, we can clearly see first the first harmonic, then also higher harmonics, i.e., harmonics corresponding to $3f$, $4f$, etc.

Remaining problem. While the relation between CMV and stiffness is an empirical fact, from the theoretical viewpoint it remains somewhat a mystery: to the best of our knowledge, there is no theoretical explanation for this empirical dependence.

In this paper, we attempt to provide such a theoretical explanation.

2. A Possible Theoretical Explanation of an Empirical Correlation Between CMV and Stiffness

Analysis of the problem: towards the corresponding equations. Let us start our analysis with the extreme situation when there is no stiffness at all. Crudely speaking, the complete absence of stiffness means that particles forming the soil are completely independent from each one other: we can move some of them without affecting others.

In this extreme case, the displacement x_i of each particle i is determined by the Newton's equations

$$\frac{d^2x_i}{dt^2} = \frac{1}{m_i} \cdot F_i, \tag{1}$$

where m_i is the mass of the i-th particle and F_i is the force acting on this particle. For a vibrating compactor, the force F_i is sinusoidal with frequency f. Thus, the corresponding accelerations are also sinusoidal with this same frequency. In this extreme case, after the Fourier transform, we will get only one component – corresponding to the vibration frequency f.

Stiffness k means that, in addition to the external force F_i, the acceleration of each particle i is also influenced by the locations of other particles x_j. For example, if we move one of the particles forming the soil, other particle move as well so that the distances between the particles remain largely the same. Thus, instead of the simple Newton's equations (1), we have more complicated equations

$$\frac{d^2x_i}{dt^2} = \frac{1}{m_i} \cdot F_i + f_i(k, x_1, \ldots, x_N), \tag{2}$$

for some expression $f_i(k, x_1, \ldots, x_N)$.

Displacements are usually small. We consider the case when stiffness is also reasonably small. It is therefore reasonable to expand this expression in Taylor series and keep only the first few terms in this expansion.

With respect to k, in the first approximation, we just keep linear terms. With respect to x_j, it is known that the corresponding processes are observably non-linear (see, e.g., [1,4,9]) so we need to also take non-linear terms into account; the simplest non-linear terms are the quadratic ones, so we end up with the following approximate model:

$$\frac{d^2x_i}{dt^2} = \frac{1}{m_i} \cdot F_i + k \cdot \sum_{j=1}^{N} a_{ij} \cdot x_j + k \cdot \sum_{j=1}^{N} \sum_{\ell=1}^{N} a_{ij\ell} \cdot x_j \cdot x_k. \tag{3}$$

Solving the resulting equations. In general, the solution to the equations (3) depends on the value k: $x_i(t) = x_i(k, t)$.

When deriving the equations (3), we ignored terms which are quadratic (or of higher order) in terms of k. It is therefore reasonable, when looking for solutions to this equation, to also ignore terms which are quadratic (or of higher order) in k, i.e., to take

$$x_i(k, t) = x_i^{(0)}(t) + k \cdot x_i^{(1)}(t). \tag{4}$$

If we plug in the formula (5) into the equation (3) and ignore terms which are quadratic in k, then we end up with the equation

$$\frac{d^2x_i^{(0)}}{dt^2} + k \cdot \frac{d^2x_i^{(1)}}{dt^2} = \frac{1}{m_i} \cdot F_i + k \cdot \sum_{j=1}^{N} a_{ij} \cdot x_j^{(0)} + k \cdot \sum_{j=1}^{N} \sum_{\ell=1}^{N} a_{ij\ell} \cdot x_j^{(0)} \cdot x_\ell^{(0)}. \tag{5}$$

This formula should hold for all k, so:

- terms independent on k should be equal on both sides, and
- terms linear in k should be equal on both sides.

By equating terms in (5) that do not depend on k, we get the linear equation

$$\frac{d^2x_i^{(0)}}{dt^2} = \frac{1}{m_i} \cdot F_i, \tag{6}$$

which, for the sinusoidal force $F_i(t) = A_i \cdot \cos(\omega \cdot t + \Phi_i)$, has a similar sinusoidal form

$$x_i^{(0)}(t) = a_i \cdot \cos(\omega \cdot t + \varphi_i) \tag{7}$$

for appropriate values a_i and φ_i.

By equating terms linear in k on both sides of the equation (5), we conclude that

$$\frac{d^2 x_i^{(1)}}{dt^2} = \sum_{j=1}^{N} a_{ij} \cdot x_j^{(0)} + \sum_{j=1}^{N} \sum_{\ell=1}^{N} a_{ij\ell} \cdot x_j^{(0)} \cdot x_\ell^{(0)}. \tag{8}$$

For the sinusoidal expression (7) for $x_i^{(0)}$:

- linear terms $\sum\limits_{j=1}^{N} a_{ij} \cdot x_j^{(0)}$ in the right-hand side are sinusoidal with the same angular frequency ω (i.e., with frequency f), while
- quadratic terms $\sum\limits_{j=1}^{N} \sum\limits_{\ell=1}^{N} a_{ij\ell} \cdot x_j^{(0)} \cdot x_\ell^{(0)}$ are sinusoids with the double angular frequency 2ω (i.e., with double frequency $2f$).

Thus, the right-hand side of the equation (8) is the sum of two sinusoids corresponding to frequencies f and $2f$, and so,

$$\frac{d^2 x_i}{dt^2} = \frac{d^2 x_i^{(0)}}{dt^2} + k \cdot \frac{d^2 x_i^{(1)}}{dt^2} = A_i \cdot \cos\left(\omega \cdot t + \Phi_i\right) +$$

$$+k \cdot \left(A_i^{(1)} \cdot \cos\left(\omega \cdot t + \Phi_i^{(1)}\right) + A_i^{(2)} \cdot \cos\left(2\omega \cdot t + \Phi_i^{(2)}\right)\right). \tag{9}$$

The measured acceleration $a(t)$ is the acceleration of one of the points $a(t) = \dfrac{d^2 x_{i_0}(t)}{dt^2}$, thus the measured acceleration has the form

$$a(t) = A_{i_0}^{(0)} \cdot \cos\left(\omega \cdot t + \Phi_{i_0}^{(0)}\right) +$$

$$+k \cdot \left(A_{i_0}^{(1)} \cdot \cos\left(\omega \cdot t + \Phi_{i_0}^{(1)}\right) + A_{i_0}^{(2)} \cdot \cos\left(2\omega \cdot t + \Phi_{i_0}^{(2)}\right)\right). \tag{10}$$

In this expression, we only have terms sinusoidal with frequency f and terms sinusoidal with frequency $2f$. Thus, in this approximation, the Fourier transform of the acceleration consists of only two components:

- a component corresponding to the main frequency f (and the corresponding angular frequency ω), and
- a component corresponding to the first harmonic $2f$, with the angular frequency 2ω.

The amplitude A_2 of the first harmonic 2ω is equal to $A_2 = k \cdot A_{i_0}^{(2)}$. The amplitude A_1 of the main frequency ω is equal to $A_1 = A_{i_0}^{(1)} + k \cdot c$ for some constant c depending on the relation between the phases. Thus, the ratio of these two amplitudes has the form

$$\frac{A_2}{A_1} = \frac{k \cdot A_{i_0}^{(2)}}{A_{i_0}^{(1)} + k \cdot c}. \tag{11}$$

In all the previous formulas, we ignored terms which are quadratic (or of higher order) in terms of k. If we perform a similar simplification in the formula (11), we conclude that

$$\frac{A_2}{A_1} = k \cdot C, \tag{12}$$

where we denoted $C \stackrel{\text{def}}{=} \dfrac{A_{i_0}^{(2)}}{A_{i_0}^{(1)}}$. In other words, we conclude that the CMV ratio is, in the first approximation, indeed proportional to stiffness.

Main conclusion. We have explained why, for reasonably small stiffness levels, we can only see two Fourier components above the noise level: the component corresponding to the vibrating frequency f and the component corresponding to the first harmonic $2f$.

We have also explained the empirical fact that the CMV – the ratio of the amplitudes of the two harmonics – is proportional to the pavement stiffness.

Case of larger stiffness: analysis and corresponding additional conclusions. When the stiffness k is sufficiently large, we can no longer ignore terms which are quadratic or of higher order in terms of k. In general, the larger the stiffness level, the more terms we need to take into account to get an accurate description of the corresponding dynamics.

Also, when the stiffness k is small, then, due to the fact that the displacements $x_i(t)$ are also reasonably small, the products of k and the terms which are, e.g., cubic in $x_j(t)$ can be safely ignored. However, when k is not very small, we need to take these terms into account as well. Using the corresponding expansion of the equations (3), and taking into account more terms in the expansion of $x_i(k,t)$ in k, we end up with terms which are cubic (or higher order) in terms of the ω-sinusoids $x_i^{(0)}(t)$. These terms correspond to triple, quadruple, and higher frequencies $3f$, $4f$, etc.

This is exactly what we observe: the higher the stiffness, the more higher order harmonics we see. Thus, this additional empirical fact is also theoretically explained.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

The authors are thankful to Soheil Nazarian and Cesar Tirado for valuable discussions.

References

1. Barragan P., Nazarian S., Kreinovich V., Gholamy A., Mazari M. How to estimate resilient modulus for unbound aggregate materials: a theoretical explanation of an empirical formula // Proceedings of the 2016 World Conference on Soft Computing. Berkeley, California, May 22–25, 2016. P. 203–207.
2. Floss R., Bräu G., Gahbauer M., Griber N., Obermayer J. Dynamische Vedichtungsprüfung bei Erd-und Straß enbauten (Dynamic Cpompaction Testing in Earth and Road Construction). Prümaft für Grundbau, Boden- und Felsmechanik, Technische Universerität München, Heft 612, München, 1991.
3. Forssblad L. Compaction meter on vibrating rollers for improved compaction control // Proceedings of the International Conference on Compaction. Partis, France, 1980. V. II. P. 541–546.
4. Mazari M., Navarro E., Abdallah I., Nazarian S. Comparison of numerical and experimental responses of pavement systems using various resilient modulus models // Soils and Foundations. 2014. V. 54, N. 1. P. 36–44.
5. Mooney M.A., Adam D. Vibartory roller integrated measurement of earthwork compaction: an overview // Proceedings of the International Symposium on Field Measurements in Geomechanics FMGM'2007. Boston, Massachusetts, September 24–27, 2007.
6. Mooney M.A., Gorman P.B., Farouk E., Gonzalez J.N., Akanda A.S. Exploring Vibration-Based Intelligent Soft Compaction. Oklahoma Department of Transportation, Projet N. 2146, Final Report, 2003.
7. M.A. Mooney, Gorman P.B., Gonzalez J.N. Vibration-based health monitoring during eathwork construction // Journal of Structural Health Monitoring. 2005. V. 2, N. 4. P. 137–152.
8. Mooney M.A., Rinehart R.V., Facas N.W., Musimbi O.M., White D.J. *Intelligent Soil Compaction Systems*, National Cooperative Highway Research Program (NCHRP) Report 676, Tranportation Research Board, Washington, DC, 2010.
9. Ooi P.S.K., Archilla A.R., Sandefur K.G. Resilient modulus models for comactive cohesive soils // Transportation Research Record. 2006. N. 1874. P. 115–124.
10. Thurner H.F., Forsblad L. Compaction Meter on Vibrating Rollers // Research Bulletin N. 8022. Solna, Sweden.
11. Thurner H.F., Sandström A. A new device for instant compaction control // Proceedings of International Conference on Compaction. Paris, France, 1980. V. II. P. 611–614.
12. White D., Thompson M. Relationships between in-situ and roller-itegrated compaction measurements for granular soils // ASCE Journal of Geotechnical and Geomechanical Engineering. 2008. V. 134, N. 12. P. 1763–1770.
13. White D., Thompson M., Vennapusa P. Filed Validation of Intelligent Compaction Monoitoring Technology for Unbound Materials. Report N. MN/RC 2007-10, Minnesota Department of Transportation, St. Paul, Minnesota, 2008.

ПОЧЕМУ ПОКАЗАТЕЛЬ СТЕПЕНИ УПЛОТНЕНИЯ ГРУНТА (CMV) ЯВЛЯЕТСЯ ХОРОШЕЙ МЕРОЙ ЖЕСТКОСТИ ДОРОЖНОГО ПОКРЫТИЯ: ВОЗМОЖНОЕ ТЕОРЕТИЧЕСКОЕ ОБЪЯСНЕНИЕ

А.М. Повнук
к.ф.-м.н., преподаватель, e-mail: ampownuk@utep.edu
П. Барраган Олаге
студент, e-mail: pabarraganolague@miners.utep.edu
В. Крейнович
к.ф.-м.н., профессор, e-mail: vladik@utep.edu

Техасский университет в Эль Пасо, США

Аннотация. Для измерения жёсткости уплотнённого дорожного покрытия практики используют показатель степени уплотнения грунта (CMV) — это отношение амплитуды первой гармоники ускорения уплотнителя к амплитуде, соответствующей частоте вибрации. Многочисленные эксперименты показывают, что CMV сильно коррелирует с жёсткостью дорожного покрытия, но на данный момент не существует каких-либо убедительных теоретических объяснений этой корреляции. В данной статье мы приводим возможные теоретические объяснения этой эмпирической корреляции. Это объяснение также показывает, почему для более жёсткого материала мы наблюдаем больше гармоник высшего порядка.

Ключевые слова: дорожное покрытие, уплотнение, Compaction Meter Value (CMV).

Дата поступления в редакцию: 01.07.2016

Mathematical
Structures and Modeling
2016. N. 4(40). PP. 55–58

UDC 530.1+514.8

WHY 3-D SPACE? WHY 10-D SPACE? A POSSIBLE SIMPLE GEOMETRIC EXPLANATION

Vladik Kreinovich
Ph.D. (Phys.-Math.), Professor, e-mail: vladik@utep.edu

University of Texas at El Paso, El Paso, Texas 79968, USA

Abstract. In physics, the number of observed spatial dimensions (three) is usually taken as an empirical fact, without a deep theoretical explanation. In this paper, we provide a possible simple geometric explanation for the 3-D character of the proper space. We also provide a simple geometric explanation for the number of additional spatial dimensions that some physical theories use. Specifically, it is known that for some physical quantities, the 3-D space model with point-wise particles leads to meaningless infinities. To avoid these infinities, physicists have proposed that particles are more adequately described not as 0-D points, but rather as 1-D strings or, more generally, as multi-D "M-branes". In the corresponding *M-theory*, proper space is 10-dimensional. We provide a possible geometric explanation for the 10-D character of the corresponding space.

Keywords: spatial dimension, M-theory.

1. Why 3-D Space?

Formulation of the problem. Empirically, our space is 3-dimensional: we need three coordinates to uniquely determine each spatial location. Why three and not two or five?

Modern physics mostly takes the number of dimensions for granted, as an empirical fact, but it would nice to come up with a theoretical explanation for this number. The main objective of this paper is to provide such an explanation.

Main idea and the resulting explanation. In classical physics, the world consists of particles.

Particles interact: e.g., positively and negatively charged particles are attracted to each other. However, this does not necessarily mean that we have to go beyond the particles model: in modern physics, interaction between particles is explained as an exchange of the auxiliary particles responsible for this interaction. For example, electromagnetic forces are explained as an exchange of photons – quanta of the electromagnetic field; see, e.g., [1].

With time, particles move in space; thus, each particle forms a 1-D trajectory in space. Particles can collide; one particle can turn into several others, etc. Thus, these trajectories can intersect. So, from the topological viewpoint, trajectories

form a *graph*, with trajectories as edges and intersections of trajectories as vertices.

From the *physical* viewpoint, the only meaningful spatial locations are points on this graph. However, from the *mathematical* and *computational* viewpoint, analyzing graphs is difficult, it is easier to analyze multi-D manifolds. Thus, it is convenient to embed the graph into a higher-D space. This is similar to the fact that, from the computational viewpoint, it is easier to consider a solid body as a continuous medium instead of explicitly taking into account its discrete atom-by-atom character; see, e.g., [1].

What is the smallest dimension for which we can embed any graph into the manifold of the corresponding dimension? Clearly, the corresponding space cannot be 2-dimensional:

- while some graphs can be embedded into a plane,
- it is well known that not every graph can be embedded into a plane without creating a non-physical additional intersection.

For example, a graph with 5 vertices all of which are connected to each other cannot be thus embedded; see, e.g., [6].

However, it is also known that every finite graph can be embedded into a 3-D space without creating unnecessary intersections. This may be an explanation of why the usual physical space is 3-dimensional: this is a simplest model containing the actual graph-like space.

2. Beyond Point Particle: Why 10-D Space?

Need to go beyond point particles. At first glance, the classical model of pointwise particle is a good consistent description of the physical Universe. However, a more detailed analysis shows that in this seemingly natural model, when we try to estimate the values of some reasonable physical quantities, we get meaningless infinities.

Indeed, let us compute the overall energy of the electric field of a single pointwise charged particle with charge q. The energy density ρ is known to be proportional to the square of the electric field E: $\rho = c \cdot E^2$ for some constant c. According to the Coulomb's law, $E = \dfrac{1}{r^2}$, where r is the distance to the particle. Thus, $\rho = c \cdot E^2 = \dfrac{c \cdot q^2}{r^4}$. The overall energy ε can be obtained if we integrate this density over all spatial locations; thus,

$$\varepsilon = \int \rho(x)\, dV = \int \frac{c \cdot q^2}{r^4}\, dV.$$

Since the integrated function depends only on r, we can integrate over each sphere of radius r and get $dV = 4 \cdot \pi \cdot r^2\, dr$, thus

$$\varepsilon = \int_0^\infty \frac{c \cdot q^2}{r^4} \cdot 4 \cdot \pi \cdot r^2\, dr = c \cdot q^2 \cdot 4 \cdot \pi \cdot \int_0^\infty \frac{dr}{r^2} = -c \cdot q^2 \cdot 4 \cdot \pi \cdot \left| {}_0^\infty \frac{1}{r} = \infty. \right.$$

String etc.: a natural idea. Since point-wise 0-D particles lead to infinities, a natural idea is to assume that particles are higher-dimensional objects: 1-D strings or, more generally, multi-D "M-branes". It turns out that in the corresponding *M-theory*, we can avoid infinities if we consider a 10-D proper space (and 11-D space-time); see, e.g., [2, 8].

A possible simple geometric explanation of 10-D character of proper space. How can we explain this 10-D character without involving complicated math? let us go back to our original idea: that all we have in the world are particles.

The only difference now is that instead of 0-D particles that form 1-D trajectories as they move, now we have at least 1-D particles that, as they move, create 2-D "trajectories".

From the topological viewpoint, the resulting trajectories are already continuous, so there is no topological need to embed them into a higher-dimensional space. However, from the computational viewpoint, it may be beneficial to consider such an embedding if this will allow us to be able to deal with a simpler space – e.g., with a simple Euclidean space instead of the general curved Riemannian one.

It is known – it was originally proven by the Nobelist John Nash – that every Riemannian space can be embedded into an Euclidean space of higher dimension. The bound on this dimension has been significantly improved since Nash's original result. The best estimate so far is that every Riemannian space of dimension n can be embedded into an Euclidean space of dimension

$$N = \frac{n \cdot (n+1)}{2} + n + \max(n, 5);$$

see, e.g., [3–5, 7]. In particular, for the case $n = 2$ of trajectories formed by 1-D particles (strings), we get

$$N = \frac{2 \cdot (2+1)}{2} + 2 + \max(2, 5) = 3 + 2 + 5 = 10.$$

Thus, we indeed get a possible simple geometric explanation of the 10-D character of proper space in M-theories.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

The author is thankful to Boguslaw Stec for valuable discussions.

References

1. Feynman R., Leighton R., Sands M. The Feynman Lectures on Physics. Addison Wesley, Boston, Massachusetts, 2005.

2. Greene B. The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory. W.W. Norton & Company, New York, 2003.
3. Günther M. On the perturbation problem associated to isometric embeddings of Riemannian manifolds // Ann. Global Anal. Geom. 1989. V. 7. P. 69–77.
4. Güunther M. Zum einbettungsatz von J. Nash // Math. Nachr. 1989. V. 144. P. 165–187.
5. Günther M Isometric embeddings of Riemannian manifolds // Proceedings of the International Congress of Mathematicians ICM'1990. Kyoto, Japan. P. 1137–1143.
6. Trudeau R.J. Introduction to Graph Theory. Dover Publ., New York, 1993.
7. Villani C. Théorèmes de plogement(s) isométrique(s) de Nash // Journal de Mathématiques des Élèves de l'Ecole Normal Supérieure de Lyon. March 2016.
8. Witten E. Fivebranes and knots // Quantum Topology. 2012. V. 3, N. 1. P. 1–137.

ПОЧЕМУ ПРОСТРАНСТВО ТРЁХМЕРНОЕ? ПОЧЕМУ ДЕСЯТИМЕРНОЕ? ВОЗМОЖНОЕ ПРОСТОЕ ГЕОМЕТРИЧЕСКОЕ ОБЪЯСНЕНИЕ

В. Крейнович
к.ф.-м.н., профессор, e-mail: vladik@utep.edu

Техасский университет в Эль Пасо, США

Аннотация. В физике число наблюдаемых пространственных измерений (три) обычно принимается как эмпирический факт, без глубокого теоретического объяснения. В этой статье мы приводим возможное простое геометрическое объяснение 3-мерного характера этого пространства. Мы также предлагаем простое геометрическое объяснение ряда дополнительных пространственных измерений, которые используют некоторые физические теории. В частности, известно, что для некоторых физических величин, модель 3-мерного пространства с точечными частицами приводит к возникновению бессмысленных бесконечностей. Чтобы избежать этих бесконечностей, физики предположили, что частицы более адекватно описывать не как 0-мерные точки, а как 1-мерные струны или, в более общем плане, как многомерные "M-браны". В соответствующей M-теории, собственно пространство является 10-мерным. Мы предлагаем возможное геометрическое объяснение для 10-мерного характера соответствующего пространства.

Ключевые слова: пространственное измерение, M-теория.

Дата поступления в редакцию: 11.07.2016

Математические структуры и моделирование
2016. № 4(40). С. 59–64

УДК 517.977

МЕТОД ВОЗВРАТА И РЕАЛИЗАЦИЯ ДИНАМИЧЕСКИХ ОГРАНИЧЕНИЙ В ЗАДАЧАХ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ

Б.К. Нартов
с.н.с., к.ф.-м.н., e-mail: nartov@ofim.oscsbras.ru

Омский филиал Института математики СО РАН им. С.Л. Соболева, г. Омск

Аннотация. Рассмотрена задача оптимального управления с неполной реализацией динамических ограничений. Представленный в работе метод направленной оптимизации начальных условий в задачах управления динамическими системами — метод возврата — предназначался первоначально для оптимизации вектора начальных координат в частной модели конфликта подвижных объектов, характеристики которых ухудшались в результате взаимодействия с объектами противника и старения. Модель связывала характеристики (вектор состояния) и координаты (вектор управления) объектов дифференциальными уравнениями типа уравнений Ланчестера. Далее становилась и решалась конкретная задача оптимального управления движениями группы объектов, противодействующих другой группе объектов с заданными на интервале управления траекториями (по критерию минимизации некоторой функции конечных состояний объектов). Существенно сложнее опорной оказалась задача построения приемлемого по времени счёта и точности алгоритма оптимизации начального вектора управления, то есть начального размещения группы управляемых объектов. Найденный подход оказался весьма общим и позволяет направленно оптимизировать начальный вектор управления по меньшей мере в классе управляемых гладких систем с непрерывно дифференцируемым функционалом качества. В самом общем виде идея метода состоит в том, что для оптимизации, в смысле избранного функционала качества, начальных условий исходной задачи оптимального управления записывается вспомогательная двойственная задача и реализуется итеративный процесс, в шагах которого чередуются исходная и двойственная задачи, а в качестве части начальных условий очередного шага итерации используется часть конечных значений предыдущего шага.

Ключевые слова: динамические системы, оптимизация начальных условий, обратная задача, динамические ограничения.

1. Вводные замечания

Данная работа дополняет нашу работу [2] полным доказательством сходимости предложенного в [1] и исследовавшегося в [2–4] процесса к локальному

или глобальному оптимуму. Кроме того, приведён пример решения задачи оптимального управления гладкой динамической системой с частичной реализацией динамических ограничений на интервале управления.

2. Метод возврата

Рассмотрим динамическую систему

$$\dot{P}_i(t) = f_i(\bar{p}(t),\, \bar{a}(t),\, \bar{u}_i(t)),\; i = 1, ..., N, \tag{1}$$

где f_i — функция, непрерывно дифференцируемая на заданном интервале управления $(0,\, t_f)$;

$$\bar{p}(t) = (p_1(t),\, ...\,,\, p_N(t));$$

$\bar{a}(t)$ — заданные на $(0,\, t_f)$ временные процессы;
$\bar{u}_i(t) = (u_{i1}(t),\, ...\,,\, u_i(t))$ — i-е управление, координата в R^M .

Не оговаривая ограничений на управления, запишем для заданных начальных условий задачу оптимального управления

$$J(\bar{p}(t_f)) \to \inf, \tag{2}$$

где J непрерывно дифференцируема по $p_1(t),\, ...\,,\, p_N(t)$.

Определив далее:

$$\tilde{f}_i = -f_i,$$

$$\tilde{\bar{a}}(t) = \bar{a}(t_f - t),$$

рассмотрим динамическую систему

$$\dot{\tilde{P}}_i(t) = \tilde{f}_i(\tilde{\bar{p}}(t),\, \tilde{\bar{a}}(t),\, \tilde{\bar{u}}(t)),\;\; i = 1, ..., N \tag{3}$$

с начальными условиями

$$\tilde{\bar{p}}(0) = \bar{p}(t_f),$$

$$\tilde{\bar{u}}_i(0) = \bar{u}_i(t_f),\; i = 1, ..., N$$

и запишем для (3) задачу оптимального управления, двойственную задаче (2):

$$J(\tilde{\bar{p}}(t_f)) \to \sup\,. \tag{4}$$

Задав теперь ограничения на управления:

$$|\dot{u}_{ij}| < c_{ij},\;\; |\dot{\tilde{u}}| < c_{ij},\;\; i = 1, ..., N,\;\; j = 1, ...M$$

и рассмотрев последовательность решений задачи (2) — первый шаг — и (4) — второй шаг, — можно заметить, что

$$J(\tilde{\bar{p}}(t_f) \geqslant J(\bar{p}(0)). \tag{5}$$

Очевидно, что на втором шаге достигается по меньшей мере равенство функционалов, для чего достаточно обратить оптимальные управления, найденные на первом шаге:

$$\tilde{\bar{u}}_i(t) = \bar{u}_i^*(t_f - t),\ i = 1, ..., N.$$

Далее нас интересует реализация строгого неравенства

$$J(\tilde{\bar{p}}(t_f)) > J(\bar{p}(0)). \tag{6}$$

Примечательно, что (6) выполняется при весьма общих предположениях.

Потребуем хотя бы для одного $\tilde{u}_{ij}(0)$ двойственной обратной задачи (4) существования в R^{M} отрезка $(\tilde{u}_{ij}(0),\ \tilde{u}_{ij}(0) + \Delta u_{ij})$, смещение по которому ij-го начального управления монотонно улучшает функционал качества — при сохранении начального вектора состояния (для рассматриваемого класса f и J это попросту означает, что полученное на первом шаге — задача (2) — $\tilde{\bar{u}}(0) = (\tilde{\bar{u}}_1(0), \dots, \tilde{\bar{u}}_N(0)) = (\bar{u}_1(t_f), \dots, \bar{u}_N(t_f))$ не совпадает с локальным или глобальным оптимумом при данном начальном векторе состояния $\tilde{\bar{p}}(0) = \bar{p}(t_f)$). Теперь, обозначив приращение функционала через ΔJ и назначив, без ограничения общности, достаточно малое $\Delta u_{ij} > 0$, можно записать:

$$\Delta J(\Delta u_{ij}) \sim \Delta u_{ij}. \tag{7}$$

Отметим, что здесь и далее существенно используются непрерывная дифференцируемость f и J и конечность интервала управления $(0,\ t_f)$.

Предположим теперь, что для любого начального управления из $(\tilde{u}_{ij}(0),\ \tilde{u}_{ij}(0) + \Delta u)$ найдётся положительное $\varepsilon < c_{ij}$ и интервал управления $(0,\ \Delta t)$, в котором решение $\tilde{u}_{ij}^*(t)$ задачи (4) удовлетворяет неравенству

$$|\dot{\tilde{u}}_{ij}^*(t)| \leqslant c_{ij} - \varepsilon. \tag{8}$$

На плоскости $t \times \tilde{u}_{ij}(t)$ неравенство (8) констатирует, что при неоптимальном начальном векторе управления оптимальная траектория задачи (4) на $(0,\ \Delta t)$ содержится в меньшем секторе, симметрично вложенном в сектор кинематического ограничения $|\dot{\tilde{u}}_{ij}| < c_{ij}$.

Предположение (8) и исходное кинематическое ограничение позволяют, назначив на некотором интервале управления $(0,\ \tau)$ для начального условия $\tilde{u}_{ij}(0)$ значение $|\dot{\tilde{u}}_{ij}(t)| = c_{ij}$, реализовать слияние траектории из худшего начального условия с оптимальной траекторией лучшего начального условия. При этом $\tau \sim Du$.

Выписав далее необходимые вариации J, легко показать, что

$$\Delta J(\Delta u_{ij}) \sim (\Delta u_{ij})^2. \tag{9}$$

Поскольку ΔJ отсчитывается от лучшего значения, сравнение (7) и (9) доказывает, что предположение (8) неверно, — тем самым для решений задачи (4) при $\tilde{u}_{ij}(0)$, не совпадающем с локальным или глобальным оптимумом при

данном $\tilde{\tilde{p}}(0)$, доказано следующее: для любого положительного $\varepsilon < c_{ij}$ найдётся интервал $(0, \Delta t)$, в котором

$$|\dot{\tilde{u}}^*_{ij}(t)| > c_{ij} - \varepsilon. \tag{10}$$

Используя (10) и потребовав дополнительно только единственности управления, оптимального для данных начальных условий, мы доказываем (6) для всех управлений исходной задачи (2), принадлежащих произвольному сектору, строго содержащемуся в секторе исходных кинематических ограничений на управление.

Повторяя теперь приведённые рассуждения и условия реализации (6) для исходной задачи (2), мы получаем итеративный процесс (2), (4), (2), (4), ... с монотонным возрастанием $J(0) - J(t_f)$ исходной задачи. Отметим, что ограничение вторых производных $|\ddot{u}_{ij}|$ управлений исходной задачи (2) требует несимметричной модификации ограничений на управления в двойственной задаче (4). Доказательство монотонной сходимости процесса оптимизации в этом случае сложнее, но в целом проводится по приведённой схеме (существование предела $J(0) - J(t_f)$ и алгоритмы, сохраняющие или параллельно оптимизирующие начальный вектор состояния исходной задачи, требуют отдельного обсуждения.)

Отметим, что полученный результат позволяет строить простые геометрические примеры трёх типов решений исходной задачи оптимального управления:

– кинематические (в общем случае — динамические) ограничения реализуются на всём интервале управления;

– ограничения реализуются на правильном подмножестве интервала;

– ограничения не реализуются ни в одной точке интервала.

Приведём пример решения второго типа.

3. Пример неполной реализации динамических ограничений в задачах оптимального управления

Рассмотрим два взаимодействующих в R^3 подвижных объекта, неотрицательные характеристики которых, p и q, изменяются во времени в результате взаимодействия. Если взаимодействие объектов удовлетворяет принципу суперпозиции, то [1] можно разделить переменные — характеристики и координаты объектов — и записать:

$$\begin{cases} \dot{p} = pq\varphi_1(\bar{x}_1(t),\ \bar{x}_2(t)), \\ \dot{q} = pq\varphi_2(\bar{x}_1(t),\ \bar{x}_2(t)), \end{cases} \tag{11}$$

где не оговорён знак φ_1 и φ_2.

Естественно полагать функцию $\varphi_v,\ v = 1, 2$ симметричной, зависящей только от расстояния между подвижными объектами, т.е. $\varphi_v(\bar{x}_1;\ \bar{x}_2) = \varphi_v(\bar{x}_2,\ \bar{x}_1) = \varphi_v(|\bar{x}_1 - \bar{x}_2|)$.

В простейшем случае положительные функции $\varphi_v,\ v = 1, 2$ совпадают и монотонно убывают по $\rho = |\bar{x}_1 - \bar{x}_2|$. Приняв дополнительно $\dot{p} < 0,\ \dot{q} < 0$, приводим (11) к системе

$$\begin{cases} \dot{p} = -pq\varphi(\bar{x}(t),\ \bar{a}(t)), \\ \dot{q} = -pq\varphi(\bar{a}(t),\ \bar{x}(t)), \end{cases} \tag{12}$$

описывающей взаимодействие двух конкурирующих подвижных объектов, где $\overline{x}(t)$ – управляемая траектория объекта p, $\overline{a}(t)$ — заданная на интервале управления траектория объекта q.

Поставим задачу оптимального управления движением объекта p:

$$J(x^*) = \sup_{x \in X} J(x), \tag{13}$$

где

$$J(x) = q(0) - q(t_f) \tag{14}$$

для множества управлений

$$X = \{\bar{x}(t) || \dot{\bar{x}}(t)| \leqslant V,\ |\ddot{\bar{x}}(t)| \leqslant 2C;\ 0 \leqslant t \leqslant t_f\}\,.$$

В [3, 4] показано, что задача (12) – (14) эквивалентна задаче

$$\int_0^{t_f} \varphi(\bar{x}(t),\ \bar{a}(t))dt \underset{x \in X}{\rightarrow} \sup\,. \tag{15}$$

Зададим теперь произвольные начальные характеристики и несовпадающие начальные координаты объектов и удаляющееся движение объекта q по соединяющей объекты прямой со скоростью меньшей V. Используя (15), легко показать, что для данных условий найдётся бесконечное множество t_f, для которых максимизирующее (14) прямолинейное движение объекта p будет состоять из двух отрезков:

1) преследование объекта q до слияния с ним — с реализацией динамических ограничений;

2) сопровождение объекта q до момента t_f — с его скоростью меньшей V.

Благодарности

Работа выполнена при финансовой поддержке РФФИ, проекты № 14-08-01132 и № 14-07-00272.

Литература

1. Nartov B.K. Conflict of Moving Systems. France : AMSE Press, 1994. 87 p.
2. Нартов Б.К. Об одном методе оптимизации начальных условий в управлении динамическими системами // Математические структуры и моделирование. 2002. Вып. 9. С. 1–3.
3. Нартов Б.К. Методы траекторного управления. Новосибирск : Наука, 2003. 104 с.
4. Лебедев Г.Н., Мирзоян Л.А., Нартов Б.К., Чуканов С.Н. Управление подвижными объектами. Оперативное планирование. М. : Научтехлитиздат, 2008. 136 с.

METHOD OF RETURN AND IMPLEMENTATION OF DYNAMIC LIMITS IN THE OPTIMAL CONTROL PROBLEM

B.K. Nartov
Ph.D.(Phys.-Math.), Senior Scientist Researcher, e-mail: nartov@ofim.oscsbras.ru

Omsk Branch of Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Science, Omsk

Abstract. Presented in the paper method, designed to directionally optimize the initial conditions in problems of dynamic systems management, — a method of return — was originally designed to optimize the vector of initial coordinates in a particular model of moving objects conflict whose characteristics deteriorated as a result of interaction with the objects of the opponent and aging. The model was binding characteristics (state vector) and coordinates (control vector) of objects by Lanchester-type differential equations. Then the specific problems of optimal control of the movements of a group of objects, opposing another group of objects with the specified paths in the control interval (by the criterion of minimizing a function of the final states of objects), were set and solved. The problem of the construction of an acceptable on time and accuracy algorithm to optimize the initial control vector, i.e. the initial placement of the group of managed objects, was much more complicated. The found approach has been very general and allows us to directionally optimize the initial control vector, at least in the class of managed smooth systems with continuously differentiable quality functional. In the most general form the idea of the method is that for optimization, in terms of selected quality functional, of the initial conditions of the original optimal control problem the supporting dual problem is written and the iterative process is implemented, which steps alternate original and dual problems, and as part of the initial conditions of the next iteration the part of the final values of the previous iteration is using.

Keywords: dynamic systems, optimization of the initial conditions, the inverse problem, the dynamic limits.

Дата поступления в редакцию: 15.11.2015

Математические структуры и моделирование
2016. № 4(40). С. 65–71

УДК 530.145:316

КВАНТОВЫЙ ПОДХОД К ОПИСАНИЮ СОЦИАЛЬНОЙ СТАТИКИ И СОЦИАЛЬНОЙ ДИНАМИКИ ОГЮСТА КОНТА

А.К. Гуц
профессор, д.ф.-м.н., и.о. заведующего кафедрой социологии ОмГУ,
e-mail: aguts@mail.ru

Омский государственный университет им. Ф.М. Достоевского

Аннотация. Предложен квантовый подход, реализующий идеи Огюста Конта о социальной статике и социальной динамике. Использован аппарат квантовой космологии, позволяющий описывать социальную физику.

Ключевые слова: Огюст Конт, социальная статика, социальная динамика, социальная физика.

Введение

Социология как наука появилась первоначально под названием «социальная физика». И дело здесь в том, что Огюст Конт хотел строить общую теорию общества, т. е. исследование тех законов, которыми управляются явления общественной жизни [1, с. 4], подобно тому, как это делалось в естествознании и физике, в частности. Успехи естествознания были связаны с тем, что оно ограничило себя одним миром явлений, не задаваясь разрешением вопроса о лежащей в их основе свехчувственной сущности. Мир же явлений оно изучает путём опыта и наблюдений [1, с. 5].

Конт свою социологию, свою социальную физику разделил на статику и динамику. Образцом статики были такие дисциплины как политика, юриспруденция, политическая экономика — они имели своим предметом общество, как нечто раз и навсегда данное, а не постоянно развивающееся. Тогда же его современники использовали уже понятие «процесс общественного развития», и это привело Конта к констатации наличия социальной динамики [1, с. 7].

Однако, если социальная статика — это выявление закономерностей в обществе как всегда самому себе равного предмета, и это породило представление об общественном порядке, лежащем в основе общественного бытия [1, с. 7], т. е. идея социальной статики шла от дисциплин практичных, прагматичных, устоявшихся, прослеживаемых во всех исторических эпохах и определяющих стабильный общественный быт, то представление о социальной динамике было навеяно Конту философией истории, которую следует отнести к дисциплинам спекулятивным, метафизичным, так или иначе предполагающим существование свехчувственной сущности, именуемой часто *течением объективного времени*.

Социальная статика — это теория социальной анатомии, теория социального равновесия. Социальная динамика выясняет вопросы общественного развития. Статические законы выявляют взаимодействия между одновременными явлениями; динамические — между последовательными, не принимая во внимания для последних, наличие или отсутствие причинно-следственных связей [1, с. 17–18].

В данной статье мы попытались найти математический аппарат, который равным образом объединял как социальную статику, так и социальную динамику, но при этом основывался на вероятностных, недетерминистических характеристиках социальных явлений.

1. Социальная статика. Исторические эпохи

1.1. Уравнения социальной статики

Общественная жизнь наблюдается нами *в изменениях* и в окружении Внешего мира, называемого Природой или Вселенной. Мы не можем отрывать людей от этого окружения и поэтому должны сказать, что общество существует в *пространстве-времени*.

Следовательно, для единого описания общества и Природы надо воспользоваться, в духе Конта, естественнонаучным подходом. Нам надо породить и пространство-время M^4, и общественное бытие в нём. Точнее, бытие в 3-мерном пространстве, в котором «течёт время», идут изменения, текут общественные процессы. Воспользуемся квантовой теорией.

Пространство-время Вселенной M^4 в квантовой космологии Уилера-ДеВитта появляется как интерференция когерентной квантовой суперпозиции, или волнового пакета:

$$\Psi[{}^{(4)}\mathcal{G}, \mu, B, e, \sigma, \nu] = \int\limits_K c_k \Psi_k[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu] dk, \quad c_i \in \mathbb{C}, \tag{1}$$

где $\Psi_k[{}^{(3)}\mathcal{G}]$ — частная волновая функция, являющаяся функционалом от 3-мерной римановой геометрии ${}^{(3)}\mathcal{G} = (M^3, h_{\alpha\beta})$ и удовлетворяющая функциональному уравнению ДеВитта-Уилера [2].

$$\left[G_{\alpha\beta\gamma\delta} \frac{\delta}{\delta h_{\alpha\beta}} \frac{\delta}{\delta h_{\gamma\delta}} + \sqrt{h}\ {}^{(3)}R + \right.$$

$$\left. + \mathcal{E}(h_{\alpha\beta}, \mu, B, e, \sigma, \nu) \right] \Psi[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu] = 0, \tag{2}$$

где $\mathcal{E}(h_{\alpha\beta}, \mu, B, e, \sigma, \nu)$ — член, учитывающий вклад материальных источников μ, окружающей среды (природы) B и социальных полей e, σ и ν.

К этому уравнению нужно добавить уравнения для материальных источников μ, окружающей среды (природы) B и полей e (этносфера), σ (социосфера) и ν (ноосфера) [3].

Мы видим, что то, что считается Реальностью, существующей в *форме* четырёхмерного непрерывного континуума M^4, называемого пространством-временем, в действительности является существенно квантовой сущностью, т. е. цепью интерференционных «горных пиков» по выражению Halliwell'a [4] в суперпространстве Уилера. В двумерной модели, к примеру, волновая функция $\Psi[{}^{(4)}\mathcal{G}, \mu, B, e, \sigma, \nu]$ будет состоять из резко взметнувшихся горных пиков в минисуперпространстве вдоль единственной классической траектории (пространства-времени).

Как правило, не обсуждается смысл системы Ω, описываемой посредством волнового пакета (1), и её состояний $\Omega_k,\ k \in K$, для которых находятся соответствующие волновые функции $\Psi_k[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu]$.

Очевидно, что Ω — это Внешний мир, Квантовая реальность, а её состояния Ω_k — это формы её существования, которые в соответствии с принципами квантовой механики в процессе, именуемом в квантовой механике *наблюдением* (измерением), локализуются. Квантовая механика, а значит описываемая ею Квантовая реальность, не может обойтись без сознающих личностей, называемых физиками *наблюдателями*. *Наблюдение* системы Ω приводит к *коллапсу* волнового пакета (1):

$$\int\limits_K c_k \Psi_k[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu] d\mu(k) \to \Psi_{k'}[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu] \tag{3}$$

с вероятностью $|c_{k'}|^2$.

Наблюдения Вселенной людьми, живущими в конкретное время, в своей эпохе, переводят Вселенную в наблюдаемое состояние Ω_k. Какой смысл несут состояния Ω_k? Вполне предсказуемый – они задают социальную статику. Иначе говоря, определяют статичное неизменяемое общественное бытие.

1.2. Исторические эпохи

Примем, что каждое состояние Ω_k квантовой реальности Ω — это 3-мерный мир, в котором практически ничего не меняется; он вневременен. В этом мире находится *наблюдатель*, способный осуществлять измерения реальности, точнее, её пространственной геометрии. Иначе говоря, Ω_k — это стационарное пространство-время M^4_k, в котором осуществляется «замороженное» историческое общественное бытие. Историки такое существование называют историческими эпохами. Гёте и Шпенглер использовали термин «гештальт» [5, 6].

Каждая историческая эпоха, такая как Античность, Средневековье, Возрождение и пр. видится историками как *ограниченная* во времени форма существования человечества.

> Историческая эпоха — наиболее крупная единица исторического времени, обозначающая длительный период человеческой истории, отличающийся определённой внутренней связностью и только ему присущим уровнем развития материальной и духовной культуры ...

> Переход от одной эпохи к другой представляет собой переворот во всех сферах социальной жизни [Философский словарь].

Конечность исторической эпохи автоматически означает её сменяемость, а значит даёт возможность все исторические эпохи разместить одну за другой, **последовательно** в одном пространственно-временном лоренцевом многообразии. В этом отражена западная культурная традиция – видеть Мир изменяющейся сущностью, эволюционирующей в физическом времени t.

Ну..., а вдруг исторические эпохи не конечны во времени? И если не пытаться их втолкнуть в одно пространство-время, полагая, что бесконечность во времени проявляется всего лишь в форме редких, но устойчивых «пережитков» прошлого? В таком случае очевидным становится, что сильно доминирующие нередкие «пережитки» прошлого будут заполнять всё будущее, разрушая идею сменяемости исторических эпох, идею эволюционирующей реальности.

Как спасти идею последовательной сменяемости исторических эпох, идею эволюционирующей реальности, не пренебрегая при этом условием бесконечности статичного существования во времени каждой исторической эпохи?

Очевидно, для этого надо использовать не классическую теорию, а квантовую, и тогда эволюционирующий Мир появляется как интерференция исторических эпох, как последовательность «горных пиков», в высоту каждого из которых вносит вклад каждая историческая эпоха, как квантовый волновой пакет исторических эпох в форме (1). Удивительно, но при этом все исторические эпохи *существуют одновременно*.

2. Социальная динамика. Историческая последовательность

Как в предложенном формализме реализуется идея социальной динамики?

Время, текущее в пространстве-времени M^4, появляется извне, искусственно. Его нет в квантовой космологии ДеВитта-Уилера [8].

Просто вдоль цепи «горных пиков», обозначающих классическую траекторию и являющихся тем, что мы называем пространством-временем M^4, вводится искусственно расстояние между ними — воспринимаемое людьми как *физическое время* t. Поэтому имеем семейство 3-геометрий ${}^{(3)}\mathcal{G}(t)$, или 3-метрик $h_{\alpha\beta}(x,t)$, удовлетворяющих уравнениям Эйнштейна [6]. Интересно, что 3-геометрия каждой исторической эпохи, формирующей «цепь горных пиков», совершенно однозначно находит своё место в качестве пространственного сечения пространства-времени, и время t указывает место локализации этой 3-геометрии в 4-геометрию. В этом смысле, как пишет Уилер, «3-геометрия выступает как «носитель временно́й информации» [2, с. 37].

2.1. Уравнение, описывающее социальную динамику

Таким образом, наблюдается динамика как 3-геометрии, так и общественной жизни. Найдём уравнение, описывающее социальную динамику.

Рассматривая волновую функцию

$$\Psi[h_{\alpha\beta}(x,t),\mu,B,e,\sigma,\nu]=\Psi[{}^{(3)}\mathcal{G}(t),\mu,B,e,\sigma,\nu]$$

и полагая

$$\Psi[h_{\alpha\beta}(x,t),\mu,B,e,\sigma,\nu]=\psi[h_{\alpha\beta}(x,t)]e^{im_P S[h_{\alpha\beta}(x,t),\mu,B,e,\sigma,\nu]}, \tag{4}$$

$$\psi(t)=\psi[h_{\alpha\beta}(x,t),\mu,B,e,\sigma,\nu],$$

$$\frac{\partial}{\partial t}\psi(t)=\int \dot{h}_{\alpha\beta}(x,t)\frac{\delta}{\delta h_{\alpha\beta}(x,t)}\psi[h_{\alpha\beta}(x,t),\mu,B,e,\sigma,\nu]d^3x,$$

где $S[h_{\alpha\beta},\mu,B,e,\sigma,\nu]$ — решения уравнения Гамильтона-Якоби, m_P — масса Планка, и

$$\dot{h}_{\alpha\beta}=NG_{\alpha\beta\gamma\delta}S[h_{\gamma\delta},\mu,B,e,\sigma,\nu]+2D_{(\alpha}N_{\beta)},$$

находим, что вдоль пространства-времени, т. е. вдоль цепи «горных пиков» справедливо уравнение Шрёдингера

$$i\hbar\frac{\partial}{\partial t}\psi(t)=H_{mat}\psi(t), \tag{5}$$

где H_{mat} — гамильтониан материальных полей (подробности см. [7, p. 172]).

Заметим, что социальная динамика, поскольку она описывается уравнением Шрёдингера, является *вероятностной*. Мы естественным образом отошли от классического детерминизма, и это позволяет учитывать человеческую непредсказуемость.

В формуле (4) — это формализм так называемого полуклассического приближения, позволяющий получить уравнение Гамильтона-Якоби и реализовать идею получения классического пространства-времени как «цепи горных пиков», т. е. как результат интерференции. В случае космологии обычно считается, что все эти расчёты относятся к раннему этапу развития Вселенной, когда она имела крайне малые размеры (отсюда число m_P в формуле (4)). Для социологии появление в теоретических рассуждениях таких выражений как «планковские константы», «ранние этапы существования Вселенной» не только недопустимо, но граничит с безответственными спекуляциями. Но в действительности эти расчёты относятся к этапам созидания пространства и времени как наличного бытия из чистого бытия, которое есть чистое ничто (см. подробности в [9, с. 13]), происходящее *в малом повсеместно и всевременно*. А из множества малого складывается пространство исторической эпохи целиком вместе с его геометрией ${}^{(3)}\mathcal{G}_k$. Сама же эта геометрия несёт память о физическом времени исторической эпохи [2, с. 37]. Просто, думая о людях, т. е. находясь в рамках социологии, надо забыть о теории Большого взрыва и поставить во главу размышлений *потребность человека* в жизненном пространстве, а не заталкивать его, наравне с курами и коровами, в какой-то момент в возникшие независимо от него большие пространственные объёмы. Другими словами, социология нуждается совсем в иной, новой космогонии [9], которая идёт от человека, от сознания, а не от сингулярности и элементарных частиц, не нуждающихся в человеке (в сознании) миллиарды лет.

2.2. Почему наблюдаются изменения?

Благодаря наличию интерференционной картины — цепи «горных пиков», — существует классическое пространство-время, которое видится живущим в нём людям (наблюдателям) как «эволюционирующее», поскольку содержит вклады всех исторических эпох. Это видно в случае полуклассичекого приближения волнового пакета: если взять

$$\Psi_k[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu] = A_k e^{\frac{t}{\hbar} S_k[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu]},$$

то

$$\int\limits_K c_k \Psi_k[{}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu] d\mu(k) = \left(\int\limits_K c_k A_k d\mu(k) \right) e^{\frac{t}{\hbar} S_0}, \tag{6}$$

где

$$\forall k (S_k({}^{(3)}\mathcal{G}, \mu, B, e, \sigma, \nu) = S_0 = const)$$

– условие интерференции. Из (6) видно, как «горные пики» складываются из разных интерферирующих эпох. Благодаря этому втиснутые в единое пространство-время люди рассуждают о наблюдаемых сменах исторических эпох, помнят своих предков, раскапывают исторические артефакты и прочее. При этом каждый из этих людей принадлежит конкретной исторической эпохе Ω_k, поскольку состояниями квантовой системы Ω являются эпохи, а не интерференция в форме пространства-времени (цепи «горных пиков»).

2.3. Динамика в смене статики

Цепь «горных пиков» называем *исторической последовательностью*. В ней течёт время t, идут общественные процессы, есть всё то, что присуще социальной динамике Огюста Конта, которая получает у него «характер не исследования законов, которыми управляется последовательность общественных явлений везде и всегда, а характер философского изображения действительных судеб человечества, т. е. может быть скорее названа философской историей, а не социологией» [1, с. 21].

Статичное существование в рамках конкретной исторической эпохи возможно, но при условии отсутствия других исторических эпох, интерферирующих с данной. Поскольку мы фиксируем изменения, наблюдаем социальную динамику, то это говорит о том, что мы находимся в «цепи горных пиков», и, следовательно, пребываем в квантовой реальности, в квантовой суперпозиции. Во всяком случае, это говорит о том, что другие исторические эпохи существуют и в принципе в них можно уйти [10], совершив какое-то особое их наблюдение, или как говорят физики, произведя некоторое измерение или даже всего лишь утвердившись в намерении совершить такие измерения.

3. Заключение

Конт спрашивал: «Если статический анализ нашего общественного организма показывает, что в конце концов по всей необходимости он покоится

на некоторой системе основных мнений, то каким образом постепенные изменения такой системы могли бы не оказывать преобладающего влияния на последовательные изменения, какие представляет собой непрерывная жизнь человечества? (цит. по [1, с. 22]).

Предложенный в статье квантовый подход к социологии даёт такой ответ на этот вопрос: изменения внутри статичной исторической эпохи не оказывают влияние на динамику непрерывной жизни людей, но последовательные изменения в этой жизни происходят в силу того, что существуют другие отличные исторические эпохи, которые вносят свой вклад в историческую последовательность через квантовую интерференцию.

Литература

1. Каревъ Н. Введеніе въ изученіе соціологіи. С.-Петербургъ, 1897.
2. Уилер Дж. Предвидение Эйнштейна. М. : Мир, 1970.
3. Гуц А.К., Паутова Л.А. Глобальная этносоциология. Изд. 2, доп. М. : Книжный дом «ЛИБРОКОМ», 2009. 236 с.
4. Halliwell J.J. Introductiry lectures on quantum cosmology // In: Quantum cosmology and baby universes / Eds. S. Coleman, J.B. Hartle, T. Piian and S. Weinberg. World Scientific Publishing Co. Pte. Ltd., 1991. P. 159–244.
5. Гуц А.К. Многовариантная Вселенная и теория исторических последовательностей // Математические структуры и моделирование. 2012. № 25. С. 70–80.
6. Гуц А.К. Физика реальности. Омск : Изд-во КАН, 2012. 424 с.
7. Kiefer C. Quantum Gravity. Second Edition. — Oxford University Press, 2007. 361 p.
8. Barbour J. The nature of time. URL: `http://arxiv.org/pdf/0903.3489v1.pdf` (2009).
9. Гуц А.К. Метафизика теоретической истории // Метафизика (РУДН). 2015. № 4(18). С. 9–30.
10. Гуц А.К. Негёделевская машина времени // Математические структуры и моделирование. 2016. № 3(39). С. 47–55.

QUANTUM APPROACH TO DESCRIPTION OF SOCIAL STATICS AND SOCIAL DYNAMICS OF AUGUSTE COMTE

A.K. Guts
Dr.Sc. (Phys.-Math.), Professor, e-mail: aguts@mail.ru

Dostoevsky Omsk State University

Abstract. A quantum approach to description of the Auguste Comte's ideas on social statics and social dynamics is given. We use the apparatus of quantum cosmology, which allows us to describe the social physics.

Keywords: Auguste Comte, social statics, social dynamics, social physics.

Дата поступления в редакцию: 31.07.2016

Математические структуры и моделирование
2016. № 4(40). С. 72–80

УДК 004.932.2

ПРИМЕНЕНИЕ МЕТОДА ДИФФЕОМОРФНОГО ПРЕОБРАЗОВАНИЯ КРИВЫХ ПРИ РЕШЕНИИ ЗАДАЧ РАСПОЗНАВАНИЯ ОБРАЗОВ

С.Н. Чуканов[1]
профессор, д.т.н., ведущий научный сотрудник, e-mail: ch_sn@mail.ru
Д.Б. Абрамов[2]
аспирант, e-mail: cuntz@mail.ru
С.О. Баранов[2]
аспирант, e-mail: serj@doctor.com
С.В. Лейхтер[2]
аспирант, e-mail: leykhter@mail.ru

[1]ФГБУН Институт математики им. С.Л. Соболева СО РАН, Омский филиал

[2]ФГБОУ ВО Сибирская автомобильно-дорожная академия

Аннотация. Рассмотрена задача оценивания нормы расстояния между двумя замкнутыми гладкими кривыми при распознавании образов. Описаны диффеоморфные преобразования кривых на основе модели больших деформаций. Для оценивания нормы расстояния между двумя замкнутыми кривыми формируется функционал, соответствующий норме расстояния между двумя кривыми, и уравнение эволюции диффеоморфных преобразований. Предложен алгоритм решения уравнения диффеоморфного преобразования, построенный на основе метода PSO, который позволяет значительно сократить объём вычислительных операций по сравнению с градиентными методами решения. Разработанные в статье алгоритмы могут использоваться в биоинформатике и биометрических системах, классификации изображений и объектов, системах машинного зрения, при распознавании образов и объектов, системах трекинга.

Ключевые слова: распознавание образов, инвариантность, диффеоморфные преобразования, биометрия, метод PSO.

Введение

Распознавание объектов по изображениям независимо от их расположения, ориентации, масштаба и перспективы — является важным направлением информационных технологий в области распознавания образов и машинного зрения. В задачах математической морфологии важной является задача сопоставления близких форм, а не точное определение каждой формы; деформация сложной фигуры может привести к пониманию формы. Изучение формы и изменчивости изображения в рамках теории распознавания образов можно свести к оцениванию преобразований, которые последовательно деформируют изображения.

Вычисление многомерных нежёстких преобразований изображений привело к развитию стратегии эластичного сравнения, при этом преобразование линеаризуется относительно системы координат исходного изображения и генерируется векторное поле смещений. Стоимость преобразования измеряется функционалом — нормой разности между преобразованным исходным изображением и эталонным изображением; оптимальному преобразованию этого функционала соответствует векторное поле смещений с наибольшей гладкостью. Измерение гладкости достигается указанием нормы в пространстве векторных полей с использованием дифференциального оператора. Одним из ограничений данного подхода является то, что не гарантируется биективность преобразования. Представляет интерес вычисление диффеоморфных преобразований, которые сами являются гладкими, но и обратные преобразования сохраняют свойства гладкости. Модель больших деформаций для вычисления преобразований изображений [1] гарантирует, что преобразования, вычисленные между изображениями, диффеоморфны. При этом преобразование исходных точек области в требуемые формируется на основе зависящего от времени векторного поля скоростей, которое определяется системой обыкновенных дифференциальных уравнений (ODE).

В работе рассмотрена задача оценивания нормы расстояния между двумя замкнутыми гладкими кривыми при распознавании 2D-образов. Рассмотрены действия групп переноса, вращения и масштабирования на 2D замкнутую кривую, инварианты к действию этих групп. Для оценивания нормы расстояния между кривыми положение кривых нормализуется центрированием, приведением главных осей инерции изображения к осям системы координат и приведением к единице площади замкнутой кривой соответствующим масштабированием. Для оценивания нормы расстояния между двумя замкнутыми кривыми формируется функционал, соответствующий норме расстояния между двумя кривыми, и уравнение эволюции диффеоморфных преобразований. Предложен алгоритм решения уравнения диффеоморфного преобразования, построенный на основе метода PSO, который позволяет значительно сократить объём вычислительных операций по сравнению с градиентными методами решения.

Разработанные алгоритмы могут использоваться в биоинформатике и биометрических системах, классификации изображений и объектов, системах машинного зрения, нейровизуализации, при распознавании образов и объектов, системах трекинга. Алгоритм оценивания нормы расстояния между замкнутыми кривыми методом диффеоморфного преобразования может быть распространён на пространственные объекты (кривые, поверхности, многообразия).

1. Построение инвариантов переноса, вращения и масштабирования

Для нахождения инвариантов при распознавании образов необходимо найти группу G, действующую на множестве аргументов функции изображения. Изображение объекта может быть описано функцией $f(x, y) = 1$, если $(x, y) \in S \subset \mathbb{R}^2$ ($f(x, y) = 0$, иначе), где (x, y) — декартовы координаты

изображения с границей $c = \partial S$ множества S. Действие группы переноса на функцию $f(x,y)$ в направлении оси X: $g_{\varepsilon_x} f(x,y) = f(x+\varepsilon_x, y)$; оси Y: $g_{\varepsilon_y} f(x,y) = f(x, y+\varepsilon_y)$.

Действие группы масштабирования на функцию $f(x,y)$: $g_{\phi_s} f(x,y) = f((1+\phi_s)x, (1+\phi_s)y)$.

Действие группы вращения (поворот на угол α): $g_\alpha f(x,y) = f(x\cos\alpha - y\sin\alpha, x\sin\alpha + y\cos\alpha)$.

Инвариантность по отношению к группе переноса может быть обеспечена нахождением центра (x_0, y_0) с последующим переносом. Для действия группы переноса на функцию 2D-изображения нахождение центра изображения сводится к методу моментов [2]. Сформируем моменты порядка $(p+q)$ 2D-функции $f(x,y)$: $m_{p,q} = \int\limits_S x^p y^q f(x,y)\, dS; p, q \in \mathbb{Z}^+$; например, площадь изображения: $m_{0,0} = \int\limits_S f(x,y)\, dS$.

Центр (x_0, y_0) функции изображения $f(x,y)$ определяется из соотношений: $x_0 = m_{1,0} m_{0,0}^{-1}$; $y_0 = m_{0,1} m_{0,0}^{-1}$. Центрированная функция является инвариантной по отношению к действию группы переносов. Нормализованные моменты: $f(x+x_0, y+y_0)$ являются инвариантами масштабирования. Подействуем на $f(x,y)$ таким элементом группы масштабирования g_{ϕ_s} , что значение будет $m_{0,0} = 1$.

Для выделения определённой ориентированной системы координат построим тензор изображения: $J = \begin{pmatrix} m_{2,0} & -m_{1,1} \\ -m_{1,1} & m_{0,2} \end{pmatrix}$. При повороте объекта с матрицей направляющих косинусов $T = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$ тензор инерции изменяется по закону: $J' = T^T \cdot J \cdot T$. При повороте объекта на угол: $\alpha = 0,5 \cdot \operatorname{arctg}\left(2J_{xy}(J_{yy} - J_{xx})^{-1}\right)$, тензор инерции будет иметь диагональный вид $J^d = \operatorname{diag}\begin{pmatrix} J_x & J_y \end{pmatrix} \in \mathbb{R}^{2\times 2}$, где J_x, J_y — собственные числа тензора инерции J. При $J_x \neq J_y$ можно провести такое преобразование координат: $\begin{pmatrix} x' & y' \end{pmatrix}^T = T \begin{pmatrix} x & y \end{pmatrix}^T$ — формированием поворота T, что оси X, Y будут направлены по главным осям тензора инерции 2D-изображения.

Для нормализации изображения необходимо решить задачу нахождения центра изображения и выделенной ориентации группы вращения с последующим центрированием и масштабированием изображения.

2. Действие элементов групп на кривые

Действие матричных групп на кривые можно представить: $A \to A \circ c$, где $A : \mathbb{R}^2 \to \mathbb{R}^2$ — действие матричной группы в $\mathbb{R}^2$. Приведём примеры матричных групп [3].

- $GL(2)$ — линейная группа матриц $GL(2) = \{A \in \mathbb{R}^{2\times 2}; \det A \neq 0\}$ с законом композиции — умножением матриц.

- $SO(2) \in GL(2)$ — специальная ортогональная группа может быть представлена матрицами $SO(2) := \left\{A \in \mathbb{R}^{2\times 2} | AA^T = A^T A = \mathrm{Id}; \det(A) = 1\right\}$.
- Группа масштабирования может быть представлена диагональными матрицами $A = \begin{pmatrix} \rho & 0 \\ 0 & \rho \end{pmatrix}, \rho \in \mathbb{R}^+$.
- $SE(2)$ — специальная группа Евклида определяется полупрямым произведением $SE(2) \simeq SO(2) \otimes \mathbb{R}^2$.

Дифференцируемая кривая в $GL(2)$ — это функция: $g_t : (a,b) \to GL(2)$, для которой существует производная ${}^{dg_t}/_{dt}; \forall t \in (a,b)$. Уравнение первого порядка для элемента матричной группы: ${}^{dg_t}/_{dt} = A \cdot g_t;\ g_{t=0} = \mathrm{Id}$, где $A \in \mathbb{R}^{2\times 2}$ — матрица с постоянными элементами имеет решение: $g_t = \exp(tA)$, которое обладает групповыми свойствами.

Кривая, соединяющая элементы $g_0, g_1 \in GL(2)$, минимизирующая функционал:

$$\int_0^1 \|v_t\|_V^2 \, dt = \int_0^1 \langle Lv_t, v_t \rangle_2 \, dt; L \in \mathbb{R}^{2\times 2} \tag{1}$$

и удовлетворяющая уравнению ${}^{dg_t}/_{dt} = v_t \cdot g_t$, является решением уравнения Эйлера [4]:

$$d(Lv_t)/dt = (Lv_t) v_t^* - v_t^* (Lv_t). \tag{2}$$

3. Группа диффеоморфных преобразований

Будем считать, что замкнутые кривые принадлежат открытому подмножеству $X \subset \mathbb{R}^2$. Диффеоморфизм X является обратимым непрерывно дифференцируемым преобразованием $X \to X$; существует тождественное отображение (Id — композиция прямого и обратного диффеоморфизма). Множество диффеоморфизмов $\mathrm{Diff}(X)$ определяет структуру группы. Диффеоморфизмы изменяют количественные характеристики объектов, которые определены на множестве X. Матричные группы диффеоморфизмов имеют конечную размерность и кодируются с помощью параметров матриц. Рассмотрим группу бесконечномерных диффеоморфизмов, действующих на ограниченном множестве $X \subset \mathbb{R}^2$. Определим диффеоморфизм $g : X \to X$ с обратным элементом g^{-1} и определим группу преобразований G, как подгруппу диффеоморфизмов с законом композиции $\circ$: $g \circ g' = g(g') \in G$. Для формирования диффеоморфных отображений диффеоморфизмы рассматриваются как потоки ODE. Предположим, диффеоморфизмы $g_t(x); x \in X$ эволюционируют во времени $t \in [0,1]$ с векторным полем $v_t(\cdot)$:

$$dg_t(x)/dt = v_t(g_t(x)); g_0(x) = x. \tag{3}$$

Формированием требуемого векторного поля $v_t(\cdot)$ в любой момент времени $t \in [0 \ldots 1]$ можно добиться такого действия элементов группы $g_t(\cdot)$ на точки пространства $X \subset \mathbb{R}^2$, что $g_0(x) = x; g_1(x) = y; \forall x, y \in X$.

Допустим, что задана норма $\|v_t\|_V^2 = \langle Lv_t, v_t\rangle_2 = \int\limits_S (Lv_t)^* v_t dS$, где $\alpha_t = Lv_t; t \in [0,1]$ — момент векторного поля. Для $g_t \in G$ существуют скорости $v_t(g_t) = dg_t/dt$, минимизирующие функционал:

$$\Phi(v_t) = \int\limits_0^1 \|v_t\|_V^2 \, dt = \int\limits_0^1 \langle Lv_t, v_t\rangle_2 \, dt \tag{4}$$

на траектории, соединяющей элементы группы $g_0 = g|_{t=0}$ и $g_1 = g|_{t=1}$. Представим обратную связь между скоростью v_t и моментом α_t в форме:

$$v_t = L^{-1}\alpha_t = K\alpha_t. \tag{5}$$

Для дифференциального оператора: $L = \mathrm{id} - a\nabla^2$ в $\mathbb{R}^2$ — обратный оператор $K = L^{-1}$ аппроксимируем функцией:

$$K(x) = \beta e^{-\gamma^{-1}\|x\|^2}. \tag{6}$$

Уравнения эволюции диффеоморфизмов Эйлера-Пуанкаре можно получить решением уравнений вариационной задачи с функционалом $\Phi(v_t)$ [5]:

$$d\alpha_t/dt = -(D\alpha_t)\, v_t - \alpha_t \nabla v_t - (Dv_t)^T \alpha_t, \tag{7}$$

где $Df = (\partial f_i/\partial x_j); i,j = 1,2$. Если объектами являются точечные множества, то векторные поля в точках $x_t = (x_1(t), \ldots, x_N(t))$ принимают вид: $v_t(\cdot) = \sum\limits_{l=1}^{N} K(\cdot, x_l)\,\alpha_l$.

Уравнение вариационной задачи позволяет перемещать объекты вдоль траекторий, которым соответствуют диффеоморфные преобразования. Диффеоморфизмы не позволяют изменить топологию вдоль геодезических траекторий. Неточный вид диффеоморфизмов [6, 7] обеспечивает механизм, который позволяет при эволюции геодезических траекторий отклоняться от точных деформаций. В задаче неточного сравнения минимизируемый функционал содержит член, который оценивает точность попадания точек $g_1(x_n^0); n = 1, \ldots, N$ в требуемые позиции x_n^1:

$$\int\limits_0^1 \|v_t\|_V^2 \, dt + \sigma^{-2} \sum_{n=1}^{N} \left\| x_n^1 - g_1\left(x_n^0\right)\right\|^2, \tag{8}$$

при этом в уравнения Эйлера-Пуанкаре диффеоморфных преобразований вводится параметр σ^2:

$$\begin{aligned} &dx_k/dt - v_t(x_k) = \sigma^2\alpha_k; \\ &v_t(\cdot) = \sum_{l=1}^{N} K(\cdot, x_l)\,\alpha_l; d\alpha_k/dt = -\sum_{l=1}^{N} \nabla_1 K(x_k, x_l)\,\alpha_k^T\alpha_l, \end{aligned} \tag{9}$$

здесь $\nabla_1 K$ представляет собой градиент функции $(x,y) \to K(x,y)$ по отношению к первой координате. Примем для оператора L функцию $K(x,\cdot)$ в виде: $K(x,\cdot) = e^{-\gamma^{-1}\|x-(\cdot)\|^2}$. Тогда:

$$\nabla_1 K(x_k, x_l) = -2\gamma^{-1}(x_k - x_l)\, e^{-\gamma^{-1}\|x_k - x_l\|^2}.$$

4. Решение задачи методом PSO

При решении уравнения (9) необходимо определить краевые условия $\alpha_0 = (\alpha_1(0), \ldots, \alpha_N(0))$ и $\alpha_1 = (\alpha_1(1), \ldots, \alpha_N(1))$ при известных $x_0 = (x_1(0), \ldots, x_N(0))$ и $x_1 = (x_1(1), \ldots, x_N(1))$. Применение градиентных методов решения задачи (9) требует значительного количества вычислительных операций. Для решения этой задачи в работе предлагается применение метода пристрелки (shooting) с использованием алгоритма PSO (particle swarm optimization). Метод пристрелки заключается в нахождении такого начального вектора $\alpha_0 = (\alpha_1(0), \ldots, \alpha_N(0))$, что значение функционала (8) минимизируется.

Метод PSO основан на имитации поведения роя насекомых и был предложен J. Kennedy в 1995 году [8]. В контексте многопараметрической оптимизации рой (swarm) имеет фиксированный размер; каждая частица первоначально расположена в случайных местах в многомерном пространстве проектирования. Частицы имеют две характеристики: положение и скорость. Положение частицы определяется значением целевой функции. Частицы обмениваются информацией (лучшими позициями) и могут корректировать свои позиции и скорости. Алгоритм метода PSO приведён в приложении.

Пример. Рассмотрим пример диффеоморфного преобразования замкнутой кривой — окружности единичного радиуса (эллипс с эксцентриситетом $\varepsilon = 0$ и длиной окружности 2π) в отрезок прямой длиной π (эллипс с $\varepsilon = 1$) за единичный период времени. Для этого выберем $N = 12$ точек на эллипсе, соответствующих параметру $\theta_i = 2\pi i N^{-1}; i = 1, \ldots, N$. Выберем параметр уравнения диффеоморфных преобразований: $\sigma^2 = 10^{-4}$; параметр метод PSO: $\vartheta = 0,7$; число частиц: 10. В таблице 1 представлены результаты моделирования диффеоморфных преобразований точек эллипса от значения эксцентриситета $\varepsilon = 0$ до $\varepsilon = 1$ для четырёх точек (из 12) замкнутой кривой.

Таблица 1. Результаты моделирования диффеоморфных преобразований

t	x_0^0	x_3^0	x_6^0	x_9^0	x_0^0	ε
0	(0,00;1,00)	(1,00;0,00)	(0,00;-1,00)	(-1,00;0,00)	(0,00;1,00)	0,000
2	(0,01;0,77)	(1,41;-0,01)	(-0,01;-0,77)	(-1,43;0,02)	(0,01;0,77)	0,840
4	(0,02;0,55)	(1,83;-0,01)	(-0,02;-0,55)	(-1,84;0,02)	(0,02;0,55)	0,954
6	(0,02;0,36)	(2,22;-0,01)	(-0,02;-0,36)	(-2,22;0,03)	(0,02;0,36)	0,987
8	(0,02;0,18)	(2,56;-0,01)	(-0,02;-0,18)	(-2,54;0,03)	(0,02;0,18)	0,998
10	(0,03;0,00)	(2,85;-0,02)	(-0,03;0,00)	(-2,82;0,05)	(0,03;0,00)	1,000
	x_0^1	x_3^1	x_6^1	x_9^1	x_0^1	
x^1	(0,00;0,00)	(3,14;0,00)	(0,00;0,00)	(-3,14;0,00	(0,00;0,00)	1,000

В результате получена дисперсия неточности попадания

$\sum_{n=1}^{12} \left\|x_n^1 - g_1\left(x_n^0\right)\right\|^2 = 0,34$ и среднее отклонение одной точки от цели $\sqrt{12^{-1} \sum_{n=1}^{12} \left\|x_n^1 - g_1\left(x_n^0\right)\right\|^2} = 0,17$. Для повышения точности попадания необходимо увеличить число итераций и количество частиц в методе PSO, а также уменьшить параметр дисперсии σ^2.

Заключение

Рассмотрена задача оценивания расстояния между замкнутыми 2D кривыми. Представлены методы нахождения инвариантов к действию групп переноса, вращения и масштабирования на замкнутую кривую, не зависящие от координатного описания изображения. Для оценивания нормы расстояния между двумя замкнутыми кривыми формируется функционал, соответствующий норме расстояния между двумя кривыми, и уравнение эволюции диффеоморфных преобразований, полученное решением вариационной задачи. Предложен алгоритм решения уравнения диффеоморфного преобразования, построенный на основе метода PSO, который позволяет значительно сократить объём вычислительных операций по сравнению с градиентными методами решения. В дальнейшем алгоритм решения уравнения диффеоморфного преобразования будет распространен на 3D объекты: точечные множества, кривые и поверхности. Следует рассмотреть задачу распознавания динамически изменяющихся объектов методом решения уравнений диффеоморфного преобразования.

Приложение. Метод PSO [9]

Рассмотрим задачу оптимизации (максимизации) без ограничений: $\mathrm{Maximize} f\left(\mathbf{X}\right); \mathbf{X}^{(l)} \leqslant \mathbf{X} \leqslant \mathbf{X}^{(u)}$, где $\mathbf{X}^{(l)}, \mathbf{X}^{(u)}$ — нижняя (lower) и верхняя (upper) границы $\mathbf{X}$. Пусть число частиц N. Процедура PSO применяется с использованием следующих шагов.

1. Сформируем случайное начальное множество $\mathbf{X}_1\left(0\right), \ldots, \mathbf{X}_N\left(0\right)$. Положение и скорость частицы j при итерации i: $\mathbf{X}_j^{(i)}, \mathbf{V}_j^{(i)}$, соответственно. Определим значение целевой функции: $f\left[\mathbf{X}_1\left(0\right), \ldots, \mathbf{X}_N\left(0\right)\right]$.

2. Найдём скорости частиц. Начальные скорости всех частиц принимаются равными нулю и номер итерации: $i = 1$.

3. На итерации i найдём параметры $\mathbf{X}_j^{(i)}, \mathbf{V}_j^{(i)}$ частицы j:

(a) Историческое лучшее значение положения $\mathbf{X}_j^{(i)}$: $\mathbf{P}_{best,j}$ с лучшим значением целевой функции $f\left[\mathbf{X}_j^{(i)}\right]$ частицы j на всех предыдущих итерациях. Историческое лучшее значение положения $\mathbf{X}_j^{(i)}$: $\mathbf{G}_{best}$ с лучшим значением целевой функции $f\left[\mathbf{X}_j^{(i)}\right]$ на всех предыдущих итерациях для всех N частиц; (b) найдём скорость частицы j на итерации i:

$$\mathbf{V}_j^{(i)} = \vartheta \cdot \mathbf{V}_j^{(i-1)} + c_1 r_1 \left[\mathbf{P}_{best,j} - \mathbf{X}_j^{(i-1)}\right] + c_2 r_2 \left[\mathbf{G}_{best} - \mathbf{X}_j^{(i-1)}\right] + c_3 \cdot r_3 \cdot 2^{-i/i_0},$$

где c_1, c_2, c_3 — скорости обучения, $r_1, r_2, r_3 \in [0 \dots 1]$ — равномерно случайно распределённые числа; (c) найдём положение частицы j на итерации i: $\mathbf{X}_j^{(i)} = \mathbf{X}_j^{(i-1)} + \mathbf{V}_j^{(i)}$ и соответствующее значение целевой функции $f\left[\mathbf{X}_1^{(i)}, \dots, \mathbf{X}_N^{(i)}\right]$.

4. Шаг 3 повторяется с $i = i + 1$ и новыми значениями $\mathbf{P}_{best,j}, \mathbf{G}_{best}$. Процесс продолжается до тех пор, пока все частицы не сойдутся к значению, обеспечивающему оптимум целевой функции.

Литература

1. Beg M.F. et al. Computing large deformation metric mappings via geodesic flows of diffeomorphisms // International journal of computer vision. 2005. T. 61, N. 2. P. 139–157.
2. Чуканов С.Н. Преобразование Фурье функции трехмерного изображения, инвариантное к действию групп вращения и переноса // Автометрия. 2008. Т. 44, № 3. С. 80–87.
3. Baker A. Matrix groups: An introduction to Lie group theory. Springer, 2012.
4. Arnold V.I., Khesin B.A. Topological methods in hydrodynamics. Springer, 1998.
5. Holm D.D. et al. Geometric mechanics and symmetry: from finite to infinite dimensions. London : Oxford University Press, 2009.
6. Miller M.I., Trouve A., Younes L. Geodesic shooting for computational anatomy // Journal of mathematical imaging and vision. 2006. Т. 24, № 2. С. 209–228.
7. Bruveris M., Holm D.D. Geometry of image registration: The diffeomorphism group and momentum maps // Geometry, Mechanics, and Dynamics. Springer New York, 2015. P. 19–56.
8. Kennedy J. et al. Swarm intelligence. Morgan Kaufmann, 2001.
9. Yang X. S. Nature-inspired optimization algorithms. Elsevier, 2014.

APPLICATION OF DIFFEOMORPHIC TRANSFORM OF CURVES FOR SOLVING PATTERN RECOGNITION PROBLEMS

S.N. Chukanov[1]
Dr.Sc.(Eng.), Professor, Senior Scientist Researcher, e-mail: ch_sn@mail.ru
D.B. Abramov[2]
Graduate Student, e-mail: cuntz@mail.ru
S.O. Baranov[2]
Graduate Student, e-mail: serj@doctor.com
S.V. Lejhter[2]
Graduate Student, e-mail: leykhter@mail.ru

[1]Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Omsk branch
[2]Department of "Computer Information Automated Systems", State Automobile and Highway Academy

Abstract. The problem of estimating the norm of the distance between the two closed smooth curves for pattern recognition is considered. Diffeomorphic transformations of curves based on the model of large deformations are described. For estimating of the norm of the distance between two closed curves the functional, corresponding normalized distance between the two curves, and the equation of diffeomorphic transformations evolution are formed. An algorithm for solving the equation of diffeomorphic transformation is proposed, built on the basis of PSO which can significantly reduce the number of computing operations compared with gradient methods for solving. The developed algorithms can be used in bioinformatics and biometrics systems, classification of images and objects, machine vision systems, for pattern recognition and object tracking systems.

Keywords: pattern recognition, invariance, diffeomorphic transformation, biometrics, PSO method.

Дата поступления в редакцию: 16.08.2016

Математические структуры и моделирование
2016. № 4(40). С. 81–89

УДК 519.237.5+616.12

СТРУКТУРНОЕ, ЭНТРОПИЙНОЕ МОДЕЛИРОВАНИЕ И КОРРЕЛЯЦИОННЫЙ АНАЛИЗ АРТЕРИАЛЬНОЙ ГИПЕРТЕНЗИИ

В.А. Шовин
научный сотрудник, e-mail: v.shovin@mail.ru

ФГБУН Институт математики им. С.Л. Соболева СО РАН, Омский филиал

Аннотация. На базе численных методов нелинейной оптимизации с условиями проведена оценка параметров экспертной структурной модели нормальной гемодинамики для пациентов с артериальной гипертензией начальной стадии до и после специального физиолечения. Незначительное изменение согласованности экспериментальных данных, полученных после физиолечения, со структурной моделью нормальной гемодинамики показывает отсутствие эффективности данного физиолечения для нормализации регуляции артериального давления. Энтропийное моделирование также показало отсутствие снижения энтропии самоорганизации и рост общей энтропии системы после лечения. Детальный корреляционный анализ выявил дезорганизацию взаимосвязей параметров, характеризующих нормальное функциональное состояния общей гемодинамики.

Ключевые слова: артериальная гипертензия, структурные уравнения, энтропийное моделирование, корреляционный анализ.

Введение

Структурные уравнения являются формой описания зависимостей между измеряемыми и латентными (не измеряемыми) переменными исследуемого объекта. Метод моделирования отношений между несколькими измеренными и латентными переменными оформился в 1970-х гг. в работах статистиков К. Йорескога и Д. Сорбома [1], социологов Г. Блэлока, О. Дункана [2,3], эконометриста А. Голдбергера [4] и психометриста П. Бентлера [5]. В общем случае такие зависимости могут иметь нелинейный характер функций модели.

Энтропийное моделирование позволяет на базе выборки показателей, характеризующих стохастическую систему в различных состояниях, оценить эффект самоорганизации и изменения дисперсии переменных [6]. В то же время более детальный анализ корреляций между отдельными переменными системы может выявить непосредственный вклад показателей в изменение энтропии самоорганизации.

Целью данной работы является построение структурной модели нормальной гемодинамики и оценка качества влияния процедур физиотерапии на нормали-

зацию функционального состояния гемодинамики на базе энтропийного моделирования, корреляционного анализа и анализа изменения согласованности структурной модели нормальной гемодинамики с экспериментальными данными пациентов с артериальной гипертензией до и после процедур физиолечения.

Задачи данного исследования представлены следующими пунктами:

- сформировать структурную модель нормальной гемодинамики на базе экспертных медицинских данных;

- оценить параметры экспертной структурной модели нормальной гемодинамики и согласованность модели с экспериментальными данными для лиц с артериальной гипертензией до и после процедур специального физиолечения;

- провести энтропийное моделирование артериальной гипертензии на базе выборки показателей как многомерной стохастической системы;

- провести детальный корреляционный анализ показателей и выявить непосредственную роль тех или иных показателей в изменениях функционального состояния общей гемодинамики.

В качестве экспериментальных данных выступали различные параметры, характеризующие состояние пациентов с артериальной гипертензией. Поскольку экспериментальные данные представлены выборкой значений измеряемых переменных у различных исследуемых объектов, для оценки параметров и значений латентных переменных модели, задаваемой структурными уравнениями, может быть использован критерий минимальных невязок как сумма невязок модели, вычисленная для всей выборки различных объектов. Дополнительно на параметры и значения латентных переменных могут быть заданы ограничительные условия. Для решения задачи минимизации невязок модели предлагается использовать метод безусловной нелинейной оптимизации: метод конфигураций. Оценка параметров экспертной структурной модели нормальной гемодинамики осуществлялась для пациентов с артериальной гипертензией начальной стадии до и после специального физиолечения.

Альтернативой данному методу оценки нормализации функционального состояния гемодинамики является энтропийное моделирование и корреляционный анализ.

1. Математическая постановка задачи оценки параметров структурных моделей

В теории структурных уравнений используются следующие типы матриц.

Матрица $Z \underset{m\times n}{\leftrightarrow} z_{ij}$ — матрица значений измеряемых переменных у исследуемых объектов или состояний объекта размерности $m \times n$, где m — число измеряемых параметров, n — число объектов или состояний объекта (объём выборки).

Матрица $P \underset{g\times n}{\leftrightarrow} p_{ij}$ — матрица значений латентных переменных объектов размерности $g \times n$, где g — число латентных параметров.

Матрица $A \underset{k\times s}{\leftrightarrow} a_{ij}$ — матрица параметров структурных уравнений размерности $k \times s$, где k — число структурных уравнений, s — число параметров в

структурных уравнениях.

Система структурных уравнений задаётся в виде:

$$\begin{cases} f_1\left(a_{11}, a_{12}, \ldots, a_{1s}; p_{1t}, p_{2t}, \ldots, p_{gt}; z_{1t}, z_{2t}, \ldots, z_{mt}\right) + \varepsilon_{1t} = 0, \\ f_2\left(a_{21}, a_{22}, \ldots, a_{2s}; p_{1t}, p_{2t}, \ldots, p_{gt}; z_{1t}, z_{2t}, \ldots, z_{mt}\right) + \varepsilon_{2t} = 0, \\ \vdots \\ f_k\left(a_{k1}, a_{k2}, \ldots, a_{ks}; p_{1t}, p_{2t}, \ldots, p_{gt}; z_{1t}, z_{2t}, \ldots, z_{mt}\right) + \varepsilon_{kt} = 0, \end{cases}$$

где $f_1, f_2, \ldots, f_k$ — в общем случае нелинейные функции своих переменных, $\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{kt}$ — невязки модели для t-го объекта или состояния объекта.

На значения параметров и значения латентных переменных могут накладываться дополнительные условия в виде равенств и неравенств.

Оптимальными значениями параметров и латентных переменных считаются те значения, которые минимизируют абсолютные значения невязок модели и удовлетворяют всем дополнительным условиям.

2. Методы оптимизации

Оптимизацию критерия суммы невязок модели, как функции от независимых параметров и латентных переменных с ограничениями, предлагается осуществлять методом штрафных функций [7]. В качестве метода безусловной оптимизации метода штрафных функций был выбран метод конфигураций [8].

3. Структурная модель нормальной гемодинамики

В научном исследовании был использован вид структурной модели для описания регуляции артериального давления в норме [9]:

$$
\begin{cases}
x_{1t} = a_{11}x_{3t} + a_{12}x_{4t} + a_{13} + \varepsilon_{1t}, \\
x_{2t} = a_{21}x_{3t} + a_{22}x_{4t} + a_{23} + \varepsilon_{2t}, \\
x_{4t} = a_{31}p_{1t} + a_{32} + \varepsilon_{3t}, \\
p_{1t} = a_{41}x_{3t} + a_{42}p_{2t} + a_{43} + \varepsilon_{4t}, \\
p_{2t} = a_{51}x_{5t} + a_{52} + \varepsilon_{5t}, \\
x_{3t} = a_{61}x_{6t} + a_{62}x_{7t} + a_{63} + \varepsilon_{6t}, \\
x_{7t} = a_{71}p_{3t} + a_{72} + \varepsilon_{7t}, \\
x_{8t} = a_{16,1}p_{6t} + a_{16,2} + \varepsilon_{16t}, \\
p_{3t} = a_{81}x_{8t} + a_{82} + \varepsilon_{8t}, \\
x_{5t} = a_{91}p_{4t} + a_{92} + \varepsilon_{9t}, \\
p_{4t} = a_{10,1}x_{1t} + a_{10,2}x_{2t} + a_{10,3} + \varepsilon_{10,t}, \\
p_{5t} = a_{11,1}x_{5t} + a_{11,2}p_{6t} + a_{11,3}p_{7t} + a_{11,4} + \varepsilon_{11,t}, \\
p_{7t} = a_{12,1}p_{6t} + a_{12,2}x_{8t} + a_{12,3} + \varepsilon_{12,t} \\
x_{6t} = a_{13,1}p_{5t} + a_{13,2} + \varepsilon_{13,t} \\
p_{6t} = a_{14,1}x_{5t} + a_{14,2} + \varepsilon_{14,t} \\
p_{7t} = a_{15,1}x_{9t} + a_{15,2}x_{10t} + a_{15,3}x_{11t} + a_{15,4}x_{12t} + a_{15,5}x_{13t} + a_{15,6}x_{14t} + a_{15,7} + \varepsilon_{15,t}.
\end{cases}
$$

Где измеряемые переменные:
систолическое артериальное давление (САД) — x_1,
диастолическое артериальное давление (ДАД) — x_2,
минутный объём сердца (МОС) — x_3,
общее периферическое сосудистое сопротивление (ОПСС) — x_4,
индекс Кердо — x_5,
ударный объём (УО) – x_6,
частота сердечных сокращений (ЧСС) – x_7,
индекс Хильдебрандта — x_8,
конечно-систолический размер левого желудочка (КСР) – x_9,
конечно-систолический объём левого желудочка (КСО) – x_{10},
конечно-диастолический размер левого желудочка (КДР) – x_{11},
конечно-диастолический объём левого желудочка (КДО) — x_{12},
фракция выброса левого желудочка (ФВ) – x_{13},
фракция укорочения левого желудочка (ФУ) — x_{14};
латентных переменные:
венозный возврат — p_1,
тонус вен — p_2,
темп деполяризации водителя ритма — p_3,
барорецепторы — p_4,
сократимость сердечной мышцы — p_5,
адреналин — p_6,

структурно-геометрическое состояние сердца — p_7.

Схема регуляции артериального давления, соответствующая данной структурной модели, представлена на рис. 1.

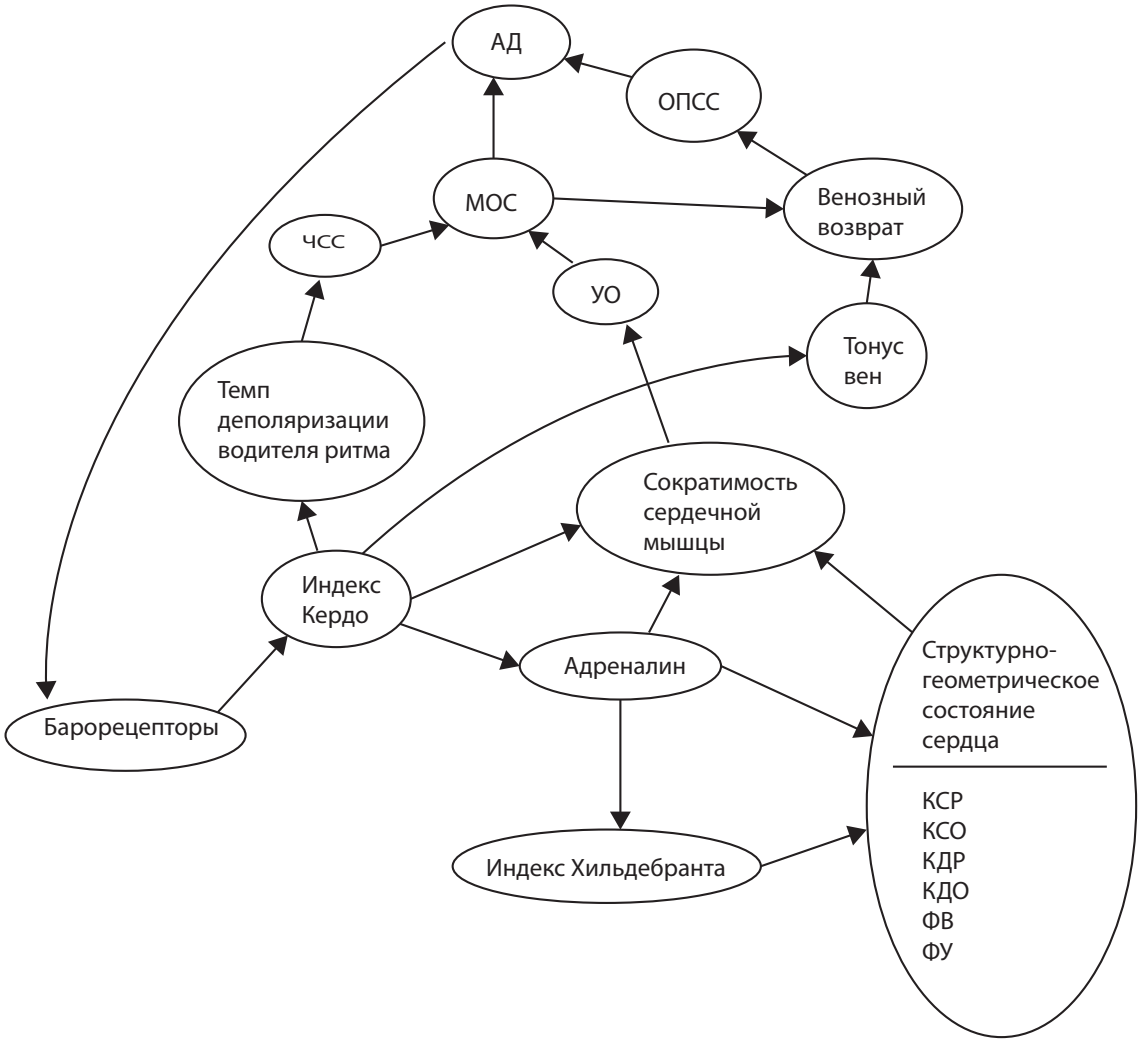


Рис. 1. Схема регуляции артериального давления сердечно-сосудистой и нервной системами

Данная схема описывает влияние изменения одних переменных от монотонного изменения других при нормальной регуляции артериального давления. Например, при росте частоты сердечных сокращений и увеличении ударного объёма при стрессовой ситуации растёт минутный объём сердца, который в случае повышенного общего сосудистого сопротивления, например, при избыточном весе, приводит к повышенному артериальному давлению. В случае нормального состояния гемодинамики значимая взаимообусловленность показателей должна приводить к компенсационным процессам для нормализации артериального давления. Поэтому важно подтвердить или опровергнуть эффективность тех или иных медицинских процедур для выявления положительного влияния на значимое увеличение взаимообусловленности различных показате-

лей гемодинамики, которое в свою очередь характеризует рост динамической компенсации физиологических систем организма для стабилизации артериального давления.

4. Энтропийное моделирование и корреляционный анализ

Приращение энтропии многомерной стохастической системы может происходить за счёт изменения дисперсий и корреляций показателей, характеризующих систему [6]:

$\Delta H\left(\bar{Y}\right) = \Delta H\left(\bar{Y}\right)_{\Sigma} + \Delta H\left(\bar{Y}\right)_{R}$, где

$\bar{Y} = \left(Y_1,\ Y_2,\ \ldots,\ Y_m \right)$ — многомерный вектор показателей системы.

Доказано, что если вектор $\bar{Y}$ является гаусовским, то

$$\Delta H\left(\bar{Y}\right)_{\Sigma} = \sum_{k=1}^{m} \ln \frac{\sigma_{Y_k^{(2)}}}{\sigma_{Y_k^{(1)}}},$$

$$\Delta H\left(\bar{Y}\right)_{R} = \frac{1}{2} \ln \frac{|R_{\bar{Y}^{(2)}}|}{|R_{\bar{Y}^{(1)}}|},$$

где $\sigma_{Y_k^{(i)}}$ — дисперсия k-го показателя в i-ом состоянии, $|R_{\bar{Y}^{(i)}}|$ — определитель корреляционной матрицы $R_{\bar{Y}^{(i)}}$ многомерного вектора показателей в i-ом состоянии.

5. Численный эксперимент

Численный эксперимент данного научного исследования заключался в тестировании метода оценки параметров структурных уравнений в рамках структурной модели нормальной гемодинамики на базе 15 биофизических показателя измеренных у 131-го пациента с артериальной гипертензией начальной стадии:

1) *вес,*
2) *индекс массы тела (ИМТ),*
3) *частота дыхания (ЧД),*
4) *сегментоядерные нейтрофилы (С),*
5) *лимфоциты (Л),*
6) *конечно-систолический размер левого желудочка (КСР),*
7) *конечно-систолический объём левого желудочка (КСО),*
8) *конечно-диастолический размер левого желудочка (КДР),*
9) *конечно-диастолический объём левого желудочка (КДО),*
10) *ударный объём (УО),*
11) *минутный объём сердца (МОС),*
12) *общее периферическое сосудистое сопротивление (ОПСС),*
13) *индекс Хильдебрандта (ИХ),*
14) *фракция выброса левого желудочка (ФВ),*
15) *фракция укорочения левого желудочка (ФУ).*

Все показатели были проверены на нормальность распределения.

Для 131 лица с артериальной гипертензией начальной стадии до и после специального физиолечения были оценены параметры структурных уравнений. Распределение невязок модели нормальной гемодинамики для лиц с артериальной гипертензией до и после физиолечения представлены на рис. 2, 3.

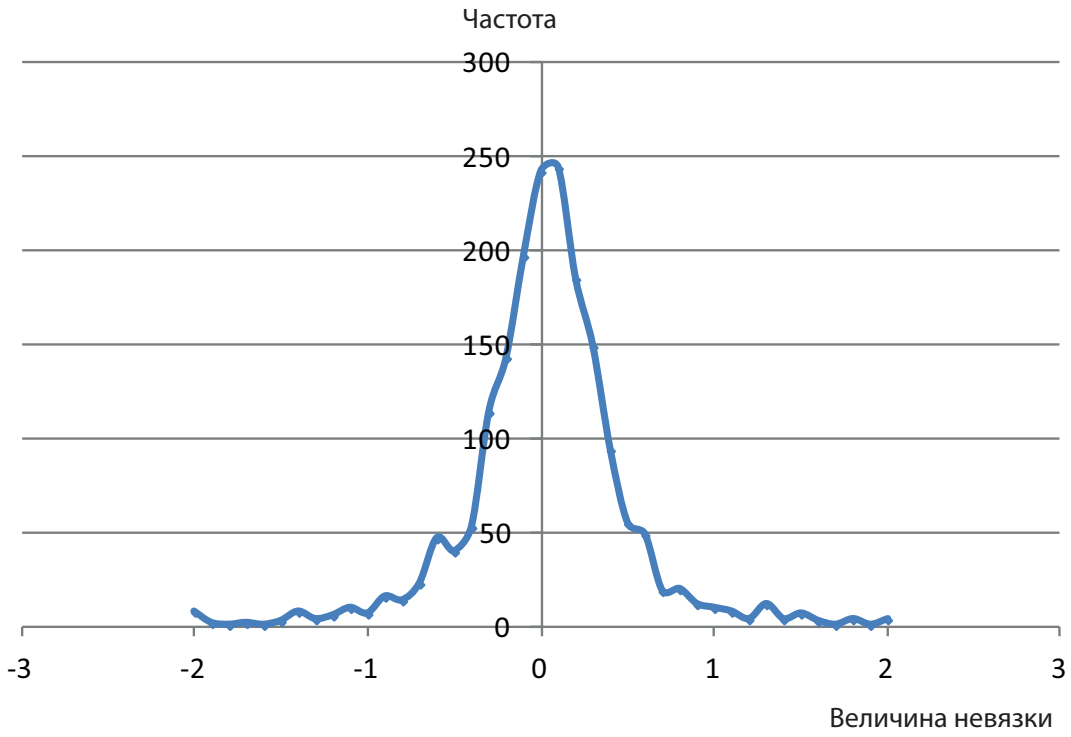


Рис. 2. Распределение невязок модели нормальной гемодинамики до физиолечения

На распределениях заметно, что частота невязок, близких к нулевым, после лечения незначительно увеличилась. По всей видимости, это говорит об отсутствии нормализации регуляции артериального давления после процедур специального физиолечения.

Энтропийное моделирование показало, что $\Delta H\left(\bar{Y}\right)_{\Sigma} = -3,78$, $\Delta H\left(\bar{Y}\right)_{R} = 4,39$, $\Delta H\left(\bar{Y}\right) = 0,6$, т.е. снижение энтропии хаотичности, увеличение энтропии самоорганизации и увеличение общей энтропии многомерной стохастической системы.

Детальный анализ отдельных коэффициентов корреляций показал, что параметр ОПСС после лечения перестал коррелировать с МОС и структурными показателями сердца. Это свидетельствует о том, что проведённое лечение не привело к нормализации функционального состояния общей гемодинамики, а понижение АД обусловлено воздействием на ОПСС, уменьшением ОПСС. Однако после лечения с индексом работы сердца начал коррелировать параметр индекс Кердо, что свидетельствует о нормализации баланса симпатической и парасимпатической систем за счёт благоприятного воздействия на ОПСС.

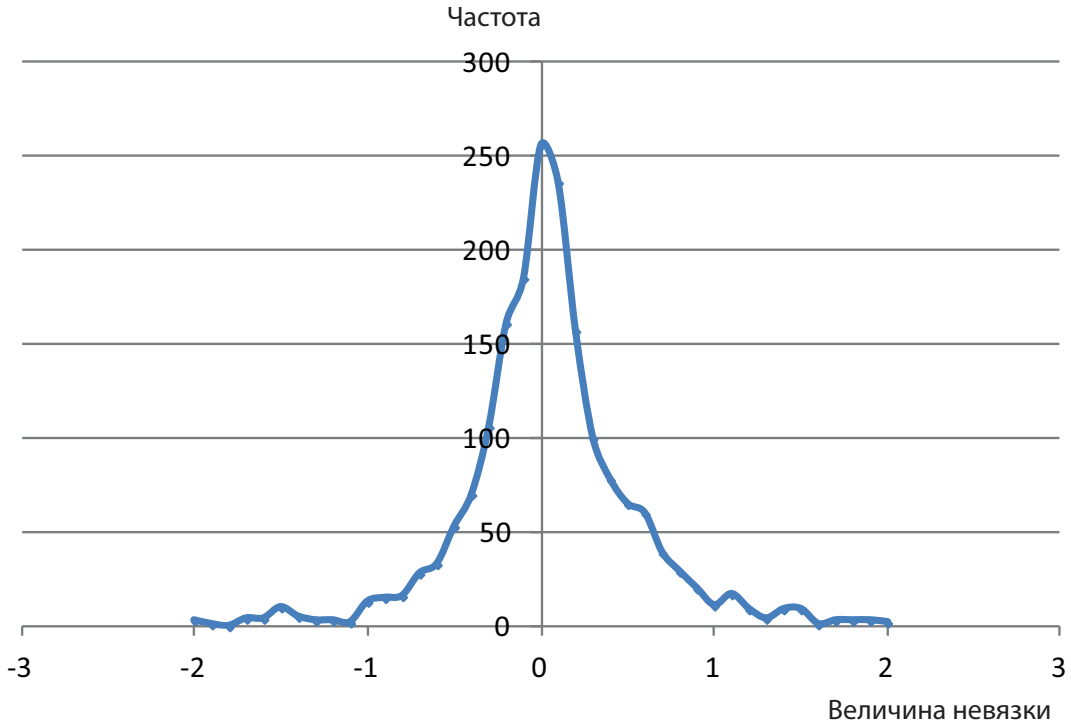


Рис. 3. Распределение невязок модели нормальной гемодинамики после физиолечения

6. Заключение

На базе экспертных сведений была сформирована структурная модель нормальной гемодинамики. Для лиц с артериальной гипертензией начальной стадии были оценены параметры и невязки модели до и после специального физиолечения. Незначительное изменение распределения невязок модели после физиолечения говорит об отсутствии нормализации регуляции артериального давления. Независимо было проведено энтропийное моделирование и выявлен рост энтропии самоорганизации и общей энтропии при снижении энтропии, обусловленной разбросом значений показателей выборки. Детальный корреляционный анализ обозначил главный вклад в отсутствие нормализации функционального состояния общей гемодинамики, а именно, нарушение взаимосвязи общего периферического сосудистого сопротивления с минутным объёмом сердца и структурными показателями сердца.

Литература

1. Joreskog K.G., Sorbom D. Advances in factor analysis and structural equation models. Edited by Jay Magidson, Cambridge, Mass.: Abt Books, 1979.
2. Blau P.M., Duncan O.D., Tyree A. The American Occupational Structure. New York : Wiley and Sons, 1967.
3. Blalock H. Theory construction. Englewood Cliffs, New Jersey : Prentice-Hall, 1968.

4. Joreskog K.G., Goldberger A.S. Estimation of a model with multiple indicators and multiple causes of a single latent variable // Journal of the American Statistical Association. 1975. N. 70(351). P. 631–639.
5. Bentler P.M. Multivariate analysis with latent variables: Causal modeling // Annual review of psychology. 1980. N. 31(1). P. 419–456.
6. Тырсин А.Н., Ворфоломеева О.В. Исследование динамики многомерных стохастических систем на основе энтропийного моделирования // Информ. и её примен. 2013. Т. 7, Вып. 4. С. 3–10.
7. Банди Б. Методы оптимизации. Вводный курс. М. : Радио и связь, 1988.
8. Кокуев А.Г. Оптимальное управление. Поиск экстремумов многомерных функций. АГТУ — Астрахань, 2011. 34 с.
9. Багаев С.Н. и др. Система кровообращения и артериальная гипертония: биофизические и генетико-физиологические механизмы, математическое и компьютерное моделирование. Новосибирск : Изд-во СО РАН, 2008. 252 с.

STRUCTURAL, ENTROPY MODELING AND CORRELATION ANALYSIS OF HYPERTENSION

V.A. Shovin
Scientist Researcher, e-mail: v.shovin@mail.ru

Omsk Branch of the Institution of the Russian Academy of Sciences Institute of Mathematics. S. Siberian Branch of RAS

Abstract. Estimates of the parameters of expert structural model of normal hemodynamics was implemented by numerical methods for nonlinear optimization with conditions. Slight increased agreement between experimental data after physiotherapy and structural model of normal hemodynamic showed lack of physiotherapy efficiency for the normalization of blood pressure regulation. Entropy modeling also showed no reduction in entropy of self-organization and growth of the total entropy of the system after treatment. Detailed analysis revealed a disorganization of correlation between the parameters of normal functional status of the hemodynamics.

Keywords: hypertension, structural equations, entropy modeling, correlation analysis.

Дата поступления в редакцию: 13.08.2016

Математические структуры и моделирование
2016. № 4(40). С. 90–95

УДК 004.912

ПРИМЕНЕНИЕ МЕТОДА АНАЛИЗА ИЕРАРХИЙ СОМЕСТНО С АЛГОРИТМОМ КЛАСТЕРИЗАЦИИ В ОБРАБОКЕ ДАННЫХ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

А.Н. Мироненко
к.т.н., доцент, e-mail: mironim84@mail.ru

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В работе рассматривается возможность применения известного в математике метода анализа иерархий совместно с алгоритмом кластеризации FOREL для классификации субъектов. Смысл объединения заключается в том, что, применяя метод анализа иерархий, а именно принятия решений в условиях определённости, мы подготавливаем данные для дальнейшей работы с ними, а алгоритмом кластеризации (таксономии) происходит их непосредственная обработка. Работу предлагаемого подхода можно условно разделить на два этапа: этап обучения и этап работы. Было проведено компьютерное моделирование, проверяющее состоятельность предлагаемого подхода.

Ключевые слова: метод анализа иерархий, МАИ, кластеризация, таксономия.

Введение

В работе предлагается исследовать возможность применения теории игр, а именно метода анализа иерархий, с целью подготовки данных для их последующей кластеризации одним из существующих алгоритмов.

В статье [1] рассматриваются возможности метода анализа иерархий, способы его применения и их особенности. Одним из важных элементов метода анализа иерархий являются матрицы попарных сравнений (pairwise comparison matrices). В статье [2] автором описывается процесс их построения и нормализации. В работе [3] проводится исследование проблемы реверса рангов (rank reversal), т.е. изменения ранжирования альтернатив выбора при их удалении или добавлении, даётся математическое описание данной проблемы и приводится доказательство её существования.

Кластерный анализ достаточно подробно был рассмотрен в статье [4]. Автор рассматривает различные методы кластеризации, а также отмечает важную роль выбора координат центра таксонов и критерия схожести (расстояние от центра таксона до точек, которые будут считаться принадлежащими таксону).

Кроме того, показывается, насколько результат кластеризации чувствителен к выбору функции расстояния, использующейся для определения близости точек.

Идея объединения кластеризации и других математических методов рассмотрена в статье [5]. В ней также исследуется возможность совместного применения метода главных компонент, иерархической кластеризации и строгой кластеризации (Principal component methods — hierarchical clustering — partitional clustering) с целью лучшей визуализации данных. Метод главных компонент применяется для предварительной обработки, а методы иерархической и строгой кластеризации — для представления данных.

1. Постановка задачи

Актуальность повышения качества, поиска новых методов и методологий социологического исследования не вызывает сомнения. Наиболее востребованным является поиск возможных применений методов математического моделирования и информационных технологий для сбора и анализа данных.

В рамках социологических исследований и дальнейшей обработки данных решается задача отнесения субъекта к той или иной группе. Кроме того, может решаться и другая задача — выделение нетипичных субъектов, то есть тех, которые нельзя отнести ни к одной из групп. Данная задача называется одноклассовой классификацией, обнаружением нетипичностей или новизны (novelty detection) [6].

С целью поиска новых методов и подходов к социологическим исследованиям, а именно с целью решения задачи классификации субъектов и обнаружения нетипичностей, предлагается объединить хорошо изученный в математике метода анализа иерархий (МАИ) с одним из алгоритмов кластерного анализа и исследовать результаты данного объединения на практике.

2. Теория

Применение МАИ совместно с алгоритмом кластеризации для классификации субъектов можно условно разделить на два этапа: этап подготовки данных (алгоритм формирования групп) и непосредственно сама классификация (алгоритм определения принадлежности субъекта к группе).

Для решения задачи классификации субъектов предлагается использовать алгоритм кластеризации FOREL. Алгоритм работает с точками на n-мерном пространстве, т.е. нам необходимо представить субъекты, которые мы хотим классифицировать в виде точек с n-координатами.

Прежде чем приступить к классификации, необходимо подготовить данные для работы с ними. Для этого используется МАИ. Перед субъектом ставится задача с определёнными критериями выбора и альтернативами её решения. Затем выполняется следующий алгоритм.

Алгоритм формирования групп:

1. Субъект для каждого из критериев указывает его важность относительно других;

2. Вычисляются относительные веса критериев;
3. Субъект указывает, насколько каждая из альтернатив предпочтительнее других в пределах каждого критерия;
4. Вычисляются относительные веса альтернативных решений;
5. Вычисляются комбинаторные весовые коэффициенты для каждого из решений;
6. Используя полученные весовые коэффициенты как координаты, получаем точку в n-мерном пространстве;
7. Повторяем шаги с 1 по 6 для всех субъектов;
8. Для полученного множества точек при помощи алгоритма FOREL решается задача кластеризации;
9. Для каждого таксона определяются координаты центра масс;
10. Таксоны упорядочиваются по величине G=Y/X, где X и Y — координаты центра масс таксона.

После формирования таксонов проводится анализ каждого из них с целью определить, какую группу субъектов он характеризует. Определяется, какое количество субъектов того или иного класса попало в тот или иной таксон в процентном соотношении от общего числа субъектов.

Алгоритм определения принадлежности субъекта к группе:

1. Перед новым субъектом, который мы хотим классифицировать, ставится задача с определёнными критериями выбора и альтернативами её решения.
2. Субъект выполняет шаги 1–6 алгоритма формирования данных для последующей классификации
3. Определяется принадлежность субъекта (n-мерной точки) к одному из таксонов.

3. Результаты эксперемента

Для проверки состоятельности предлагаемого метода был проведён эксперимент. Целью эксперимента является определение субъекта к группе людей с техническим складом ума или же с гуманитарным. Для этого было подготовлено три различных анкеты, в которых перед анкетируемыми ставится задача с некоторыми критериями выбора и альтернативами её решения. Всего в анкетировании приняло участие более 115 студентов различных факультетов ОмГУ им. Ф.М. Достоевского.

Затем выполнялись шаги алгоритма формирования групп.

В результате проведения данного этапа эксперимента были получены следующие кластеризации с различным количеством таксонов, например, рис. 1.

На рис. 1 слева представлена кластеризация с параметром критерия схожести 0.2, справа — с параметром 0.7. Так как нам необходимо получить 2 таксона, мы экспериментально подбираем параметры, при которых получим нужную нам кластеризацию. Для нашего эксперимента критерий схожести равен 0.2. На рис. 1 таксономия представлена слева.

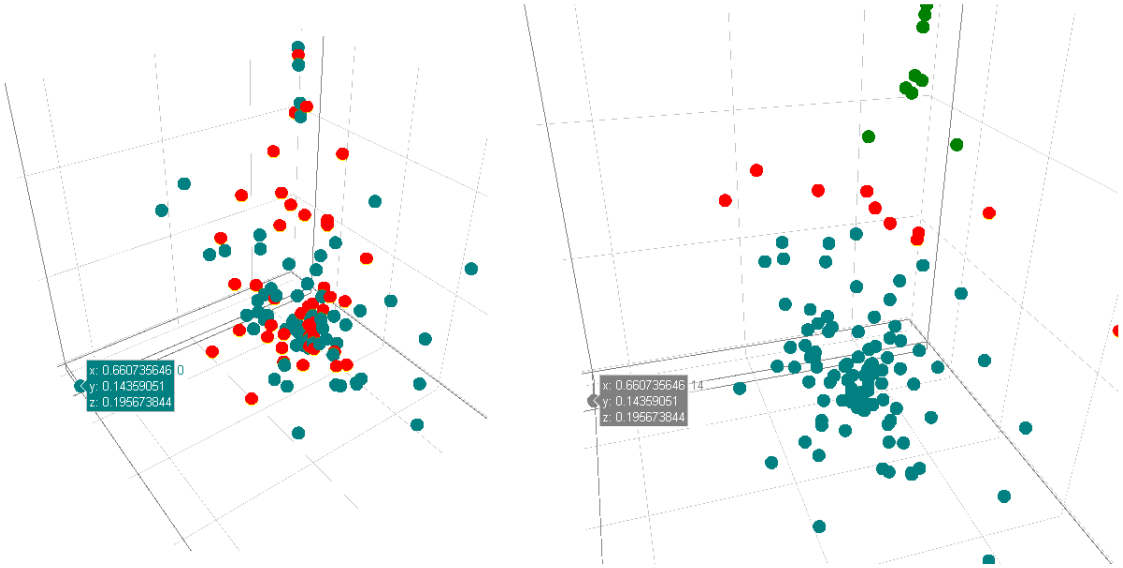


Рис. 1. Пример кластеризации

Уже на данном этапе проведения эксперимента можно видеть, что таксоны располагаются очень близко, при этом получается, что один экземпляр может одновременно попадать в несколько разных таксонов, что не желательно в связи с тем, что усложняет процедуру определения принадлежности к той или иной группе, которой соответствует таксон.

4. Обсуждение результатов

В результате эксперимента возникли некоторые сложности, связанные с тем, что при выборе критерия схожести, который бы давал нужную нам кластеризацию, мы получили очень близко расположенные таксоны, что существенно затрудняет классификацию новых объектов в силу того, что не удалось получить чётко сформированные группы. Это может быть связано с неудачным выбором алгоритма кластеризации, так как алгоритм FOREL, как правило, применяется в случаях, когда число таксонов, на которые необходимо осуществить разбиение выборки, заранее не известно. В нашем эксперименте число таксонов было известно, и чтобы получить необходимое количество, пришлось изменять критерии схожести, что в свою очередь могло негативно сказаться на результатах эксперимента.

При этом в результате эксперимента подтвердилось предположение о возможном применении МАИ и алгоритма кластеризации, то есть, возможность применять МАИ для подготовки данных с целью их последующей кластеризации.

Выводы и заключение

В работе был предложен метод совместного применения метода анализа иерархий (принятие решения в условиях определённости) и кластерного анализа. Был проведён эксперимент, позволяющий проверить возможность такого объединения.

Эксперимент показал, что такое применение возможно, но требует дополнительного исследования. Необходимо выбрать более удачный алгоритм кластеризации в случае, когда нам необходимо получить классификацию субъектов по чётко определённым группам и выявить принадлежность нового субъекта к той или иной группе.

В случае, когда количество групп, по которым необходимо осуществить классификацию, нам неизвестно, применение метода возможно с алгоритмом FOREL, но необходимо дополнительное исследование каждой из групп для определения, какие объекты она содержит. В этом случае можно решать задачу обнаружения нетипичностей, определять объекты, которые невозможно отнести ни к какой группе.

Литература

1. Thomas L. Saaty Decision making with the Analytic Hierarchy Process // International Journal of Services Sciences. 01/2008. N. 1(01). P. 83–98.
2. Farkas A. The Analysis of the Principal Eigenvector of Pairwise Comparison Matrices // Acta Polytechnica Hungarica. 2007. V. 4, Issue 2.
3. Недашковская Н.И. Метод анализа иерархий в методологии сценарного анализа решения задач предвидения // Восточно-Европейский журнал передовых технологий. 2010. № 9(46), Т. 4.
4. Cluster Analysis. URL: `https://www.qualtrics.com/wp-content/uploads/2013/05/Cluster-Analysis.pdf` (дата обращения: 10.05.2016).
5. Husson F., Josse J., Pages J. Principal component methods — hierarchical clustering — partitional clustering: why would we need to choose for visualizing data? // Technical Report — Agrocampus. 09/2010.
6. Метод одноклассовой классификации интервальных данных с использованием треугольного ядра, основанный на теории Демстера-Шефера // Официальный сайт Санкт-Петербургского государственного лесотехнического университета им. С.М. Кирова URL: `http://spbftu.ru/UserFiles/Image/izvesti/22-210.pdf` (дата обращения: 10.05.2016).

APPLYING THE ANALYTIC HIERARCHY PROCESS IN CONJUNCTION WITH CLUSTERING ALGORITHM TO CLASSIFY DIFFERENT SUBJECTS

A.N. Mironenko
Ph.D.(Eng.), Associate Professor, e-mail: mironim84@mail.ru

Dostoevsky Omsk State University

Abstract. This paper examines the possibility of applying the analytic hierarchy process, known in mathematics, in conjunction with the FOREL clustering algorithm to classify different subjects. By term "conjunction" we mean a process when the analytic hierarchy process (namely decision making under certainty) is used for preparation of data for further work with them, and the clustering algorithm (taxonomy) is used for direct processing of the data. The proposed approach can be divided into two stages: the training stage and the work stage. We carried out a computer simulation which verifies validity of the proposed approach.

Keywords: analytic hierarchy process, clustering, taxonomy.

Дата поступления в редакцию: 09.10.2016

Математические структуры и моделирование
2016. № 4(40). С. 96–101

УДК 004.896

ПРОГРАММА CHATBOT — ЧАТ-БОТ ИЛИ ВИРТУАЛЬНЫЙ СОБЕСЕДНИК

В.А. Шовин
научный сотрудник, e-mail: v.shovin@mail.ru

ФГБУН Институт математики им. С.Л. Соболева СО РАН, Омский филиал

Аннотация. Разработана программа виртуального собеседника или чат-бота на базе внешнего API, алгоритма поиска ответов в базе знаний расширенной разметки AIML, а также рекуррентной нейронной сети. Алгоритм позволяет находить ответы к вопросам релевантные вопросам из базы знаний. Процедура сортировки релевантных ответов включает в себя поиск по регулярному выражению, поиск по тематике, поиск по истории и поиск по максимальному совпадению слов в вопросах. Для повышения качества поиска релевантных ответов в интерпретатор внедрён модуль морфологического анализатора отдельных слов. Рекуррентная нейронная сеть задана на множестве слов всех вопросов и ответов базы знаний.

Ключевые слова: чат-бот, виртуальный собеседник, AIML, рекуррентная нейронная сеть.

Введение

На сегодняшний день остаётся актуальным создание программ, имитирующих общение человека. Простейшей моделью общения является база вопросов и ответов к ним [1]. В данном случае возникает проблема описания базы знаний и реализация программы-интерпретатора. Язык разметки базы знаний может включать в себя паттерны вопросов и соответствующие им шаблоны ответов, также предысторию диалогов к ним и название соответствующей темы общения.

Чат-бот может выполнять дополнительные функции, например, такие как поиск музыки, картинок, фактов; также отражать калькулятор, прогноз погоды, вывод курса валют. Большинство таких функций имеют реализацию в интернете и доступны в качестве внешнего API.

Альтернативным вариантом создания программы виртуального собеседника является использование алгоритмов машинного обучения на базе диалогов общения, а именно искусственные нейронные сети. Подходящей моделью ИНС является рекуррентная нейронная сеть, способная хранить, обобщать и прогнозировать различные последовательности. В данной работе в качестве элементов последовательности предлагается использовать индексы, соответствующие словам в базе знаний вопросов и ответов.

1. AIML

Одним из форматов разметки базы знаний является стандарт языка разметки AIML (Artificial Intelligence Markup Language). Ключевыми словами в языке являются category, pattern и template:

```
<aiml>
    <category>
        <pattern>Привет! Как дела?</pattern>
        <template>
            <random>
                <li>Привет. Нормально.</li>
                <li>Привет. Отлично. Как у тебя?</li>
            </random>
        </template>
    </category>
</aiml>
```

Тег category является родительским к тегам pattern и template, хранящим шаблоны вопроса и ответов. Тег random позволяет указать несколько ответов к вопросу, выбираемых интерпретатором случайным образом. В работе предлагается ввести дополнительные теги, соответствующие истории и теме разговора.

```
<aiml>
    <category>
        <pattern>Чем занимаешься?</pattern>
        <pattern>Что делаешь?</pattern>
        <template>
            <random>
                <li>Читаю книгу.</li>
                <li>Смотрю телевизор.</li>
            </random>
        </template>
        <history>Привет! Как дела? Привет. Нормально.</history>
        <theme>Дела</theme>
    </category>
</aiml>
```

Несколько тегов pattern позволяют описать различные варианты вопросов, соответствующие данной категории, на которые должны последовать одни и те же варианты ответов. Тег history хранит историю диалога, предшествующую данному вопросу. Тег theme хранит название темы разговора. Данные теги позволяют интерпретатору подобрать паттерн вопроса, соответствующий предыстории диалога и теме общения, что должно сказаться на улучшении качества имитации общения посредством чат-бота.

Интерпретатор языка разметки должен позволять находить наиболее релевантные вопросы следующими дополняющими друг друга способами:

1. Поиск по всей фразе вопроса на основе регулярного выражения.
2. Поиск по количеству совпадающих слов в вопросе и паттернах.
3. Поиск по совпадению текущей темы и тем категорий.
4. Поиск по количеству совпадающих слов в текущей предыстории разговора и предысториях категорий.

Соответствующий алгоритм выбора лучшего совпадения формируется на базе алгоритма сортировки:

```
sortMatches = allMatches.sort(function(a, b) {
    if(a.pattern == inputText && b.pattern != inputText)
        return -1;
    if(b.pattern == inputText && a.pattern != inputText)
        return 1;
    if(a.matches < b.matches)
        return 1;
    if(a.matches > b.matches)
        return -1;
    if(a.theme == bot.theme && b.theme != bot.theme)
        return -1;
    if(b.theme == bot.theme && a.theme != bot.theme)
        return 1;
    if(a.historyMatches < b.historyMatches)
        return 1;
    if(a.historyMatches > b.historyMatches)
        return -1;
    return 0;
})
```

В результате выбирается первый элемент отсортированного массива совпадений.

Для повышения качества поиска релевантных ответов в процедуру сравнения слов был внедрён модуль морфологического анализатора, позволяющий находить базовые формы слов. Тем самым сравнение происходит по базовым формам слов, что исключает несоответствия слов, связанные с их склонениями.

2. Алгоритм классификации тем

Предлагается использовать алгоритм определения названий тем категорий, когда определена лишь их часть.

1. Пусть множество категорий, в которых определены темы, — T. Множество категорий, в которых темы не определены, — D. Элементам данных множеств ставится в соответствие объединение строк всех значений паттернов и шаблонов категории.

2. Вводится порог детерминации, например, $p = 70\%$.

3. Последовательно перебираются элементы множества D. Выбирается подмножество V элементов множества T, в котором p — процент уникальных слов — определён.

4. Если V пусто, то тема определяется строкой элемента множества D.

5. Если V не пусто, то выбирается элемент из V, для которого отношение количества совпадающих уникальных слов к общему количеству уникальных слов в строке этого элемента максимально. Соответствующая тема данного элемента V определяется как тема элемента D.

3. Рекуррентная нейронная сеть

Рекуррентная нейронная сеть — вид многослойного перцептрона, у которого сигналы с нейронов выходного слоя поступают на дополнительные нейроны входного слоя, т.н. нейроны контекста.

Входной вектор сигнала поступает на группу нейронов INPUT, на группе нейронов CONTEXT нулевой сигнал. Далее сигнал распространяется в группу нейронов скрытого слоя HIDDEN, а затем преобразуется ими и попадает на нейроны выходного слоя OUTPUT. На следующей итерации вместе с вектором сигнала INPUT на контекстную группу нейронов поступают копии сигналов с выходного слоя OUTPUT прошлой итерации (рис. 1).

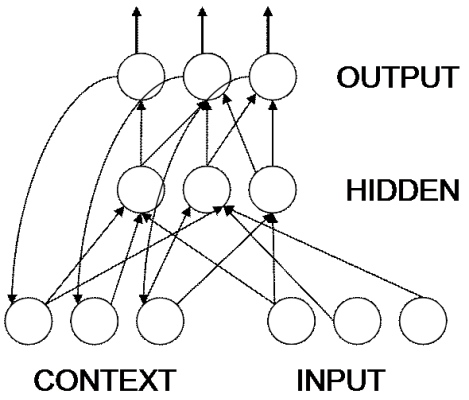


Рис. 1. Общий вид структуры рекуррентной нейронной сети.

Структура рекуррентной нейронной сети для запоминания предложений имеет следующий вид:

Слои CONTEXT, INPUT и OUTPUT имеют по одному нейрону, значение сигнала на выходе которого ставится в соответствие индексу слова в наборе слов. Дополнительно вводится слово __end__, соответствующее концу предложения [2]. Сеть последовательно обучается предложениям вида:

«Привет. Как дела? __end__ Привет. Нормально. __end__».

Получение ответов на вопросы рекуррентной нейронной сетью происходит по следующей схеме (рис. 2)

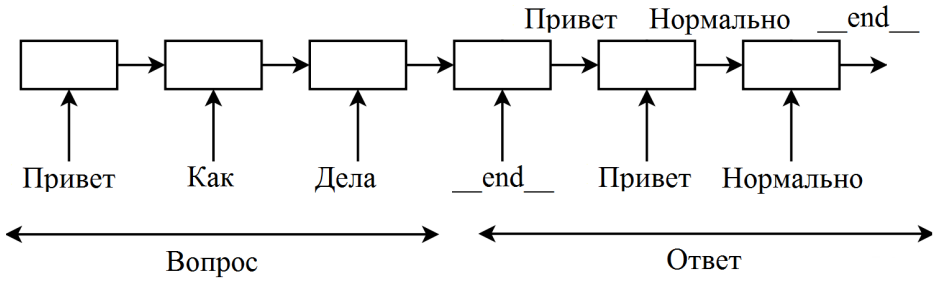


Рис. 2. Получение ответа на вопрос рекуррентной нейронной сетью.

Объём слоёв HIDDEN должен позволять запоминать весь набор предложений. Сеть обучается методом обратного распространения ошибки.

4. Программная реализация

Программа чат-бота была реализована в качестве Android-приложения с возрастным ограничением 18+. В приложении доступно несколько режимов чат-бота:

1) на базе внешнего API www.pandorabots.com;
2) на базе оригинального обобщения языка AIML;
3) на базе рекуррентной нейронной сети.

В качестве морфологического анализатора была использована свободная JavaScript-библиотека для обработки текстов на русском языке Az.js.

Для создания и обучения рекуррентной нейронной сети использовалась свободная JavaScript-библиотека RecurrentJS.

Во всех режимах доступен автоперевод ответов на язык вопросов пользователя на базе сервисов Yandex Translate API и Bing Translate API. Также в качестве внешних сервисов используется сервис поиска картинок Custom Search API и Bing Image Search API. Поиск знаний реализован на базе Google Knowledge Graph Search API. Поиск музыки реализован на базе SoundCloud API. Калькулятор, прогноз погоды, курс валют, время реализованы на базе Wolfram|Alpha API.

Приложение доступно по адресу:
`https://play.google.com/store/apps/details?id=svlab.chatbot`

5. Заключение

Программно реализованы режимы чат-бота на базе внешнего API, расширенной разметки AIML и рекуррентной нейронной сети. Расширение разметки AIML включает в себя новые теги theme и history для более эффективного

поиска релевантных вопросов и ответов. В интерпретатор внедрён модуль морфологического анализатора отдельных слов и приведения их в базовую форму в процедуре поиска релевантных ответов. Рекуррентная нейронная сеть позволяет получать ответы на вопросы, которых не было в базе знаний, с помощью способности сети к обобщению. Приложение ChatBot доступно для платформы Android в Play Маркете.

Литература

1. Провотар А.И., Клочко К.А. Особенности и проблемы виртуального общения с помощью чат-ботов // Научные труды Винницкого национального технического университета. 2013. № 3. С. 2.
2. Vinyals O. Quoc Le A Neural Conversational Model // arXiv preprint arXiv:1506.05869, 22 Jul 2015.

CHATBOT PROGRAM — CHATBOT OR VIRTUAL COMPANION

V.A. Shovin
Scientist Researcher, e-mail: v.shovin@mail.ru

Omsk Branch of the Institution of the Russian Academy of Sciences Institute of Mathematics. S. Siberian Branch of RAS

Abstract. The program of the virtual person or chatbot based on an external API, algorithm of search of responses in the knowledge base of expanded markup AIML, as well as the recurrent neural network are developed. The algorithm allows us to find answers to relevant questions from the knowledge base. The procedure of relevant responses sorting includes regular expression search, search by category, search by history and search for the best match of words in questions. The recurrent neural network is defined on the set of words of Q & A knowledge base

Keywords: chatbot, virtual companion, AIML, recurrent neural network.

Дата поступления в редакцию: 28.06.2016

Математические структуры и моделирование
2016. № 4(40). С. 102–115

УДК 004.7

САМООРГАНИЗУЮЩИЕСЯ MESH-СЕТИ ДЛЯ ЧАСТНОГО ИСПОЛЬЗОВАНИЯ

С.В. Гусс
преподаватель, e-mail: infoguss@gmail.com

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В статье рассматриваются особенности самоорганизующихся mesh-сетей, приводятся необходимые определения и пояснения, выделяются главные характеристики. Даётся небольшой обзор примеров существующих в мире mesh-сетей общего использования и проприетарных технологий, в основе которых лежат принципы самоорганизации. Обсуждается вопрос создания и развёртывания частной mesh-сети, проблем, возникающих с этим и способов их решения.

Ключевые слова: mesh-сети, самоорганизующиеся сети, 802.11s, WMN.

Введение

Самоорганизующаяся сеть (ad hoc network) — беспроводная, динамическая, децентрализованная, мобильная сеть, не имеющая постоянной структуры (площадь покрытия такой сети не обязана быть постоянной; сеть с переменной топологией).

Из представленного выше определения можно выделить четыре главных свойства самоорганизующейся сети:

1. **Беспроводная**. Можно использовать существующие протоколы, стандарты и технологии беспроводной связи, такие, как, например, IEEE 802.11 Wi-Fi (для локальных и городских сетей), IEEE 802.15.1 Bluetooth (для бытовых устройств), IEEE 802.15.4 Zigbee (для датчиков).

2. **Динамическая**. Сеть настраивается сама, автоматически, без участия человека. Требует обмена управляющей, а в некоторых случаях и статистической информацией между узлами, участвующими в организации сети приёма и передачи данных (например, для балансирования нагрузки и отправки сведений об изменении топологии сети).

3. **Децентрализованная**. В таких сетях нет единого управляющего центра. Каждое устройство сети (абонент) — активный участник процесса организации приёма и передачи данных между сетевыми узлами. В частных случаях абонент может находиться только в одном из режимов: простой клиент (station), точка доступа (access point), прямое соединение (peer-to-peer).

4. **Мобильная**. Узлы, составляющие сеть, могут перемещаться в пространстве, могут выбывать из сети, а новые устройства, в свою очередь, присоединяться к сети и участвовать в её организации. Это вполне могут быть персональные компьютеры, ноутбуки, смартфоны, планшеты, интеллектуальные датчики и другие устройства (подойдёт и Raspberry PI, пример для справки — проект Meshberry [1]), способные принимать и передавать данные, основываясь на установленных в данный момент правилах (на которые влияет структура сети, её связность).

В типичной локальной самоорганизующейся сети устройства пользователей просто подключаются к точкам доступа, которые могут перемещаться в пространстве, как и сами пользователи со своими устройствами. Естественно, что здесь не обойтись без беспроводной технологии подключения, особенно если учесть, что сами точки доступа могут находиться высоко в воздухе или постоянно перемещаться (если, например, установлены в патрульных машинах, ну, или в квадрокоптерах, доставляющих товары, заказанные в интернет-магазине). Такую сеть практически невозможно настроить статически, её топология меняется, а устройства подключаются и отключаются, причём как пользовательские, так и промежуточные сетевые, составляющие основу сети. Поэтому функции настройки и маршрутизации должны быть автоматизированы и оптимизированы по времени организации работоспособности.

Один из типов самоорганизующихся сетей — mesh-сети (ячеистые сети). Это одноранговые (P2P, peer-to-peer) распределённые сети, в которых каждый абонент соединяется со своими ближайшими соседями и может принимать на себя функцию маршрутизатора. Подобные сети способны реализовать высокую отказоустойчивость и применяются в таких областях, как [2]: военная связь, интеллектуальные транспортные системы, локальные сети, беспроводные сенсорные сети; также могут быть использованы в бизнесе, образовании, сфере развлечений, промышленности и коммерции.

В стандарте IEEE 802.11s (Wireless LAN Mesh Network Technology) представлено наиболее полное описание самоорганизующихся сетей.

1. Актуальность mesh-сетей

Mesh-сети могут интегрировать в себе различные сетевые и радиотехнологии. Самый распространённый на сегодняшний день стандарт беспроводного соединения устройств — Wi-Fi. Поэтому и сами mesh-сети строятся в основном на этой технологии. А используются такие сети преимущественно для организации локальных (LAN) и городских (MAN) сетей.

Особая актуальность mesh-сетей определяется развитием микроэлектроники, появлением множества различных устройств, способных работать автономно долгое время, имеющих особенность многократной смены режима (online-нахождения в сети и offline-выхода из сети) и нуждающихся в обмене информацией со своим окружением, а возможно и с управляющим или информационным центром (поддержка концепции IoT — Internet of Things, Интернета вещей).

Одно из преимуществ mesh-сетей — независимость. Можно создать свою мобильную сеть передачи данных, которую никто не контролирует, и всё время оставаться на связи. Чем больше абонентов — тем плотнее и надёжнее сеть. Таким образом, можно всегда оставаться на связи в местах, где отсутствует сетевая инфраструктура. Это может оказаться весьма полезно в районах повышенного риска (где вынуждены работать специальные бригады), в местах дикой или неосвоенной природы (где проводят исследования учёные, археологи, геологи, туристы) и удалённых населённых пунктах, где абсолютно каждое абонентское устройство (например, смартфон местного жителя или станция, установленная в транспортном средстве участкового) может принимать участие в процессе передачи важной информации до адресата.

По замыслу, в полноценных mesh-сетях нельзя перехватить трафик и запретить распространение информации. Это, в свою очередь, может противоречить государственным законам конкретной страны или региона. В то же время, государственные структуры и военные ведомства по этим же самым причинам заинтересованы в освоении и организации подобных сетей.

Примеры общественных mesh-сетей:

1. **Guifi** [3]. Каталония, Валенсия. Создана в ответ на отсутствие «доступного» (по цене и качеству) интернет-провайдера. Есть специальные удалённые серверы доступа в Интернет, есть mesh-сети, ряд абонентов которых имеют доступ к этим серверам, а через них, в свою очередь, до Интернета могут добраться и другие участники сети.

2. **AWMN** (Athens Wireless Metropolitan Network) [4]. Греция, Афины. Для маршрутизации используются протоколы BGP (Border Gateway Protocol) и OLSR (Optimized Link-State Routing).

3. **WasabiNet** [5]. США, Сент-Луис. Один из ярких примеров городских сетей, покрывающих отдельные улицы и предлагающих доступ в Интернет как бесплатно, так и за деньги, с определёнными услугами и более высокой скоростью доступа. Для маршрутизации также используется OLSR.

4. **OLPC** (One Laptop Per Child) [6]. Страны третьего мира. Цель — возможность организовать классную/аудиторную mesh-сеть, не используя специальное коммутационное оборудование, а только ноутбуки детей/школьников, выдаваемые им в рамках проекта.

5. **Hyperboria** (прошлое название — Project Meshnet) [7]. В рамках проекта реализован протокол Cjdns. С его помощью можно построить свою инфраструктуру обмена информацией. Но самое главное, протокол решает проблему оптимального перераспределения трафика и перенаправления нагрузки [8]. В отличие от OLSR и других протоколов (например, B.A.T.M.A.N.) позволяет объединять отдельные сети и шифровать трафик. Проект также доступен и для русскоязычного сообщества [9].

Лидирующим протоколом на сегодняшний день является Cjdns. Иногда Wi-Fi mesh-сети называют Cjdns-сетями, а протокол Cjdns «движком маршрутизации». Не менее популярен и протокол OLSR (RFC 3626) — протокол маршрутизации для сетей MANET (Mobile Ad hoc Network, RFC 2501). MANET — полностью децентрализованная самоорганизующаяся сеть со случайным соединением узлов.

2. Примеры технологий на базе mesh-сетей

Далее будет представлен небольшой обзор организации и возможностей четырёх готовых, проприетарных технологий самоорганизующихся mesh-сетей для совершенно разных сфер применения, чтобы подчеркнуть универсальный характер данного подхода к построению беспроводных соединений.

Z-Wave [10]

Это технология «интеллектуального дома». Каждое устройство Z-Wave сети является одновременно и приёмником, и передатчиком. Пользователь технологии имеет полный контроль над бытовыми устройствами своего дома (освещение, электроприборы, кондиционирование, отопление, видеонаблюдение). Хозяин дома, имея специальный пульт, может управлять автоматизированной системой. Сама сеть состоит из электронных бытовых приборов обслуживания, контроллеров, датчиков (движения, света, температуры, . . .) и других модулей. Есть возможность дистанционного управления через Интернет (например, по смартфону). Таким образом, если хозяин дома забыл выключить утюг или не закрыл дверь гаража, это всегда можно проверить, более того, удалённо можно даже выключить забытое устройство, закрыть дверь, спрятать что-нибудь, перекрыть кран, погасить свет (если система сама до этого «не додумалась»).

Всем управлением занимается специально разработанная микросхема Single Chip и модуль с радиочастотным трактом и антенной. В качестве процессорного ядра выступает хорошо известный и проверенный временем 8051-совместимый контроллер (вычислитель).

Компания Sigma Designs предлагает комплект для разработчиков Z-Wave Development Kit для создания и программирования своей собственной беспроводной mesh-сети на базе протокола Z-Wave.

Типы узлов в Z-Wave mesh-сети: 1) controllers — могут осуществлять маршрутизацию; 2) slaves — посылают, принимают и исполняют команды, могут работать как ретрансляторы сигналов. Максимальное количество узлов — 232. Для упорядочивания устройств используются специальные идентификаторы — Home ID и Node ID. Устройства с абсолютно уникальными Node ID, где бы они не находилось, в пределах дома должны иметь одинаковый Home ID.

Ruckus [11]

Это очередной пример «умной» mesh-сети для создания экономичной, высокопроизводительной, беспроводной локальной сети. Производитель основным пользователем технологии видит коммерческие предприятия и заявляет о таких свойствах, как самоорганизация, самооптимизация и самовосстановление.

Устройства сети:

1. Контроллер сети ZoneDirector. Центральный элемент, через который конфигурируется и управляется вся сеть. Через ZoneDirector администратор сети может просмотреть карту топологии сети и клиентов, внести необходимые изменения.

2. Точки доступа Root AP (корневая точка доступа соединяется с ZoneDirector) и Mesh AP (точка доступа сети, выбирающая оптимальный путь для передачи сигнала другим узлам и корневой точке доступа).

AirTies Mesh [12]

Технология решает проблему слабого сигнала в беспроводной локальной сети и ограниченного покрытия (например, в многоэтажных бетонных зданиях). Устройство подключается не к одной, а к нескольким точкам доступа, действующим в режиме репитера. Точки доступа образуют mesh-сеть, устройство пользователя подключается к этой сети через точку с самым сильным сигналом. Цель mesh-сети — предоставить оптимальный маршрут до маршрутизатора (чтобы получить доступ к другим сетям, в частности к Интернету).

RAELink3 Mesh [13]

Это «портативный модем, обеспечивающий беспроводную связь большого радиуса действия между удалёнными портативными газоанализаторами и базовой станцией компьютера для комплексного мониторинга» [14]. К модему могут подключаться до 8 газоанализаторов и до 64 газоанализаторов, оборудованных модемом. Три режима работы: 1) дистанционный — для подключения к базовой станции; 2) базовая станция — для работы в качестве главного модема; 3) ретранслятор.

3. Маршрутизация

Поскольку основной характеристикой mesh-сетей является переменная топология, самый актуальный вопрос, который необходимо решить для её организации, — как эффективно построить маршрут между источником сообщений и адресатом.

Протоколы маршрутизации делят на два класса:

1. *Проактивные* (табличные). Каждый узел строит свою таблицу маршрутизации и делится информацией об изменении топологии сети со своими соседями. Примеры: OLSR, DSDV (Destination-Sequenced Distance Vector).

2. *Реактивные* (работающие по запросу). Таблицы маршрутизации не строятся, маршрут составляется по мере необходимости. Используется широковещание для определения пути отправки сообщения. Примеры: DSR (Dynamic Source Routing), AODV (Ad hoc On-Demand Distance Vector).

На практике используют гибридные протоколы. Крупная сеть делится на подсети. Внутри подсетей на узлах создаются таблицы маршрутизации. Маршрут через подсети составляется через широковещательные запросы и определение наиболее актуального в данный момент пути. Примеры гибридных протоколов: HWMP (Hybrid Wireless Mesh Protocol).

Существуют также протоколы геомаршрутизации, которые предполагают брать информацию о местонахождении участников сетевого взаимодействия через спутниковые системы (ГЛОНАСС, ГЛОНАСС/GPS). Примеры протоколов: GAF (Geographical Adaptive Fidelity), GPSR (Greedy Perimeter Stateless Routing), LAR (Location-Aided Routing).

Также как и для обычных компьютерных сетей (с постоянной топологией) протоколы маршрутизации могут делиться на протоколы вектора расстояния и протоколы состояния канала (с комплексной метрикой маршрутов). Вопрос выбора правильной метрической системы весьма актуален для создания сетей особого назначения.

Обзор актуальных протоколов маршрутизации, их преимуществ и недостатков можно найти в работе [15]. В работе [16] протоколы анализируются по определённой системе критериев, учитывающей функциональные, топологические, дистанционные, ресурсные и другие особенности организации самоорганизующейся сети. Более подробно о возможных критериях и метриках выбора оптимальных маршрутов в mesh-сетях можно узнать в работе [17]. Кратко можно отметить, что для построения маршрута могут учитываться [17]: длина пути, надёжность, задержки, пропускная способность, загрузка и стоимость передачи данных.

4. Создание и развёртывание mesh-сети

Идея самоорганизующейся сети заключается в том, что если, к примеру, вы живёте в многоэтажном многоквартирном доме, и в каждой (или почти в каждой) квартире есть маршрутизатор (беспроводная точка доступа с функцией маршрутизации и выходом в Интернет, которые сегодня весьма распространены), то эти маршрутизаторы можно объединить в mesh-сеть. Преимущества сети будут состоять в том, что если у одного из пользователей пропадёт Интернет или «упадёт» скорость, можно будет воспользоваться Интернетом соседей (выбор оптимального маршрута). Или если сеть не перегружена, то большой объём информации можно загрузить одновременно через несколько каналов. Такая схема будет работать, если соседи находятся в дружеских отношениях. В реальности это можно организовать через посредника — управляющую компанию, которая сама развернёт сеть и будет предоставлять абонентские услуги доступа к оборудованию. В случае с предприятием или небольшой организацией вопрос собственности уже менее принципиален, а организацией сети может заняться специальный отдел, отвечающий за вопросы автоматизации и предоставления доступа к компьютерной сети.

Проблем становится больше, когда сеть находится в постоянном движении, например, сеть транспортных средств путешественников или просто геологов-исследователей, перемещающихся целевыми группами. В этом случае на помощь приходят компании со своим готовым сетевым оборудованием и предлагают услуги. Сами разработки (реализация технологии) являются проприетарными и не выставляются на всеобщее обозрение.

На сегодняшний день нет полной и точной информации о том, как на основе

имеющихся свободных программных и аппаратных средств создать и развернуть свою mesh-сеть, готовую предоставить полноценные сервисы самоорганизующейся частной сети своим пользователям.

Теоретически, и даже практически, вопрос о настройке существующих специальных операционных систем коммуникационного оборудования и протоколов в той или иной степени раскрыт. Можно без труда объединить в ячеистую топологию Wi-Fi маршрутизаторы и даже подключить к ним абонентов. Проблема состоит в том, что для обычных пользователей, которым нужен готовый, надёжный функционал и набор сервисов, вся эта информация практически ни о чём не говорит, за исключением тех, кто как-то связан с областью компьютерных сетей и информационных технологий. Более того, современный пользователь не хочет покупать специализированное оборудование или тратить деньги на его аренду. Необходимо задействовать то, что у него уже есть. Для жизни в современном развивающемся обществе человеку просто необходим смартфон для ориентации в городе, для связи с друзьями и близкими, для доступа к обслуживающей население информационной среде. Поэтому можно считать, что у каждого, кому понадобилась своя mesh-сеть, найдётся смартфон. Самая распространённая операционная система для смартфонов — Android. Следовательно, для удовлетворения спроса имеет смысл ориентироваться на пользователей, у которых всегда с собой есть смартфон с операционной системой Android.

Можно найти довольно много советов о том, как подключиться к «сети будущего», где нет цензуры и вся информация пользователя доступна только тем, кому он её предоставляет (адресатам). Но это всего лишь предполагаемая замена или альтернатива сети Интернет. Если требуется развернуть безопасную сеть, можно воспользоваться технологией и протоколом Cjdns. Весь трафик сети будет шифроваться автоматически. Протокол гарантирует приватность (никто не перехватит личную информацию, предмет разговора будет сокрыт), но не гарантирует анонимность (всегда можно узнать, кто и с кем общается). Для работы с сетью устройство должно поддерживать работу протокола IPv6. Проблема Cjdns состоит в том, что вы создаёте не совсем свою сеть, а становитесь частью проекта Hyperboria, так называемого независимого Интернета будущего. Тем не менее, это пример mesh-сети, стало быть, в своих разработках можно ориентироваться на огромный опыт, чтобы избежать проблем. Особого внимания заслуживает android-проект Open Garden (набор приложений) [18], позволяющий без особых проблем подключиться к mesh-сети пользователям приложения, тем не менее у приложения всё та же направленность — «общий открытый Интернет», не для частного использования.

Когда речь заходит о mesh-сетях, часто можно встретить такую технологию, как OpenWrt [19]. Немало обсуждений встречается в Интернете по поводу того, как настроить на маршрутизаторе операционную систему OpenWrt для создания mesh-сети. Более того, маршрутизаторы на OpenWrt могут соединяться с Cjdns-узлами, а это значит, что технологии разделяют общую идеологию самоорганизующейся сети. Беда состоит в том, что OpenWrt ориентирован не на мобильные, а на коммуникационные устройства, которые объединяются в mesh-сеть, а устройства пользователей пользуются преимуществами такой сети. Та-

ким образом мобильность реализуется только по отношению к пользователям, перемещающимся в радиусе охвата установленных точек доступа. Вопрос, таким образом, остаётся открытым — насколько реально организовать частную сеть группы людей (тех же геологов как пример перемещающихся по различным местам пользователей возможной сети), имеющих при себе только смартфоны (например, на базе Android, которые могут работать как маршрутизатор, обычно современные Android-смартфоны поддерживают данный функционал). Поскольку в каждом смартфоне есть Wi-Fi и Bluetooth, то рождается аналогия с проводной сетью — по Bluetooth (как замена сериального/последовательного соединения) объединить промежуточные устройства (центральные смартфоны-маршрутизаторы), а по Wi-Fi — открыть доступ участникам, абонентам сети. Или же, наоборот, если два маршрутизатора находятся далеко друг к другу, то соединить их по Wi-Fi, а подключение абонентов осуществлять по Bluetooth. Если развить тему и предположить, что группа геологов перемещается в лесу и у них есть квадрокоптер (или подобное устройство), на который можно установить Wi-Fi ретранслятор, то идея становится вполне жизнеспособной.

Что же в этом случае делать разработчикам для удовлетворения пользовательской потребности? Из вышесказанного следует, что нужно разработать некое универсальное промежуточное программное обеспечение для поддержания надёжной mesh-связи между абонентскими устройствами, которые подключаются друг к другу посредством доступных коммуникационных технологий (Wi-Fi, Bluetooth, ZigBee, IrDA). Причём необходимо учитывать, как будет происходить обмен информацией о топологии и подключениях, как вся эта работа по организации многосвязной сети будет влиять на продолжительность работы аккумуляторов пользовательских устройств. Нужен хорошо продуманный протокол, который учитывает все ограничения устройств и предполагает возможность соответствующей подстройки.

Можно предложить и более амбициозную идею. Учитывая уровень развития электроники и наличия современных программируемых микропроцессорных средств (тех же микроконтроллеров и более интегрированных решений), реализовать на низком уровне подобный представленному выше протокол. Протокол, который можно будет интегрировать в миниатюрное устройство (назовём его UIWRT — Universal Intellectual Wireless Receiver/Transmitter, универсальный беспроводной приёмо-передатчик), способное участвовать в беспроводных соединениях. Пользовательские устройства будут подключаться к умным UIWRT, а те в свою очередь организовывать сеть, mesh-сеть.

5. Моделирование

Моделирование работы самоорганизующейся сети остаётся актуальной задачей, поскольку далеко не все места планета покрыты общедоступными системами связи и в каждой сети могут возникать свои потребности. Одна из универсальных проблем — информирование населения (или группы людей) в условиях чрезвычайных ситуаций. В связи с чем становится обоснованной потребность разработки целого программно-аппартного комплекса для «организа-

ции локального информационного пространства в рамках зоны чрезвычайной ситуации с возможностью взаимодействия её с внешними информационными сетями» [20], а также возможность «развернуть сеть в сроки адекватные ситуации» [20]. Такая система должна обеспечить быстрое развёртывание с минимальными расчётами, оперативное самовосстановление, доступ к мобильным устройствам.

Для построения своей частной mesh-сети можно опираться на стандарт IEEE 802.11s. Несмотря на то, что стандарт рекомендует использовать протокол маршрутизации HWMP (по умолчанию), он не ограничивает использование других протоколов. В частности, можно использовать свою собственную разработку. Поскольку возможности доступных мобильных устройств не ограничиваются технологией Wi-Fi, стоит также воспользоваться рекомендациями стандарта IEEE 802.21 для организации взаимодействия сетей разных типов (802.3, 802.11, 802.15, . . .). С другой стороны и сам стандарт рекомендует использовать несколько частотных каналов. Таким образом, на устройстве может быть несколько различных интерфейсов передачи.

Ведутся также работы над открытым стандартом open80211s [21].

Роли устройств (согласно 802.11s): 1) как в стандартной 802.11 сети: а) станция (STA) — пользователь сети, не участвующий в организации инфраструктуры, просто подключается к б) точке доступа (AP); 2) в mesh-сети появляются ко всему: а) узлы mesh-сетей (MP), реализующие mesh-службы для поддержки необходимого функционала, б) точки доступа mesh-сети (MAP) — MP, поддерживающие функции AP, в) порталы mesh-сети (MPP) — это MP, открывающий доступ к внешним сетям.

Формат кадра остался тем же (что и в общем стандарте 802.11), только в поле данных добавляется mesh-заголовок (mesh-флаги, время жизни mesh-пакета, номер mesh-пакета в последовательности, mesh-расширение адреса).

Вопросы, требующие решения для эффективного взаимодействия: маршрутизация, балансирование нагрузки, передача служебных сообщений.

Учитывая всё разнообразие задач, встающих на пути организации эффективной mesh-сети, нужно подумать также над выбором соответствующего задаче алгоритма назначения канала. Алгоритм должен определить, какой интерфейс следует использовать для конкретного случая, учитывая характеристики канала связи. В качестве примера можно привести алгоритм Hyacinth [22, 23]. Для определения канала алгоритм сравнивает пропускную способность соединения с ожидаемой нагрузкой. Для вычисления общей пропускной способности алгоритм учитывает объём трафика и загрузку каналов на маршрутах. Пропускная способность всех соединений не должна быть меньше ожидаемой нагрузки.

Алгоритмы, работающие в mesh-сети должны быть основаны на математической модели. Пример математической модели для организации mesh-сети стандарта IEEE 802.16 (беспроводной городской сети) представлен в работе [24]. В работе рассматривается задача распределения подканалов для оценки характеристик возможных конфигураций сети. Для сравнения полученных результатов используются графы Кенига. Для построения модели сети, согласно [24], следует учитывать: неоднородность устройств и подсетей, динамическое распреде-

ление частотных ресурсов, обеспечение показателей QoS (качество обслуживания), режим управления (распределённый или централизованный), влияние интерференции между устройствами, выделение радиоканалов, протокол маршрутизации, технологические особенности (дальность связи, интенсивность трафика). Чем полнее будет математическое описание, тем проще будет создавать реализующий протокол.

6. Прототипирование

Естественное продолжение этапа моделирования — прототипирование или создание экспериментального устройства, реализующего модель для проверки на практике целесообразности дальнейшей разработки.

Далее будет представлен обзор существующих на рынке, готовых к непосредственному использованию модулей аппаратного обеспечения для реализации беспроводных специализированных устройств на базе технологии Wi-Fi. Модули пригодны не только для прототипирования, но также и для проектирования готовых устройств. В обзор включены недорогие по цене решения (согласно распространённому мнению сообщества разработчиков встраиваемых устройств). Более дорогой вариант — использование Raspberry PI и подключаемых к ней внешних модулей. Использование Raspberry PI не всегда целесообразно ввиду более сильного (порой значительно) расходования ресурсов электропитания проектируемого устройства по сравнению со специализированными, менее производительными, микроконтроллерными средствами.

CC3100MOD [25]. Специализированный сетевой MCU (Microcontroller Unit, микроконтроллер) для IoT. Это SiP модуль (System in Package). Оборудован беспроводным интерфейсом Wi-Fi 802.11 b/g/n. Скорость передачи полезных данных до 16 Мбит/с (для UDP) и до 12 Мбит/с (для TCP). Может работать в режимах station (клиент), AP (точка доступа) или Wi-Fi Direct (для независимой работы без маршрутизатора). Выполнен в виде микросхемы на базе ядра ARM. Реализует стек протоколов Wi-Fi, TCP/IP (версия 4), шифрование (WPA2, SSL 3.0, TLS 1.2), сервисы (ARP, ICMP, DHCP, DNS). Можно одновременно открыть 8 сокетов или 2 защищённых соединения. Предполагается, что встраиваемое решение на базе процессора может работать год и более от двух батарей AA. Загрузка программ возможна через UART и SPI. Для реализации всех возможностей рекомендуется подключение внешней памяти Serial Flash через SPI-интерфейс. Эта память используется для хранения пользовательских данных, файлов настройки, сертификатов. Для тактирования используются две частоты (два тактовых генератора) — 1) для часов реального времени (32768 Гц), 2) для внутреннего процессора и подсистемы WLAN (40 МГц). Поскольку микроконтроллер поставляется с управляющим драйвером, от разработчика программных приложений не требуется реализовывать с нуля стандартные протоколы или службы, а можно просто использовать высокоуровневый интерфейс команд. Полное описание процессора и принципов его работы можно найти в [26, 27].

ESP8266 [28]. Микроконтроллер, оборудованный беспроводным интерфей-

сом Wi-Fi 802.11 b/g/n. Режимы: Wi-Fi Direct, soft-AP (software enabled Access Point, программно реализованная точка доступа). Поддерживаются WEP и WPA/WPA2. Управление модулем осуществляется через UART посредством набора AT-команд. Модуль можно перепрограммировать (перепрошивать). Интегрированный стек TCP/IP. Встроенный датчик температуры. Программы выполняются из внешней памяти (микроконтроллер не имеет энергонезависимой памяти программ), подключаемой через SPI. Присутствует 10-битный АЦП. На базе микроконтроллера выполнен модуль ESP8285 со встроенной флэш-памятью в 1 Мбайт. ESP8266 построен на 32-битном ядре Tensilica Xtensa L106 на 80 MHz. Дальнейшее развитие — **ESP32** [29] (двухядерный процессор Tensilica LX108 с частотой до 240 МГц, поддержка Bluetooth).

Particle Photon [30]. Модуль представлен в виде небольшой платы. Плата включает в себя 1) микроконтроллер STM32 ARM Cortex M3 на 120 МГц, 2) Wi-Fi модуль Broadcom BCM43362, реализующий стандарты 802.11b/g/n, 3) 1 Мбайт флэш-памяти. Работает на базе операционной системы FreeRTOS. Поддерживает режим soft-AP. Интерфейсы: SPI, UART, CAN. Для программирования загрузчика или прошивки можно использовать JTAG.

Omega2 [31]. Микрокомпьютер на базе Linux. Wi-Fi-модуль реализует стандарты 802.11b/g/n. Имеются два приёмопередатчика UART и интерфейс SPI. Процессор работает на частоте 580 МГц. 16 Мбайт памяти для хранения программ.

Из современных популярных модулей можно также выделить Oak [32], C.H.I.P. [33], Electric Imp [34].

Все рассмотренные модули выполняют примерно одни и те же функции, характерные для IoT. Разработчику доступны готовые контроллерные платы со всем необходимым микропрограммным обеспечением, предоставляющим свой API. Для быстрых экспериментов вполне могут подойти Particle Photon, Omega2 или решения на подобие Oak или Electric Imp. Для большей свободы реализации стоит использовать микроконтроллеры серии CC3100, ESP8266 или ESP32, распространённые и доступные для свободного приобретения на территории Российской Федерации.

7. Заключение

В статье кратко была рассмотрена проблема организации частной самоорганизующейся mesh-сети. Дано общее определение самоорганизующейся сети и особенностей mesh-сетей, поддерживающих идеи самоорганизации. Было показано, что проблема построения частной беспроводной mesh-сети всё ещё актуальна и ближайшее время будут появляться новые задачи, возможные решения и готовые технологии. Проведённый анализ задач организации частной mesh-сети установил необходимость создания оптимальной математической модели самоорганизующейся сети, созданной с учётом имеющихся в данной области стандартов, руководств и рекомендаций. В конце статьи был представлен обзор возможных аппаратных решений для прототипирования промежуточного сетевого оборудования для расширения охвата и увеличения возможностей проек-

тируемой mesh-сети.

Литература

1. NYC Mesh. A community-owned resilient network. URL: https://nycmesh.net/meshberry/ (Дата обращения: 24.09.2016).
2. Беспроводные самоорганизующиеся сети. URL: http://crossgroup.su/solutions/adhoc.html (Дата обращения: 24.09.2016).
3. Xarxa de Telecomunicacions de Comunus Oberta, Lliure I Neutral. URL: https://guifi.net (Дата обращения: 24.09.2016).
4. Athens Wireless Metropolitan Network (AWMN). URL: http://www.awmn.net (Дата обращения: 24.09.2016).
5. WasabiNet. URL: http://gowasabi.net/sitemap (Дата обращения: 24.09.2016).
6. One Laptop per Child. URL: http://one.laptop.org (Дата обращения: 24.09.2016).
7. Hyperboria. URL: https://hyperboria.net (Дата обращения: 24.09.2016).
8. Project Meshnet Documentation. URL: https://docs.meshwith.me/project-goals-ru.html (Дата обращения: 24.09.2016).
9. Cjdroute.net Hyperboria ex CJDNS.RU. URL: https://cjdroute.net (Дата обращения: 24.09.2016).
10. Z-Wave Russia — Умный дом. URL: http://z-wave.ru (Дата обращения: 24.09.2016).
11. Ruckus Wireless — лучшие беспроводные WiFi решения. URL: http://ruckus-wireless.ru (Дата обращения: 24.09.2016).
12. AirTies Wireless Networks — Technology. URL: http://www.airties.com/technology.html (Дата обращения: 24.09.2016).
13. RAELink 3 Mesh. Portable wireless transmitter with integrated GPS. URL: http://www.raesystems.com/products/raelink-3-mesh (Дата обращения: 24.09.2016).
14. Руководство пользователя RAELink3 Mesh. URL: http://www.raesystems.com/sites/default/files/content/resources/Manual_RAELink3_Mesh_Manual_RevA_3_RU.pdf (Дата обращения: 24.09.2016).
15. Павлов А.А., Датьев И.О. Протоколы маршрутизации в беспроводных сетях // Труды Кольского научного центра РАН. 2014. № 5(24). С. 64–75.
16. Карманов М.Л. Протокол маршрутизации для ad-hoc сетей // Вестник Южно-Уральского государственного университета. 2009. № 26(159). С. 47–51.
17. Вишневский В., Лаконцев Д., Сафонов А., Шпилев С. Mesh-сети стандарта IEEE 802.11s: протоколы маршрутизации // Первая миля. 2009. Том 10, № 1. С. 16–21.
18. Open Garden — Connecting the next billion mobile devices. URL: http://www.opengarden.com (Дата обращения: 24.09.2016).
19. OpenWrt. URL: https://openwrt.org (Дата обращения: 24.09.2016).
20. Мельников М.И., Ковтун А.С. Самоорганизующаяся сеть оперативного взаимодействия для нужд населения и специальных служб // Доклады Томского государственного университета систем управления и радиоэлектроники. 2014. № 2(32). С. 281–286.

21. open80211s. URL: http://www.o11s.org (Дата обращения: 24.09.2016).
22. Легков К.Е., Донченко А.А. Беспроводные mesh-сети специального назначения // T-Comm — Телекоммуникации и Транспорт. 2009. № 2. С. 36–37.
23. Ashish Raniwala, Tzi Chiueh. Architecture and Algorithms for an IEEE 802.11-Based Multi-Channel Wireless Mesh Network. URL: http://www.ecsl.cs.sunysb.edu/tr/hyacinth-infocom.pdf (Дата обращения: 24.09.2016).
24. Гаркуша С.В., Гаркуша Е.В., Еременко А.С. Модель распределения подканалов в беспроводной mesh-сети стандарта IEEE 802.16, представленной в виде гиперграфа // Збірник наукових праць Харківського університету Повітряних Сил. 2015. Вип. 2(43). С. 33–38.
25. CC3100MOD. SimpleLink Certified Wi-Fi Network Processor, Internet-of-Things Module Solution for MCU Application. URL: http://www.ti.com/product/CC3100MOD (Дата обращения: 24.09.2016).
26. CC3100 — сетевой процессор для «Интернета вещей». Часть I. URL: http://www.compel.ru/lib/ne/2014/10/2-cc3100-setevoy-protsessor-dlya-interneta-veshhey-chast-i (Дата обращения: 24.09.2016).
27. CC3100 — сетевой процессор для «Интернета вещей». Часть II. URL: http://www.compel.ru/lib/ne/2015/2/7-cc3100-setevoy-protsessor-dlya-interneta-veshhey-chast-ii (Дата обращения: 24.09.2016).
28. Everything ESP8266. URL: http://www.esp8266.com (Дата обращения: 24.09.2016).
29. The Internet of Things with ESP32. URL: http://esp32.net (Дата обращения: 24.09.2016).
30. Particle: Ship your IoT product. URL: https://www.particle.io (Дата обращения: 24.09.2016).
31. Omega2: $5 Linux Computer with Wi-Fi, Made for IoT. URL: https://www.indiegogo.com/projects/omega2-5-linux-computer-with-wi-fi-made-for-iot#/ (Дата обращения: 24.09.2016).
32. Oak by Digistump. URL: http://digistump.com/oak/ (Дата обращения: 24.09.2016).
33. C.H.I.P.s — Get CHIP — The World's First Nine Dollar Computer. URL: https://getchip.com/pages/chip/ (Дата обращения: 24.09.2016).
34. The Electric Imp Platform. URL: https://electricimp.com/platform/ (Дата обращения: 24.09.2016).

PRIVATE WIRELESS MESH NETWORKS

S.V. Guss
Tutor, e-mail: infoguss@gmail.com

Dostoevsky Omsk State University

Abstract. This paper is about private wireless mesh networks. There are definitions and comments for main features of ad hoc networks, survey of public mesh networks and proprietary technologies, routing, modeling and prototyping.

Keywords: mesh networks, ad hoc networks, 802.11s, WMN.

Дата поступления в редакцию: 26.09.2016

Математические структуры и моделирование
2016. № 4(40). С. 116–121

УДК 004.55, 004.657, 004.658.6, 007, 378.141.21

РАЗРАБОТКА В ОМГУ НОВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ПРИЁМА В ВУЗ

Т.А. Погромская
к.т.н., начальник отдела web-технологий управления информатизации ОмГУ,
e-mail: pta@omsu.ru

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В статье рассматривается опыт разработки и перехода на новую информационную систему приёма абитуриентов в вуз в Омском государственном университете им. Ф.М. Достоевского. Система основана на прежней разработке, внедрённой в четырёх российских вузах, использует её структуру данных и логику, но имеет новый веб-интерфейс.

Ключевые слова: информационная система, разработка, приёмная комиссия, абитуриент, веб-интерфейс.

Актуальность

Любое учебное заведение начинается с приёма контингента обучающихся. Процедура эта сложная и ответственная, особенно если речь идёт о конкурсном отборе, поэтому автоматизация процессов приёма всегда остаётся актуальной задачей.

Как было

С 2000 года в Омском государственном университете эксплуатируется информационно-аналитическая система «Абитуриент» (далее — ИАС «Абитуриент») — собственная разработка университета для приёма в вуз, реализованная на СУБД Visual FoxPro [3]. На протяжении многих лет происходило наращивание функционала системы, она прошла внедрение помимо ОмГУ ещё в трёх вузах: Томском, Челябинском, Горно-Алтайском государственных университетах. Развитие системы происходило по спирали [6], ежегодно выходила новая версия, настроенная на изменённые правила приёма. Многолетняя эксплуатация системы подтвердила правильность заложенных в ней алгоритмов, позволяющих автоматизировать все этапы проведения приёма в вуз. Идеология и технология созданной системы позволяют проводить и межвузовский конкурсный отбор. Причём каждый вуз может проводить приём заявлений и вступительные испытания по собственным правилам, а для одновременного конкурсного отбора в несколько вузов достаточно предоставить линейно упорядоченные по набранным баллам и дополнительным критериям списки абитуриентов [9].

Информационное пространство вуза, действовавшее в ОмГУ до 2014 года, представлено на рисунке ниже (см. рис. 1):

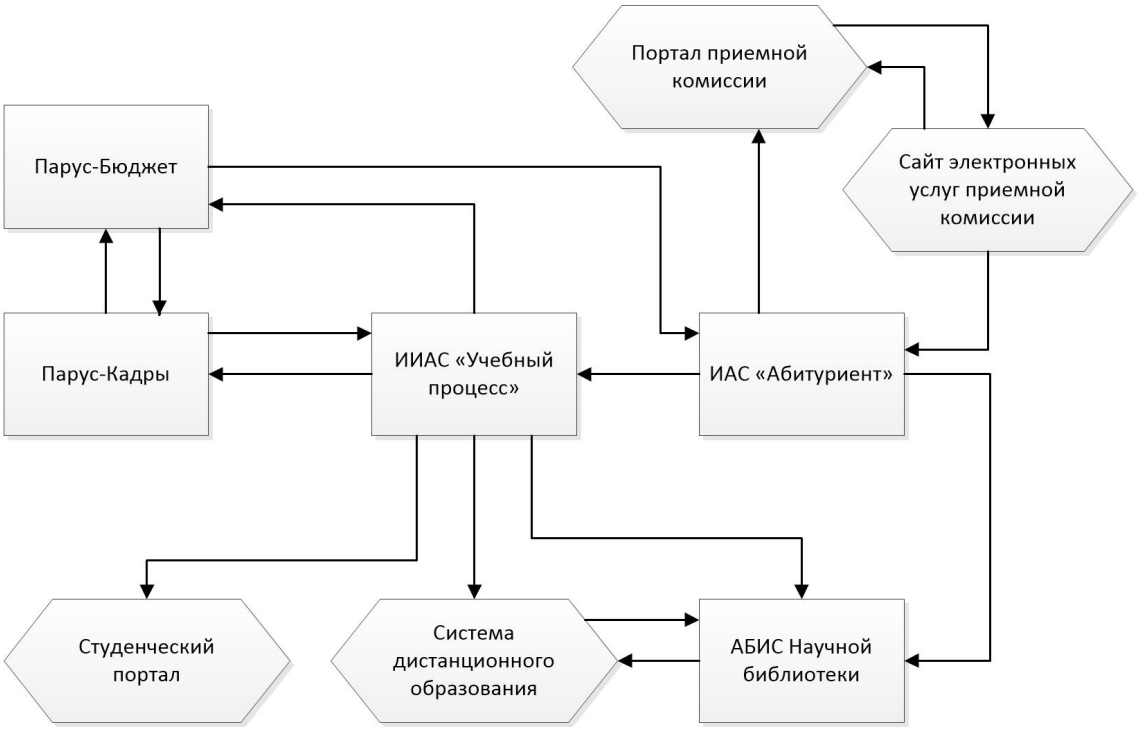


Рис. 1. Схема взаимодействия информационных систем ОмГУ до 2014 года

ИАС «Абитуриент» использовала собственную базу данных. Сведения из неё (о ходе приёма, о результатах экзаменов, о зачислении и выпуске приказов) с помощью ODBC-драйвера публиковались на портале приёмной комиссии в среде Lotus Notes/Domino [4]. По окончании приёма данные о зачисленных абитуриентах выгружались в виде xls-файлов для передачи в университетскую библиотечную систему, а по той же технологии ODBC связываясь с СУБД Oracle, загружались в интегрированную информационно-аналитическую систему «Учебный процесс» — основную систему для последующей работы со студенческим контингентом.

Сама же ИАС «Абитуриент» получала сведения о заключённых абитуриентами договорах на внеплановое обучение из финансовой системы «Парус-Бюджет», основанной также на СУБД Oracle. И принимала данные с сайта электронных услуг приёмной комиссии о поданных дистанционно заявлениях [5]. В обоих случаях использовалась всё та же технология обмена данными через ODBC.

Основное неудобство состояло в том, что приходилось инициировать действия по обмену сведениями с внешними системами (нажать кнопку в форме или вызвать соответствующий пункт меню) — эти функции возлагались на администраторов информационных систем и сайтов. Человеческое участие всегда было «слабым звеном» в цепочке передачи данных.

Как стало

В последние годы появились изменения в бизнес-логике процессов приёма, и поскольку каждые 2-3 года информационные технологии меняются примерно на 80% [8], назрела необходимость перехода на новый технологический уровень. В 2014 году решено было начать разработку новой системы, ориентированной на веб-интерфейс — кроссплатформенный, «лёгкий» и гибкий.

К основной схеме базы данных ИИАС «Учебный процесс» была добавлена схема для хранения данных новой системы приёма в вуз (условно — «Новый Абитуриент»). Она с небольшими изменениями продублировала схему данных из Visual FoxPro. Для переходного этапа была реализована технология синхронизации информации (периодическая, инициируемая администратором, по ODBC) между базами данных старой и новой систем. Новый портал приёмной комиссии, реализованный на Drupal, с первого года своей эксплуатации использовал схему данных новой системы для динамической публикации сведений о проведении приёмной кампании 2014 года и последующих [7]. Сервисы сайта электронных услуг для абитуриентов (в том числе дистанционная подача заявлений в вуз) также используют эту базу данных для своей работы. Публикация данных на портале приёмной комиссии и функционирование сервисов электронных услуг выполняются автоматически, без участия администраторов систем и сайтов.

Разработка системы «Новый Абитуриент» была разбита на этапы и условные подсистемы:

1. Разработка архитектуры системы, определение ролей и разрешений для пользователей системы (основные роли в системе «Новый Абитуриент» см. на рис. 2).

2. Разработка подсистемы просмотра данных об абитуриенте — реализация экранных форм для просмотра всей информации об абитуриенте, имеющейся в системе.

3. Разработка подсистемы работы со справочниками — реализация экранных форм для просмотра и редактирования данных, хранимых в таблицах настройки приёмной кампании и самой системы.

4. Разработка подсистемы ввода данных об абитуриентах — реализация экранных форм для работы операторов по приёму документов.

5. Разработка подсистемы формирования ведомостей. Эта часть процесса приёма останется актуальна для нашего вуза, поскольку у нас есть не только вступительные испытания, проводимые вузом самостоятельно, а также творческие экзамены факультета культуры и экзамены для поступления в магистратуру, аспирантуру и университетский колледж.

6. Разработка подсистемы построения рейтингов. Эта часть также актуальна для региональных вузов, поскольку единой информационной системы приёма в России нет, а поступающих необходимо как-то ранжировать и информировать о складывающейся конкурсной ситуации, чтобы абитуриенты могли принимать правильное решение о том, как выразить своё согласие на окончательное зачисление.

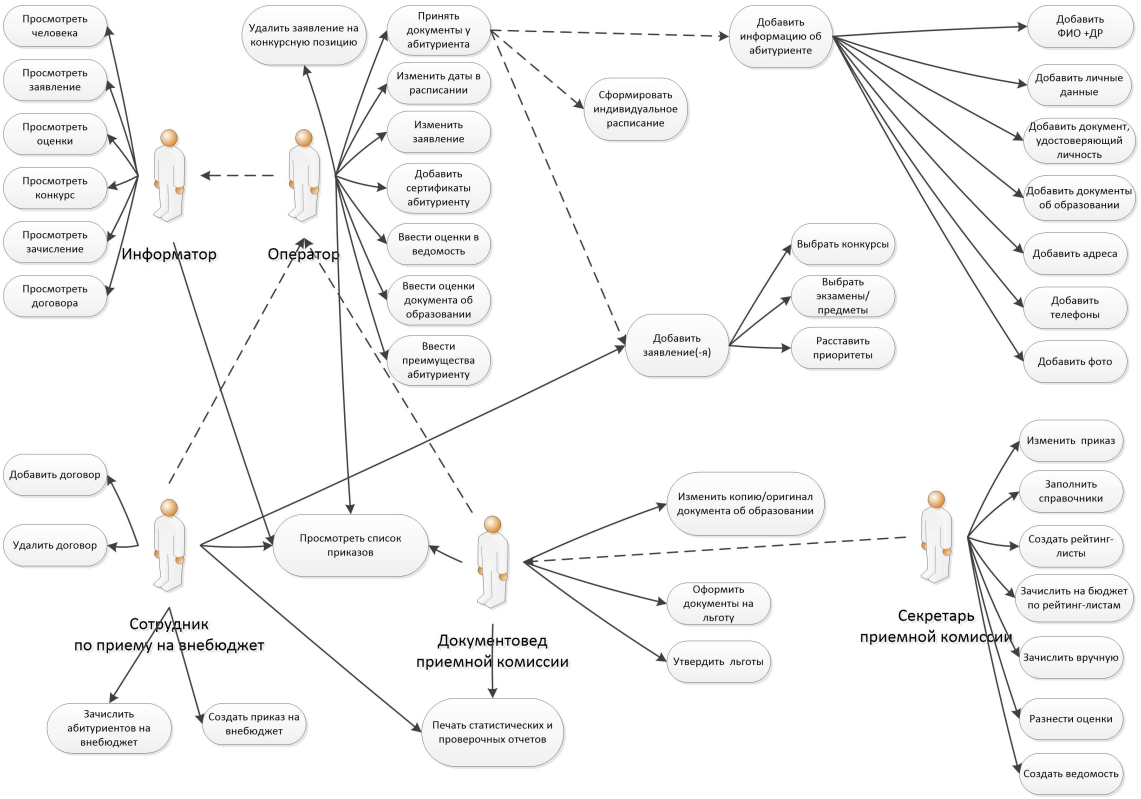


Рис. 2. UseCase-диаграмма новой информационной системы приёма

7. Разработка подсистемы издания приказов. Поскольку Порядок приёма в вузы ежегодно меняется, а процедура зачисления для вузов, возможно, скоро будет сводиться к изданию приказов по полученным из ФИС ГИА и Приёма спискам рекомендованных к зачислению, подсистема издания приказов примет свой окончательный вид после реализации предыдущих этапов.

Для разработки веб-интерфейса системы «Новый Абитуриент» используется объектно-ориентированный PHP-фреймворк Yii [2], реализующий парадигму MVC (Model-View-Controller — «модель-представление-контроллер» — схема использования нескольких шаблонов проектирования, с помощью которых модель приложения, пользовательский интерфейс и взаимодействие с пользователем разделены на три отдельных компонента таким образом, чтобы модификация одного из компонентов оказывала минимальное воздействие на остальные [1]).

Как будет

В период приёма 2015 года новая система прошла опытную эксплуатацию в части подсистемы просмотра данных об абитуриенте. По итогам эксплуатации были доработаны экранные формы для просмотра личных данных. Во время приёма 2016 года планируется расширить круг пользователей этой подсистемы. Также планируется, что опытную эксплуатацию в 2016 году пройдёт подси-

стема работы со справочниками, доступная пользователям с ролью «Секретарь приёмной комиссии», — для просмотра и редактирования данных по настройке системы.

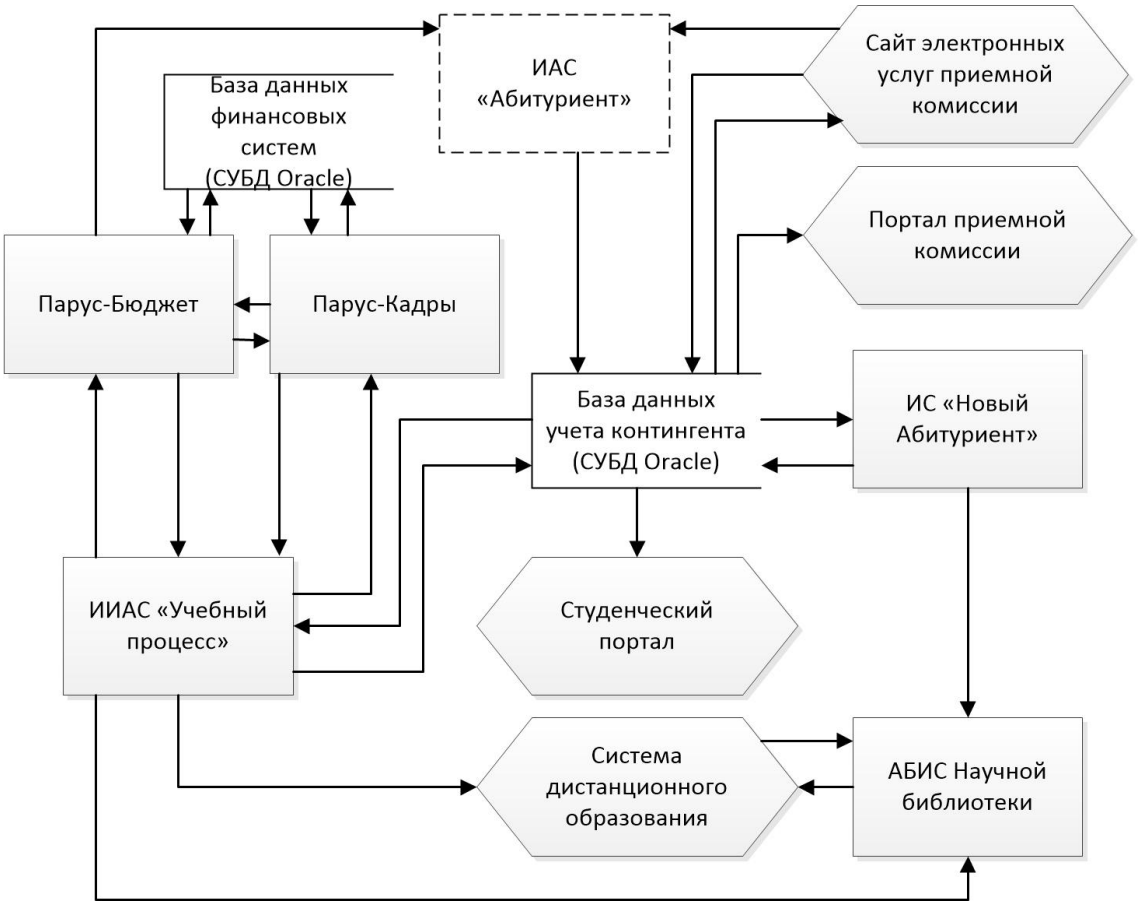


Рис. 3. Схема взаимодействия информационных систем ОмГУ с 2014 года

Как видно из рисунка 3, теперь в любой момент ИАС «Абитуриент» можно исключить из схемы взаимодействия информационных систем. Планируется перейти на «Новый Абитуриент» и полностью отказаться от старой системы ориентировочно к приёмной кампании 2017 года (как только новая информационная система будет реализована до подсистем ввода данных по приёму документов и формирования ведомостей).

Благодарности

Выражаю благодарность коллективу разработчиков управления информатизации ОмГУ и лично начальнику управления Епанчинцевой Ольге Леонидовне за совместную работу по созданию новой системы приёма в Омский государственный университет им. Ф.М. Достоевского.

Литература

1. Model-View-Controller. Материал из Википедиа. URL: `https://ru.wikipedia.org/wiki/Model-View-Controller` (дата обращения: 20.06.2016).
2. Yiiframework. URL: `http://www.yiiframework.com/about/` (дата обращения: 20.06.2016).
3. А.с. 2003611045 РФ, Роспатент. Информационно-аналитическая система «Абитуриент» (ИАС «Абитуриент») / Горнева И.С., Епанчинцева О.Л., Захаров А.М., Картешкина Е.В., Костюшина Е.А., Погромская Т.А., Рапаева И.А., Сергеева Т.И. (RU). № 2003610757: Заяв. 07.04.2003. Опубл. 30.04.2003. Бюл. № 3(44). С. 98.
4. Епанчинцева О.Л., Земсков И.А. Интеграция информационных ресурсов ОмГУ в сеть интернет // Математические структуры и моделирование. 2000. № 6. С. 143–146.
5. Епанчинцева О.Л., Погромская Т.А., Редреев Д.Г. Дистанционная подача документов в вуз. ИАС «Абитуриент» // Компьютерные учебные программы и инновации. 2005. № 2. С. 14.
6. Жизненный цикл программного обеспечения. URL: `http://qaevolution.ru/zhiznennyj-cikl-programmnogo-obespecheniya/` (дата обращения: 20.06.2016).
7. Погромская Т.А. Опыт создания новой версии портала приёмной комиссии вуза (на примере ОмГУ) // Математические структуры и моделирование. 2016. № 1(37). С. 91–96.
8. Всяких Е.И. [и др.] Практика и проблематика моделирования бизнес-процессов. М. : ДМК Пресс; М. : Компания АйТи, 2008. 246 с.
9. Хорошевский М.В., Епанчинцева О.Л. [и др.] Информационные технологии в приёмной кампании ОмГУ // Открытое и дистанционное образование. 2002. № 1. С. 85–90.

DEVELOPMENT IN OMSU OF THE NEW INFORMATION SYSTEM FOR ADMISSION TO UNIVERSITY

T.A. Pogromskaya
Ph.D.(Eng.), Head of the Department of Web-Based Technologies of Information Management of OmSU, e-mail: pta@omsu.ru

Dostoevsky Omsk State University

Abstract. The article describes the experience of the development and transition to a new information system for reception of entrants to higher education in Dostoevsky Omsk State University. The system is based on the same program, introduced in four Russian universities, it uses old data structure and logic, but has a new web interface.

Keywords: information system, development, reception of entrants, applicant, web-interface.

Дата поступления в редакцию: 04.08.2016

Математические структуры и моделирование
2016. № 4(40). С. 122–132

УДК 004.9

РАЗРАБОТКА ЭЛЕКТРОННОЙ ОБРАЗОВАТЕЛЬНОЙ СРЕДЫ ВУЗА

С.В. Белим
профессор, д.ф.-м.н., e-mail: sbelim@mail.ru
И.Б. Ларионов
доцент, к.т.н., e-mail: me@g0gi.ch
Ю.С. Ракицкий
доцент, к.т.н., e-mail: yrakitsky@gmail.com

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В статье представлен один из возможных подходов формирования электронной образовательной среды. Рассмотрены вопросы предоставления образовательных ресурсов, а также оперативной информации об учебном процессе. Отдельное внимание уделено вопросам разграничения доступа и аутентификации при запросе на доступ к ресурсам.

Ключевые слова: электронная образовательная среда, образование, учебный процесс, учебные материалы, библиотека, электронные образовательные ресурсы.

Введение

Необходимость создания электронной образовательной среды вуза обусловлена современными тенденциями образования, которые нашли отражение в требованиях ст. 16 Федерального закона №273-ФЗ «Об образовании в Российской Федерации» и Федеральных государственных образовательных стандартах высшего образования.

Разработка и внедрение электронной образовательной среды вуза позволяет решать следующие задачи:

- предоставление единого авторизованного доступа к собственным информационным ресурсам вуза (изданиям вуза, методическим и справочным материалам вуза) для обучающихся и работников вуза с любого устройства, подключённого к сети Интернет;
- предоставление единого авторизованного доступа к электронным библиотечным системам и электронным подписным изданиям, с которыми заключён договор вуза;
- предоставление единого авторизованного доступа к текущей информации об учебном процессе (график учебного процесса, расписание и т.д.) для обучающихся и работников вуза с любого устройства, подключённого к сети Интернет.

Внедрение электронной образовательной среды вуза предоставляет обучающимся и работникам вуза ряд новых возможностей, таких как:

- удалённый доступ к библиотечным и информационным ресурсам университета, который приводит к повышению качества образования и интенсификации научной работы;
- удалённый доступ к текущей информации об организации учебного процесса;
- оперативное информирование обучающихся и работников об изменении в учебном процессе;
- оперативное предоставление учебных материалов обучающимся со стороны преподавателей.

Так, в работе [1] рассматривается применение электронной образовательной среды для совершенствования образовательного процесса в соответствии с миссией, приоритетами, стратегией, системой аккредитационных и инновационных показателей вуза. В работе [2] отмечается, что при создании информационно-образовательной среды необходимо учитывать ряд принципов: приоритетное внимание к мотивационному обеспечению процесса обучения и самообучения; опора на процессы саморазвития и индивидуализация обучения; постепенное расширение сферы самостоятельности обучающихся и уменьшение доли педагогического руководства ими; обеспечение принятия обучающимися некоей роли в учебном процессе; обучение рациональным способам учебной деятельности и самостоятельному приобретению знаний. В работе [3] информационно-образовательная среда характеризуется образовательными ресурсами. Под ними понимают различного вида содержательную учебную информацию (дидактическая, методическая, справочная, нормативная, организационная и др.), необходимую для эффективного управления педагогическим процессом с гарантированным качеством подготовки специалиста в учебное и внеучебное время. В работе [4] предлагается использование информационно-образовательной среды как основы внедрения электронных образовательных изданий и ресурсов в образовательный процесс университета. В работе [5] отмечается, что образовательная среда должна строиться как многокомпонентная система, содержащая в себе компоненты учебной, внеучебной, научно-исследовательской деятельности, измерения, контроля и оценки результатов обучения. Основными требованиями к компонентам, входящим в состав среды, являются наличие чёткой методики их использования в учебном процессе, взаимосвязи с телекоммуникационными ресурсами. Информационные ресурсы должны отвечать стандартным требованиям, предъявляемым к образовательному процессу. Формирование образовательной среды создаёт дополнительные условия для анализа показателей образовательного процесса, позволяет получить целостное представление о состоянии системы образования, о качественных и количественных изменениях в ней. В работе [6] указано, что информационно-образовательная среда современного учебного заведения — это не только компьютеры, сеть, коммутационные устройства, офисная техника, программное обеспечение, «облачные» сервисы, но и люди (преподаватели, учителя, студенты, ученики), которые работают в этой среде. Поэтому от квалификации и мотивации людей во многом зависит

эффективность использования информационной среды. В работе [7] в перечень задач, которые необходимо решить в процессе разработки и поддержки информационно-образовательной среды, авторы включают представление технических, информационно-технологических, методических, ресурсных средств и создание условий их использования для повышения эффективности процесса обучения и достижения новых образовательных результатов. В работе [8] отмечается, что компьютерные технологии призваны стать не дополнительным компонентом в обучении, а неотъемлемой частью целостного образовательного процесса, значительно повышающей его эффективность. Информационные и коммуникационные технологии с каждым днём всё больше проникают в различные сферы образовательной деятельности. В большинстве случаев использование средств информатизации оказывает положительное влияние на интенсификацию работы преподавателей вузов, а также на качество обучения студентов. В работе [9] указано, что важным этапом информатизации учебных заведений стало внедрение в вузах локальных вычислительных сетей и создание общих информационных ресурсов. Появились новые понятия: информационная система вуза, система электронного документооборота, электронная библиотека. Произошли изменения и в учебном процессе.

Главной проблемой внедрения электронных образовательных сред является интеграция с уже существующими информационными системами вуза, обеспечивающими непрерывность учебного процесса. Описываемая в данной статье электронная образовательная среда была реализована в ОмГУ им. Ф.М. Достоевского и интегрирована с различными информационными системами, действующими в вузе.

1. Состав и основные функции электронной образовательной среды вуза

Электронная образовательная среда вуза (ЭОСВ) включает в себя сервер аутентификации и модули, обеспечивающие доступ к информационным ресурсам.

Сервер аутентификации решает следующие задачи:

- обеспечение авторизации пользователей при запросе доступа к информационным ресурсам через различные модули информационной системы;
- поддержка и периодическая актуализация базы данных пользователей (обучающихся и работников вуза). База данных содержит три типа пользователей «Обучающиеся», «Преподаватели», «Иные работники».

Для обеспечения доступа к информационным ресурсам ЭОСВ содержит следующие модули:

- электронная библиотека вуза;
- модуль удалённого доступа к электронным ресурсам библиотеки;
- информационный модуль об учебном процессе.

Состав ЭОСВ допускает расширение списка модулей, обеспечивающих доступ к информационным ресурсам.

Электронная библиотека вуза обеспечивает авторизованный доступ к электронным изданиям вуза с любого устройства, подключённого к сети Интернет.

Модуль удалённого доступа к электронным ресурсам библиотеки решает следующие задачи:

- обеспечивает авторизованный доступ к электронным библиотечным системам, с которыми заключён договор вуза, с любого устройства, подключённого к сети Интернет;
- обеспечивает авторизованный доступ к электронным версиям периодических изданий, с которыми заключён договор вуза, с любого устройства, подключённого к сети Интернет.

Информационный модуль об учебном процессе решает следующие задачи:

- обеспечивает авторизованный доступ к данным о расписании занятий, консультаций и экзаменов с любого устройства, подключённого к сети Интернет;
- обеспечивает авторизованный доступ к данным о графике учебного процесса;
- обеспечивает авторизованный доступ к объявлениям и сообщениям кафедры, деканата, учебного отдела и иных подразделений вуза, актуальных для данного обучающегося.

Личный кабинет обучающегося обеспечивает авторизованный доступ к данным о текущей и итоговой аттестации для каждого обучающегося, а также информации о его научных, образовательных, спортивных и иных достижениях.

2. Сервер аутентификации

Сервер аутентификации предназначен для авторизации пользователей при запросе доступа к информационным ресурсам через различные модули информационной системы. Для обеспечения доступности ЭОСВ необходима поддержка и периодическая актуализация базы данных пользователей, для чего требуется взаимодействие с информационными системами отдела кадров работников и отдела кадров студентов. Наиболее простой способ состоит в периодической автоматической выгрузке информации из баз данных отделов кадров и загрузке в базу данных сервера аутентификации.

Сервер аутентификации включает в себя базу данных обучающихся и работников вуза, модуль регистрации пользователей, модуль идентификации пользователей, модуль восстановления пароля.

База данных содержит отдельную запись для каждого пользователя. Актуализация базы данных производится не реже одного раза в неделю. Каждому пользователю системы сопоставляется запись, содержащая:

1. Уникальный идентификатор (ID, число, 4 байта). Для пользователей группы «Обучающийся» в качестве идентификатора выступает номер зачётной книжки. Для пользователей групп «Преподаватели», «Иные работники» идентификаторы формируются системой на основе генератора псевдослу-

чайной последовательности и передаются пользователям через корпоративную почту или деканат факультета на твёрдом носителе.
2. Фамилия (FN, строка).
3. Имя (SN, строка).
4. Отчество (TN, строка).
5. Логин (email, строка). В качестве логина используется адрес электронной почты, определяемый самим пользователем.
6. Хэш-значение пароля (H, число, 256 бит). В качестве алгоритма хэширования используется ГОСТ Р 34.11-2012.
7. Тип пользователя. (T, число). Возможны три типа пользователей: «Обучающиеся» (T=0), «Преподаватели» (T=1), «Иные работники» (T=2).
8. Факультет (D, строка).
9. Статус (S, число). Если пользователь зарегистрирован S=1, если не зарегистрирован S=0.

Регистрация пользователя осуществляется с любого устройства, подключённого к сети Интернет, с использованием любого web-браузера. Регистрация нового пользователя осуществляется на отдельной странице, ссылка на которую присутствует на главной странице ЭОСВ. Возможность регистрации в системе предоставляется обучающимся и работникам вуза, не прошедшим регистрацию ранее. Процедура регистрации пользователя состоит из следующих шагов:

1. Пользователю предлагается заполнить поля «Фамилия», «Имя», «Отчество», «Идентификатор», «Адрес электронной почты» и выбрать факультет из выпадающего списка. Все поля являются обязательными для заполнения. Перед отправкой данных система проверяет правильность заполнения полей. При неверном или неполном заполнении полей система выдаёт предупреждение и предлагает повторное заполнение полей, в которых обнаружены ошибки.
2. Данные, отправляются для проверки на сервер аутентификации. Сервер аутентификации осуществляет проверку наличия в базе данных пользователя с представленными данными. Если соответствующая запись отсутствует, то пользователю выдаётся сообщение о невозможности регистрации в системе. Если запись присутствует, но имеет статус зарегистрированного пользователя, то пользователю выдаётся сообщение о том, что регистрация невозможна, так как данный пользователь уже зарегистрирован в системе.
3. Если запись в базе данных присутствует и имеет статус незарегистрированного пользователя, то на адрес электронной почты, который ввёл пользователь, высылается сообщение со ссылкой. Пользователю предлагается перейти по данной ссылке и дважды ввести пароль его учётной записи. Система проверяет совпадение двух введённых паролей, после чего хэш-значение пароля пересылается серверу.
4. При совпадении псевдослучайной строки, введённой пользователем, со значением, высланным ранее сервером, производятся необходимые изменения записи в базе данных. Устанавливается статус «пользователь зарегистрирован».

При запросе пользователя на восстановление пароля выполняются следующие шаги:

1. Пользователю предлагается ввести фамилию, имя, отчество, адрес электронной почты.
2. Сервер проверяет наличие соответствующей записи в базе данных со статусом «зарегистрирован».
3. Если запись в базе данных отсутствует, то пользователю выдаётся соответствующее сообщение.
4. Если запись в базе данных присутствует, то пользователю на адрес электронной почты отправляется псевдослучайная строка.
5. Пользователю предлагается ввести строку, отправленную ему на адрес электронной почты, и дважды – новый пароль.
6. В базе данных обновляется поле, содержащее хэш-значение пароля.

Сервер обрабатывает запросы на аутентификацию, поступающие от модулей информационной системы. Протокол авторизации пользователей и взаимной аутентификации сервисов использует алгоритм электронной цифровой подписи ГОСТ Р 34.10–2012, который, в свою очередь, использует алгоритм получения дайджеста сообщения ГОСТ Р 34.11–2012. В протоколе используется уникальный идентификатор сессии. Идентификатор имеет длину 1024 символа и представляет собой 512-битное число в шестнадцатеричном представлении. Модули информационной системы, при необходимости аутентификации пользователя, направляют запрос серверу аутентификации, содержащий логин пользователя и хэш-значение пароля, подписанные электронной цифровой подписью, связанной с параметрами сессии и временной меткой. При успешной аутентификации пользователя сервер возвращает модулю токен на доступ к системе. При отказе в аутентификации модулю передаётся сообщение «Отказ».

При необходимости авторизованного доступа пользователя к информационным ресурсам, контролируемым каким-либо из модулей, модуль запрашивает логин пользователя и пароль, которые отправляет серверу аутентификации с помощью криптографического протокола, основанного на цифровой подписи. При получении от модуля запроса на аутентификацию сервер проверяет наличие пользователя в базе данных, а также правильность имени пользователя и пароля. В случае успешной аутентификации пользователя сервер отправляет модулю токен на доступ к ресурсам. В случае отказа в аутентификации модулю посылается соответствующее сообщение.

3. Электронная библиотека вуза

Модуль «Электронная библиотека вуза» предназначен для предоставления доступа к каталогу электронных изданий вуза неавторизованным пользователям, а также для предоставления доступа к электронным изданиям вуза авторизованным пользователям. Для обеспечения доступности ресурсов библиотеки вуза необходима взаимосвязь с информационной системой библиотеки вуза, в которой располагается каталог электронных изданий.

Модуль «Электронная библиотека вуза» осуществляет взаимосвязь ЭОСВ с

электронным каталогом библиотеки вуза. Взаимодействие пользователя с электронным каталогом библиотеки вуза осуществляется с любого устройства, подключённого к сети Интернет, с использованием любого web-браузера. Процедура взаимодействия пользователя с электронной библиотекой вуза зависит от статуса пользователя (зарегистрированный пользователь, не зарегистрированный пользователь). Не зарегистрированный пользователь имеет возможность просматривать каталог электронных изданий вуза, но не имеет возможности загружать файлы. Для пользователей, зарегистрированных в ЭОСВ, процедура взаимодействия с электронным каталогом библиотеки вуза состоит из следующих шагов:

1. Пользователю отображается список доступных для загрузки изданий в виде ссылок на файлы в заданном формате. Предпочтительней использовать формат pdf, поскольку файлы в данном формате отображаются большинством современных браузеров, а также присутствует возможность установить приложение для просмотра файлов в данном формате.
2. После выбора пользователем одной из ссылок, ему предлагается пройти авторизацию. Запрос на авторизацию отображается в отдельном окне. Введённые пользователем логин и пароль отправляются на проверку серверу авторизации.
3. В случае успешной авторизации окно для ввода логина и пароля закрывается, после чего начинается процедура загрузки запрошенного файла.
4. В случае неудачной авторизации пользователю отображается соответствующее сообщение, и файл не загружается.

4. Удалённый доступ к электронным ресурсам библиотеки

Модуль «удалённый доступ к электронным ресурсам библиотеки» предназначен для предоставления удалённого доступа работникам вуза и обучающимся в вузе к внешним информационным ресурсам, с которыми заключён договор вуза. Как правило, внешние информационные ресурсы, с которыми заключён договор вуза, доступны из внутренней сети вуза. Целью модуля «удалённый доступ к электронным ресурсам библиотеки» является предоставление удалённого доступа к внешним ресурсам зарегистрированным в ЭОСВ пользователям с любого устройства, подключённого к сети Интернет, с использованием любого web-браузера.

Для пользователей, зарегистрированных в ЭОСВ, удалённый доступ к электронным ресурсам библиотеки вуза состоит из следующих шагов:

1. На сайте библиотеки вуза пользователю отображается ссылка на страницу с внешними информационными ресурсами.
2. Веб-страница, содержащая список внешних информационных ресурсов, также содержит поля для ввода логина и пароля пользователя, позволяющие аутентифицироваться пользователю. Введённые пользователем логин и пароль отправляются на проверку серверу авторизации.

3. После авторизации пользователю отображается список ссылок на внешние информационные ресурсы, с которыми в данный момент вуз заключил договор.
4. Переход пользователя по ссылкам осуществляется с правами доступа, предоставленными вузу по договору для данного ресурса. Для выполнения указанных условий в модуль от сервера авторизации передаётся токен на доступ.
5. При отказе аутентификации пользователя на странице отображается сообщение «Неверное имя пользователя или пароль», модулю от сервера авторизации вместо токена передаётся сообщение «Отказ». При этом пользователь имеет возможность повторно ввести логин и пароль.

5. Учебный процесс

Модуль «Учебный процесс» предназначен для получения работниками вуза и обучающимися в вузе информации об организации учебного процесса в вузе. Традиционно информация о расписании, график учебного процесса, объявления преподавателей, кафедр и деканата, а также иная информация доступна на информационных стендах в бумажном виде. Целью модуля «Учебный процесс» является предоставление доступа зарегистрированным и авторизованным в ЭОСВ пользователям к информации доски объявлений деканата и кафедр с любого устройства, подключённого к сети Интернет.

Для пользователей, зарегистрированных в ЭОСВ, возможен доступ к графику учебного процесса, доске объявлений и расписанию занятий. Для просмотра графика учебного процесса и расписания занятий пользователю необходимо выбрать факультет, образовательную программу и номер курса. Документы доступны в формате pdf.

Просмотр объявлений доступен после выбора факультета. На странице каждого факультета размещены актуальные объявления. Для каждого объявления указываются поля, упрощающие пользователю поиск нужного объявления: студенческая группа, которой предназначено объявление; дата размещения объявления; должностное лицо, разместившее объявление; текст объявления.

Каждое посещение пользователем страницы с объявлением фиксируется в журнале просмотра объявлений. Каждая запись о просмотре объявления содержит следующие поля: фамилия, имя и отчество пользователя, посетившего страницу объявления; дата и время посещения; факультет, на страницу объявлений которого был осуществлён вход.

Функция размещения информации в системе доступна только выделенным пользователям, наделённым соответствующими полномочиями. В современных системах управления и планирования наиболее распространённой является ролевая политика безопасности, поэтому полномочия пользователям предоставляются в рамках ролевой политики безопасности [10–13]. В системе предусмотрены следующие роли:

1. «Редактор расписания». Роль «Редактор расписания» предоставляет пользователю полномочия размещать расписание занятий, консультаций и эк-

заменов посредством загрузки файлов в формате pdf.

2. «Редактор графика». Роль «Редактор графика» предоставляет пользователю полномочия размещать график учебного процесса.
3. «Редактор объявлений». Роль «Редактор объявлений» предоставляет пользователю полномочия размещать объявления и просматривать журнал посещения объявлений.

Размещение расписаний осуществляется один раз в семестр и не требует дополнительных модулей проверки актуальности. В то время как доска объявлений должна поддерживаться в актуальном состоянии постоянно с учётом короткого «времени жизни» отдельных объявлений. Кроме этого, необходимо отслеживать процесс чтения объявлений обучающимися. Для решения этих задач в системе действует диспетчер объявлений и журнал просмотра объявлений. Каждому размещаемому объявлению присваивается внутренний уникальный идентификатор (число). Объявления хранятся в системе и отображаются пользователям до даты актуальности и в течение недели после указанной даты. После этого уничтожается само объявление и все записи журнала, связанные с ним. В журнале просмотра объявлений фиксируются все факты чтения объявления пользователями системы: имя, дата, время. Пользователь, разместивший объявление, имеет возможность получать информацию о круге пользователей, ознакомившихся с ним.

Каждый случай использования полномочий по размещению расписаний, графиков учебного процесса и объявлений заносится в журнал. Каждая запись журнала содержит следующие поля: пользователь, разместивший информацию; дата и время размещения информации; вид информации (расписание, график учебного процесса, объявление); атрибуты информации (факультет, форма обучения, курс, группа). Полномочиями на просмотр журнала размещения информации обладает администратор системы.

Заключение

В работе рассмотрены проблемы создания и интеграции ЭОСВ. Требования создания ЭОСВ сформулированы в ст. 16 Федерального закона №273-ФЗ «Об образовании в Российской Федерации» и Федеральных государственных образовательных стандартах высшего образования. При этом, при создании ЭОСВ также необходимо соблюдать требования Федерального закона №152-ФЗ «О персональных данных», разделяя информацию о обучающихся в вузе и о сотрудниках вуза. Это достигается с помощью реализации сервера аутентификации и использования алгоритма хэширования ГОСТ Р 34.11-2012. Также пользователям ЭОСВ предоставляется возможность использования ресурсов библиотеки вуза и удаленных ресурсов библиотеки вуза с помощью реализации соответствующих модулей. Кроме того, ЭОСВ предоставляет возможность оперативно получать информацию об учебном процессе как сотрудникам вуза, так и обучающимся в вузе. Таким образом, ЭОСВ повышает качество образовательного процесса в вузе и создает дополнительные условия для развития сотрудников вуза и обучающихся в вузе.

Литература

1. Вроробьев Г.А. Электронная образовательная среда инновационного университета // Высшее образование в России. 2013. № 8–9. С. 59–64.
2. Носкова Т.Н., Тумалева Е.А., Шилова О.Н. Информационные технологии в образовании и высокотехнологичная образовательная среда // Universum: Вестник Герценовского университета. 2012. № 2. С. 83–87.
3. Скибицкий Э.Г. Информационно-образовательная среда вуза: цель или средство в обеспечении качества образования. URL: `http://www.edit.muh.ru/content/mag/trudy/06_2009/06.pdf` (дата обращения: 16.05.2016).
4. Насейкина Л.Ф. Применение электронных образовательных изданий и ресурсов как компонентов развития информационно-образовательной среды универститета // Вестник ОГУ. 2011. № 2. С. 248–253.
5. Свиряева М.А., Молоткова Н.В., Анкудимова И.А. Организация информационно-образовательной среды вуза на основе технологий дистанционного обучения // Вопросы современной науки и практики. 2010. № 4–6(29). С. 180–184.
6. Голубев О.Б., Никифоров О.Ю. Развитие информационно-образовательной среды современного вуза // Инновации в образовании. ИнВестРегион. 2014. № 1. С. 57–61.
7. Гагарина Д.А., Хеннер Е.К. Структурв высокоразвитой информационно-образовательной среды инновационного университета // Университетское управление: практика и анализ. 2009. № 3. С. 69–73.
8. Еремина И.И. Формирование информационно-коммуникационной компетенции субъектов образовательного процесса в условиях информационной образовательной среды вуза // Педагогика. Психология. 2012. № 1. С. 162–169.
9. Прохоренков П.А. Этапы формирования электронной информационно-образовательной среды вуза // Международный журнал экспериментального образования. 2016. № 2. С. 291–294.
10. Белим С.В., Богаченко Н.Ф., Ракицкий Ю.С. Совмещение ролевой и мандатной политик безопасности // Проблемы обработки и защиты информации. Книга 1. Модели политик безопасности компьютерных систем. 2010. С. 117–132.
11. Белим С.В., Богаченко Н.Ф., Ракицкий Ю.С. Теоретико-графовый подход к проблеме совмещения ролевой и мандатной политик безопасности // Проблемы информационной безопасности. Компьютерные системы. 2010. № 2. С. 9–17.
12. Богаченко Н.Ф., Белим С.В., Белим С.Ю. Использование метода анализа иерархий для построения ролевой политики безопасности // Проблемы информационной безопасности. Компьютерные системы. 2013. № 3. С. 7–17.
13. Белим С.В., Богаченко Н.Ф. Применение метода анализа иерархий для оценки рисков утечки полномочий в системах с ролевым разграничением доступа // Информационно-управляющие системы. 2013. № 6(67). С. 67–72.

THE DEVELOPMENT OF ELECTRONIC EDUCATIONAL ENVIRONMENT OF HIGH SCHOOL

S.V. Belim
Dr.Sc.(Phys.-Math.), Professor, e-mail: sbelim@mail.ru
I.B. Larionov
Associate Professor, Ph.D.(Eng.), e-mail: me@g0gi.ch
Y.S. Rakitskiy
Associate Professor, Ph.D.(Eng.), e-mail: yrakitsky@gmail.com

Dostoevsky Omsk State University

Abstract. The article presents one of the possible approaches of formation of electronic educational environment. The problems of educational resources, as well as operational information about the learning process, are considered. Special attention is paid to issues of access control and authentication when requesting access to resources.

Keywords: e-learning environment, education, educational process, teaching materials, library, electronic educational resources.

Дата поступления в редакцию: 26.09.2016

Математические структуры и моделирование
2016. № 4(40). С. 133–142

УДК 004.422

АРХИТЕКТУРА МОБИЛЬНОГО КЛИЕНТА ПОД IOS ДЛЯ ДОСТУПА К ВЕБ-СЛОВАРЮ НАРОДНОЙ РЕЧИ СРЕДНЕГО ПРИИРТЫШЬЯ

И.А. Балезин
студент, e-mail: iabalezin@gmail.com
Д.Н. Лавров
к.т.н., доцент, e-mail: dmitry.lavrov72@gmail.com
М.А. Харламова
к.фил.н., доцент, e-mail: khr-spb@mail.ru

Омский государственный университет им. Ф.М. Достоевского

Аннотация. Представлена архитектура и разработано мобильное клиент-серверное приложение для доступа с iPhone к словарю констант народной речи Среднего Прииртышья с необходимым набором возможностей.

Ключевые слова: архитектура приложения, iOS, мобильное приложение, словарь народной речи.

1. Введение

Изучение национального мировидения и отражения его в различных формах народной речи — одна из важнейших задач и неотъемлемая часть лингвистической науки, поэтому создание региональных словарей, манифестирующих традиционную народную культуру, представляется безусловно актуальным.

Объектом описания нового словаря, над которым трудятся диалектологи и программисты ОмГУ им. Ф.М. Достоевского, является речь диалектоносителей полиэтнического региона Среднего Прииртышья (без дифференциации по этническим, социальным, образовательным и др. признакам). Предметом же становится описание культурно значимых констант (ключевых слов), объективирующих мировидение сельского жителя Омского Прииртышья и сохраняющих устойчивые элементы смысла на протяжении истории. Словарь имеет полевую структуру, что позволяет объёмно и полно реконструировать тот или иной фрагмент диалектной картины мира, см. подр. [1].

Электронный словарь выполняет, прежде всего, важную практическую функцию — позволяет быстро найти слово и его контекстуальное употребление. Эвристическая функция электронного словаря обеспечивает решение задач исследовательского характера, касающихся как семантики отдельного слова-репрезентанта константы, его синтагматических и парадигматических связей, так и способности слова к участию в устойчивых сочетаниях и в прецедентных высказываниях. История функционирования константы, представленная в словаре, репрезентирует своеобразие концептуализации мира носителями традици-

онной крестьянской культуры. Электронный формат словаря значительно расширяет пользовательскую аудиторию: доступен как специалистам-филологам, культурологам, историкам, этнографам, так и всем ценителям народного слова.

Мобильный клиент, представленный в данной работе, разработан для электронного веб-словаря народной речи Среднего Прииртышья. С рабочей версией веб-словаря можно ознакомиться по адресу: `http://dict.univer.omsk.su`. Представляемая статья является продолжением работ [2, 3].

2. Требования к клиенту «MiddleIrtyshDictionary»

Мобильное приложение «MiddleIrtyshDictionary» представляет собой клиент информационной платформы MediaWiki, основной задачей которого является предоставление пользователям возможности читать статьи с экрана мобильного устройства, работающего на ОС iOS. В частности речь идёт о смартфоне iPhone на iOS v 9.0.2 и выше. Приложение представляет карманный словарь констант народной речи Среднего Прииртышья. Функционал приложения дополняется возможностью работы с учётной записью. Статьи должны быть отсортированы по алфавиту, а также должен быть поиск, история открытых статей, закладки в приложении и возможность делиться статьями в соц. сетях или по e-mail.

Представлена диаграмма прецедентов (см. рисунок 1), которая отражает варианты использования разработанного клиент-приложения. Пользователь — основной исполнитель, который является пользователем приложения «MiddleIrtyshDictionary».

Перечислим основные варианты использования разработанного приложения:

1. Войти в приложение: пользователь может войти в приложение. Вход в приложение возможен посредством авторизации, если пользователь уже имеет аккаунт, или регистрации, если это новый пользователь.
2. Авторизация: пользователь вводит в форму авторизации свой логин и пароль. Если данные указаны верно, то выполняется вход в приложение.
3. Регистрация: пользователь указывает в форме регистрации желаемый логин и пароль, подтверждает ещё раз пароль и e-mail. Если пользователя с таким логином не существует, то выполняется регистрация нового пользователя и вход в приложение, в противном случае пользователю необходимо выбрать другой логин.
4. Выйти из приложения: пользователь может выйти из приложения.
5. Открыть список статей: после входа в приложение пользователю становится доступен экран со списком статей.
6. Открыть статью: на экране со списком опубликованных словарных статей пользователь выбирает любую по заголовку, открывает статью для чтения.
7. Добавить статью в закладки: если статья заинтересовала, у пользователя есть возможность добавить статью в избранные.
8. Открыть меню приложения: доступ к остальным функциям приложения пользователю предоставляется по действию свайпа вправо с левой сторо-

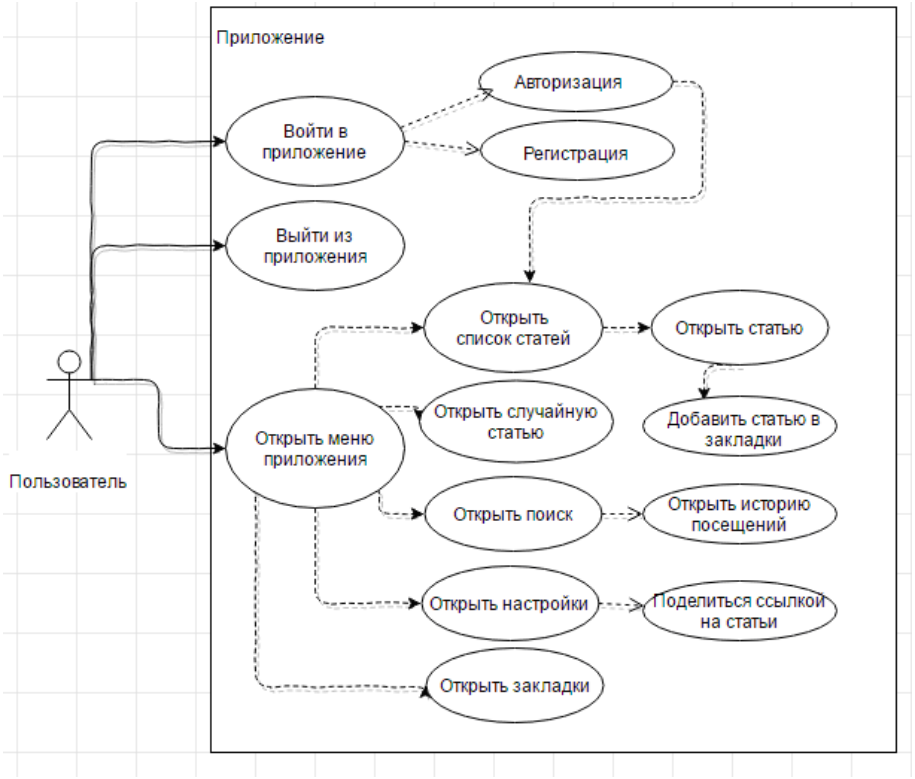


Рис. 1. Диаграмма прецедентов

ны экрана, или по нажатию на иконку меню с вызовом бокового меню.

9. Открыть случайную статью: одной из функций MediaWiki является открытие любой случайной статьи.
10. Открыть поиск: пользователь может начать поиск среди опубликованных статей по заголовку.
11. Открыть историю посещений: не возвращаясь к экрану со списком опубликованных статей, пользователь может вернуться к прочитанной статье через историю.
12. Открыть настройки: экран приложения предоставляет возможность пользователю увидеть имя учётной записи, из-под которой совершается сессия, и завершить сессию.
13. Поделиться ссылкой на статьи: пользователь может отправить сообщение кому-либо со ссылкой на сайт словаря из приложения.
14. Открыть закладки: экран приложения, содержащий перечень ссылок на статьи, которые пользователь добавляет сам при чтении статей.

3. Архитектура приложения «MiddleIrtyshDictionary»

Архитектура рассматриваемого приложения использует паттерн проектирования (см. рисунок 2) модель-представление-контроллер (MVC, model-viewcontroller) [4]. Основная идея данного паттерна заключается в том, чтобы

Рис. 2. Модель-контроллер-представление

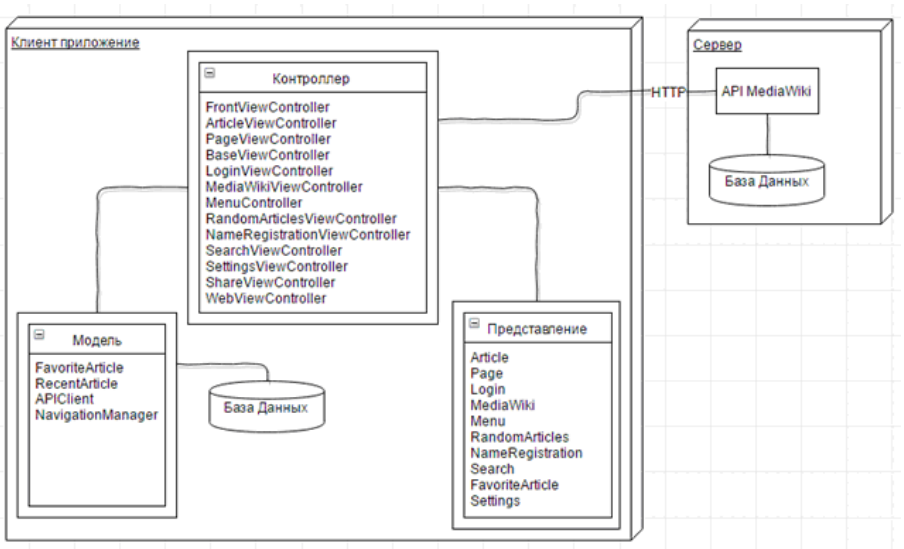


Рис. 3. Архитектура приложения

разделить модель данных приложения, пользовательский интерфейс и взаимодействие с пользователем на три отдельных компонента таким образом, чтобы модификация одного из компонентов оказывала минимальное воздействие на остальные.

Представлена архитектура разработанного приложения (см. рисунок 3).

Компонент «представление» отвечает за отображение информации. В разработанном приложении данный компонент включает следующие экраны:

- Article — экран со списком статей;
- Page — экран текущей статьи;
- Login — экран авторизации;
- MediaWiki — экран полной версии сайта;
- Menu — экран бокового меню;
- RandomArticles — экран случайной статьи;
- NameRegistration — экран регистрации;
- Search — экран поиска;
- FavoriteArticle — экран с закладками.
- Settings — экран с настройками.

Компонент «контроллер» обеспечивает связь между пользователем и системой: контролирует ввод данных пользователем и использует модель и представ-

ление для реализации необходимой реакции. В рассматриваемом приложении модуль контроллеров состоит из:

- FrontViewController — контроллер, отвечающий за анимацию бокового меню и кнопки.
- ArticleViewController — контроллер, который отвечает за взаимодействие с представлением Article.
- PageViewController — контроллер, который отвечает за взаимодействие с представлением Page.
- BaseViewController — контроллер, отвечающий за навигацию всех контроллеров.
- LoginViewController — контроллер, который отвечает за взаимодействие с представлением Login.
- MediaWikiViewController — контроллер, который отвечает за взаимодействие с представлением MediaWiki.
- MenuController — контроллер, который отвечает за взаимодействие с представлением Menu.
- RandomArticlesViewController — контроллер, который отвечает за взаимодействие с представлением RandomArticles.
- NameRegistrationViewController — контроллер, который отвечает за взаимодействие с представлением NameRegistration.
- SearchViewController — контроллер, который отвечает за взаимодействие с представлением Search.
- SettingsViewController — контроллер, который отвечает за взаимодействие с представлением Settings.
- ShareViewController — контроллер, который отвечает за взаимодействие с представлением FavoriteArticle.
- WebViewController — представляет собой область, которая может отображать HTML-контент.

Компонент «модель» предоставляет данные и методы работы с этими данными, реагирует на запросы, изменяя своё состояние.

В разработанном приложении модель состоит из:

- APIClient — модель, в которой устанавливается content-type (application/json, text/html), и методы запросов (GET, POST), а также URL, на который они будут отправляться.
- NavigationManager — модель, содержащая кнопки для навигации в отдельном баре.
- NSError+APIClient — модель, которая обрабатывает ошибки с сервера.
- NSError + NSResponsData — модель, которая показывает ответ с сервера.
- UIViewController+AlertHelpersViewController — модель, которая показывает уведомление об ошибке. А также классов, взаимодействующих с базой данных (core Data):
- FavoriteArticle — модель, работающая с добавлением элементов в избранные.
- RecentArticle — модель, работающая с добавлением элементов в историю посещений.

Таблица 1. Таблица FavoriteArticle

№	Атрибут	Семантика	Тип
1	Date	Дата добавления	DATE
2	PageID	id страницы	STRING
3	Title	Заголовок статьи	STRING
4	Url	Ссылка на статью	STRING

Для хранения закладок и истории посещений используется база данных, состоящая из двух таблиц: FavoriteArticle и RecentArticle. Описание таблицы, в которой хранятся избранные статьи, приведено в табл. 1. В аналогичной таблице хранится история посещений — Таблица RecentArticle.

4. Графический интерфейс приложения

Компонент «представление» состоит из набора экранов, созданных с помощью стандартного средства платформы OS X — Interface Builder. Interface Builder предоставляет палитры или коллекции объектов пользовательского интерфейса для разработчиков. Эти объекты содержат такие элементы, как текстовые поля, таблицы данных, слайдеры и всплывающие меню. Палитры Interface Builder являются полностью расширяемыми, то есть любой разработчик может создавать новые объекты и добавлять их к уже существующим. Для создания интерфейса разработчик просто перетаскивает элементы интерфейса с палитры на окно или меню. Конкретные объекты, которые получают сообщения, указываются в коде приложения. Таким образом, все инициализации происходят до выполнения, что ведёт к повышению производительности и делает процесс разработки более упорядоченным. Разработанное приложение включает в себя 10 экранов. При входе в приложение игрок может авторизоваться (представление LoginViewController) или создать аккаунт, т.е. зарегистрироваться (представление NameRegistrationViewController), см. рисунок 4.

После того как пользователь авторизуется, осуществляется переход на экран со списком статей, представление ArticleViewController (см. рисунок 5), на скриншоте экрана также присутствует библиотека MJNIndexView. Находясь в представлении ArticleViewController, пользователь может открыть статью (нажатием по заголовку) для чтения. После выбора элемента из списка (статьи) открывается представление PageViewController (см. рисунок 5), в представлении PageViewController имеется возможность добавить текущую статью в избранные, список добавленных статей в закладки отображается в представлении FavoriteArticle (см. рисунок 6).

Представление FavoriteArticle содержит метод segmentedControlChanged типа IBAction, что в результате является кнопкой-переключателем с «Закладок» на «Последние посещения», последнее — это список с историей открытия статей. Модели FavoriteArticle (закладки) и RecentArticle (послед-

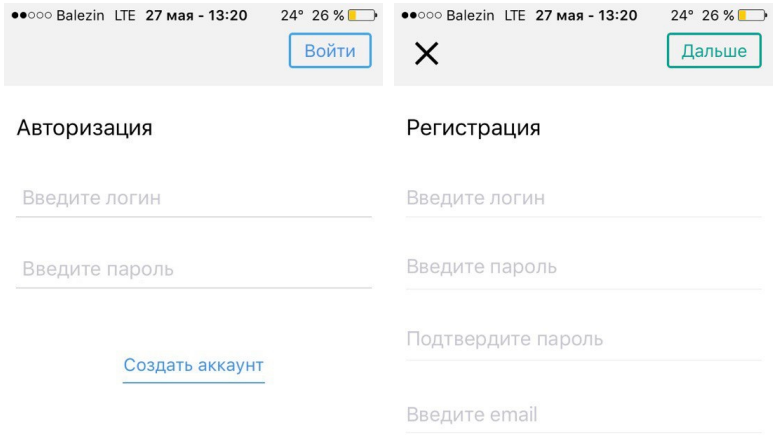


Рис. 4. Экраны LoginViewController и NameRegistrationViewController

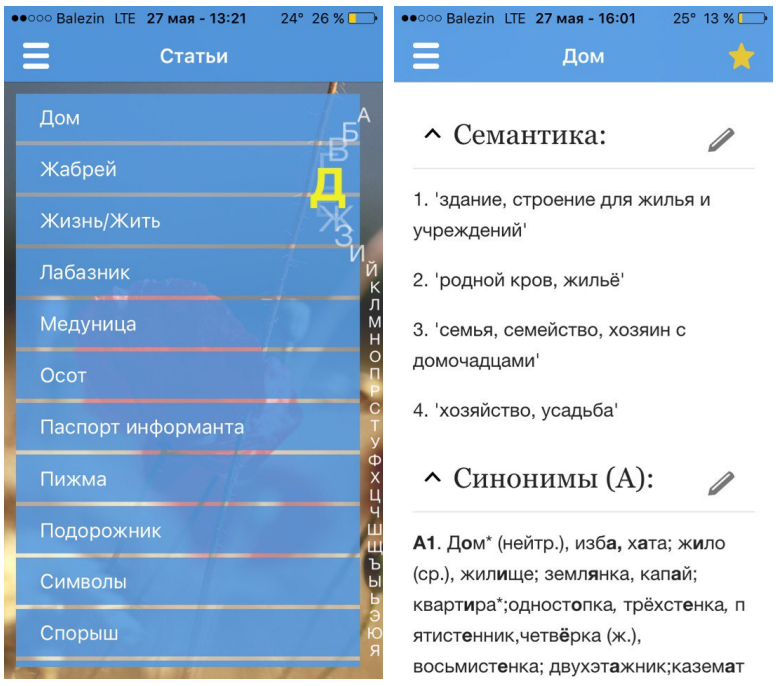


Рис. 5. Экран ArticleViewController и Экран PageViewController

ние посещения) доступны в представлении FavoriteArticle через контролер ShareViewController (см. рисунок 6). Открыть меню приложения для доступа к остальным функциям приложения пользователю предлагается свайпом вправо с левой стороны экрана или по нажатию на иконку меню. Меню является представлением MenuController, боковое меню разработано на основе библиотеки SWRevealViewController.

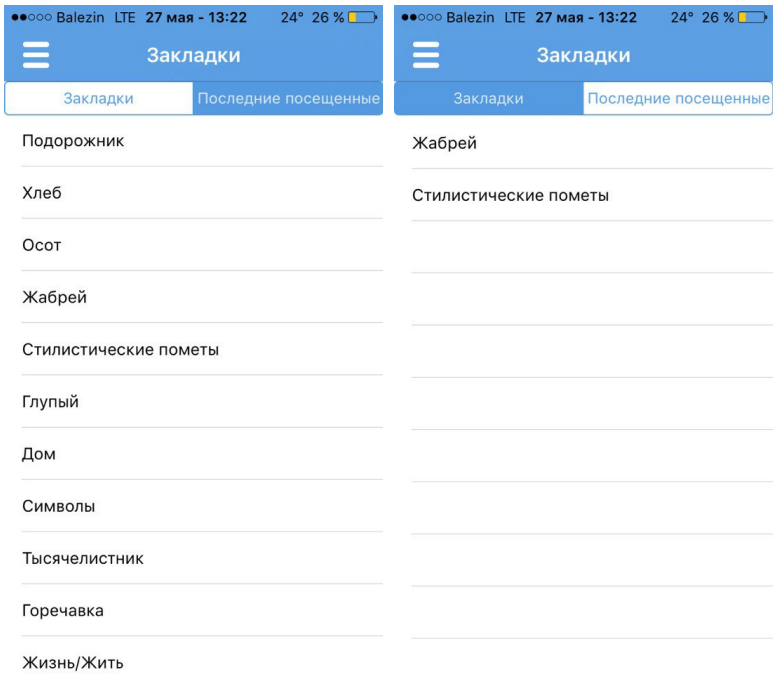


Рис. 6. Экран FavoriteArticle и Экран FavoriteArticle

Представление MenuController, меню приложения, содержит перечень ссылок в виде кнопок на функции приложения, представляет навигацию внутри приложения. Из представления MenuController уже доступны и другие экраны, например, случайная статья, поиск, полная версия, настройки. Экран со случайной статьёй, представление RandomArticlesViewController, предоставляет возможность пользователю открыть любую случайную статью из опубликованных, реализуется функция информационной системы MediaWiki. Экран поиска, представление SearchViewController, представляет возможность пользователю, начать поиск среди опубликованных статей по заголовку. Экран полной версии, представление MediaWikiViewController, открывает веб-сайт, сайт-словарь констант народной речи Среднего Прииртышья. Экран настройки, представление SettingsViewController (см. рисунок 7), предоставляет возможность пользователю увидеть имя учётной записи, из-под которой совершается сессия, и завершить сессию. Поделиться ссылкой на статьи: пользователь может отправить сообщение кому-либо со ссылкой на сайт словаря из приложения.

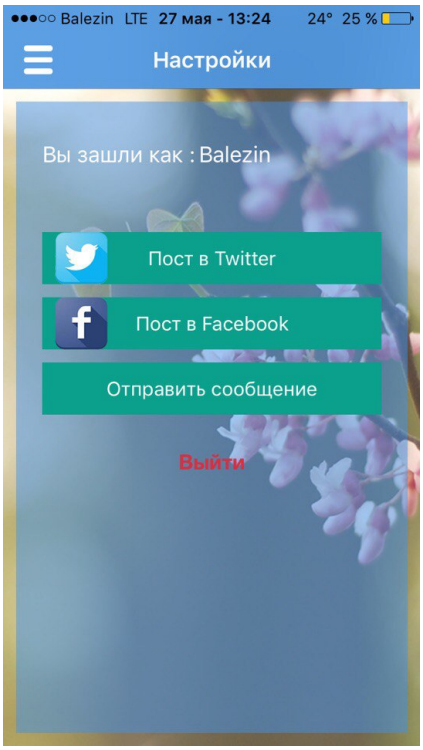


Рис. 7. Экран SettingsViewController

Заключение

Основной результат работы — это разработанная и реализованная архитектура мобильного клиента под iOS с возможностью чтения статей с сайта на небольшом экране. Проработан и реализован эргономичный, интуитивно понятный графический интерфейс мобильного приложения.

Литература

1. Харламова М.А. Константы народной речемысли и их лексикографическая интерпретация. Омск : Изд-во Ом.гос.ун-та, 2014. 290 с.
2. Лавров Д.Н., Харламова М.А. Словарь констант народной речи: выбор платформы представления // Вестник Омского университета. Омск : Ом. гос. ун-т, 2015. № 1(75). С. 213–215.
3. Балезин И.А., Лавров Д.Н. Разработка формы для заполнения словарных статей и мобильного клиента под iOS для словаря констант народной речи Среднего Прииртышья // XL региональная студенческая научно-практическая конференция «Молодёжь третьего тысячелетия». 8–30 апреля 2016 г. 8 с. В печати.
4. Piper I. Learn XCode Tools for Mac OS X and iPhone Development. USA: Appress, 2009. 450 p.

THE ARCHITECTURE OF THE MOBILE CLIENT FOR IOS TO ACCESS WEB-DICTIONARY OF FOLK SPEECH OF THE MIDDLE IRTYSH

I.A. Balezin
Student, e-mail: iabalezin@gmail.com
D.N. Lavrov
Ph.D. (Eng.), Associate Professor, e-mail: dmitry.lavrov72@gmail.com
M.A. Harlamova
Ph.D. (Philological), Associate Professor, e-mail: khr-spb@mail.ru

Dostoevsky Omsk State University

Abstract. The architecture is presented and the mobile client-server application is developed for access from iPhone to the dictionary of the constants of folk speech of the Middle Irtysh area with the necessary set of options.

Keywords: application architecture, of iOS, mobile app, dictionary of folk speech.

Дата поступления в редакцию: 31.10.2016

Математические структуры и моделирование
2016. № 4(40). С. 143–150

УДК 004.056

ДОКАЗАТЕЛЬСТВО С НУЛЕВЫМ РАЗГЛАШЕНИЕМ КАК МЕТОД АУТЕНТИФИКАЦИИ В ВЕБ-ПРИЛОЖЕНИЯХ

И.Д. Сиганов
аспирант, e-mail: ilya.siganov@gmail.com

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В статье рассмотрен альтернативный протокол аутентификации для веб-приложений, основанный на доказательстве с нулевым разглашением. Такой подход позволяет избежать многих уязвимостей, которым подвержены системы с обычной парольной аутентификацией, таких как перехват процесса входа в систему, подсматривание секрета, взлома и анализа базы данных сервера для поиска паролей. Реализованная система состоит из веб-сервера и андроид-приложения. Процесс аутентификации происходит с помощью сканирования QR-кода со страницы веб-сайта приложением–менеджером аутентификации. Достоинством системы можно считать её удобство, высокий уровень защищённости, возможность безбоязненного использования на публичных и подозрительных компьютерах.

Ключевые слова: компьютерная безопасность, аутентификация, доказательство с нулевым разглашением, веб-приложения, андроид.

Введение

Как известно, выделяют три сущности аутентификации — что субъект знает, что имеет и кто субъект есть. Дешевле всего для реализации взять первую сущность, основанную на знании какого-то секрета, например, пароля. Такие системы проще всего проектировать, на первый взгляд, и они не требуют от пользователя дополнительных устройств. Скажем так, это наиболее привычный метод аутентификации где бы то ни было — вводить пароль в систему. При всей своей простоте и преимуществах у этого подхода есть серьёзный недостаток. Безопасность полностью зависит от самого пользователя, который придумал секрет, предполагая, что сервера хранят пароли достаточно безопасно. Но люди — очень плохой источник энтропии, кроме того, им нужно помнить пароль, поэтому они придумывают очень простые секреты. К тому же этот подход плохо масштабируется на пользователях. Им приходится помнить столько паролей, сколько сервисов они использует. Всевозможные менеджеры и генераторы паролей призваны решить эти проблемы. Но не стоит забывать, что они реализуются как специальные приложения или расширения для браузеров, которые надо ещё установить на компьютер. Это не приемлемо для исполь-

зования на общедоступных компьютерах, например, в школах, университетах, интернет-кафе, на рабочих местах. Итак, проблемы:

1. Люди — плохой источник энтропии, поэтому создают очень простые секреты, которые легко взломать.
2. Люди не могут помнить сотни паролей, от этого страдает сложность паролей ещё сильнее. Кроме того, это просто не удобно.
3. Возможен перехват пароля во время его ввода.
4. Если пароль перехвачен, то его нужно сразу же менять. Но не всегда можно вовремя обнаружить перехват.
5. База данных с хэшированными паролями всегда поддаётся криптоанализу. Постоянно в СМИ появляются материалы об очередной утечке баз паролей.

Исходя из проблем, можно сформулировать критерии идеального протокола аутентификации:

1. Стойкость не зависит от выбора пользователя, т.е. система сама создаст необходимые ключи; эталон, хранимый на сервере, не должен быть уязвим к прямому взлому для вычисления самого ключа аутентификации.
2. Стойкость к пассивному прослушиванию, подглядыванию.
3. Стойкость к атаке воспроизведением.
4. Достаточно высокая скорость работы, чтобы не сильно нагружать вычислительные мощности клиента, что позволит охватить распространённые устройства.
5. Не создавать высокие нагрузки на сервер.

В общем случае аутентификация — это доказательство: Виктор должен удостовериться, что Пегги действительно знает какой-то секрет. Ограничения, которые были выписаны, можно представить в виде желания Пегги не раскрывать свой секрет, но каким-то образом доказать Виктору, что она знает секрет, причём Виктор действительно поверит Пегги. К тому же Пегги хочет, чтобы Виктор не смог потом воспользоваться доказательством и выдать себя за неё. Протоколы, реализующие данную схему, называются протоколами нулевого разглашения. Классическим объяснением этого протокола служит пример с пещерой, в которой есть дверь с паролем. Формальное определение даётся через вероятностные машины Тьюринга, которые разделяют общую ленту [3, с. 81]. Если описать простыми словами, то все сводится к тому, что у Пегги есть неограниченные вычислительные мощности, а у Виктора имеется лишь обычная машина. Виктор отправляет вопросы Пегги, на которые в силах ответить только она, но проверить правильность ответа может каждый. Предполагается также, что вероятность обмана крайне мала. К тому же есть дополнительное ограничение, называемое как раз нулевым разглашением — т.е. Виктор не сможет воспользоваться доказательствами Пегги, чтобы выдать себя за неё.

Как видно, протоколы нулевого разглашения устойчивы к пассивному прослушиванию, повторному воспроизведению и криптоанализу базы данных эталонов. К сожалению, большинство таких протоколов интерактивные и требуют

прохождения большого числа испытаний для построения доказательства. Связь между числом испытаний n и вероятностью аутентичности выражается формулой:

$$p(n) = 1 - \frac{1}{2^n}.$$

Для хоть сколько-то приемлемой величины нужно пройти хотя бы 10 испытаний. В случае растущего потока людей будет генерироваться большая нагрузки на сервер аутентификации. Чтобы этого избежать, придумывают протоколы нулевого разглашения, где необходимо не так много итераций. Одним из таких примеров может служить схема Шнорра. Это модификация схемы Эль-Гамаля и Фиата-Шамира, в основе использующая проблему дискретного логарифмирования в конечном поле. Есть, правда, некоторые вопросы в том, действительно ли этот протокол нулевого разглашения, так как нет доказательства за и против.

1. Описание системы

Концептуально разработанная система описывается взаимодействием между сервером аутентификации и клиентским приложением — менеджером аутентификации. Сервер аутентификации может работать как отдельный сервис или как микросервис, встроенный, например, в другое веб-приложение. В данной работе реализован второй подход. В любом случае сервер аутентификации имеет БД пользовательских данных, необходимых для процесса аутентификации, и дополнительные хранилища, необходимые для функционирования самого протокола. Клиентское приложение написано под ОС андроид; оно хранит список всех учётных записей пользователя в своей внутренней шифрованной БД вместе с приватными ключами и осуществляет процесс аутентификации. Клиент может работать с любыми веб-приложениями, использующими разработанный сервер аутентификации, то есть нет привязки к определённому адресу сервера и реализации веб-приложения, главное чтобы API взаимодействия с сервером аутентификации оставался прежним. Благодаря этому приложение может использоваться для любых сайтов, использующих этот сервер аутентификации. Особенностью данной реализации является то, что система разрабатывалась без жёсткой привязки к определённому протоколу аутентификации, поэтому поддерживает многие схемы. В качестве примера были реализованы протоколы s/key и схема Шнорра, но подробное описание архитектуры приложения не входит в рамки данной статьи.

2. Описание протокола на основе схемы Шнорра

На примере алгоритма Шнорра покажем детально, как передаются данные между сервером аутентификации, браузером и приложением-менеджером аутентификации. Схема Шнорра — это протокол аутентификации, основанный на сложности дискретного логарифмирования в конечном поле. В нём принимают участие две стороны — Пегги, которая хочет подтвердить свою личность, и Виктор, проверяющий её доказательство. Пегги имеет у себя два ключа, один

из которых закрытый и должен храниться в безопасном месте, а второй общедоступный. Таким образом, Виктор проверяет знание Пегги закрытого ключа по её открытому ключу. Протокол проходит в три шага и считается обладающим свойством нулевого разглашения.

Рис. 1. Диаграмма входа в систему

На рисунке 1 изображена диаграмма, иллюстрирующая процесс входа в систему. На ней пользователь — это браузер клиента. Сервер — это веб-приложение, в котором клиент хочет авторизоваться. Менеджер ключей — это мобильное приложение клиента со всеми его существующими аккаунтами. Можно заметить любопытную особенность — фактически оказалось, что в протоколе участвует три субъекта. Обычно запрос аутентификации, сама аутентификация и все действия авторизованного пользователя происходят в рамках одной сессии. Ключевой особенностью этой системы является разделение процесса аутентификации между двумя субъектами. Благодаря этому привычный менеджер паролей удалось вынести за пределы устройства, с которого будет производиться вход. Итак, запрос на вход в систему отправляется через браузер, сессию которого мы хотим аутентифицировать. Но подтверждение аутентичности вместе с идентификацией проходит через другой канал, в котором приложение на смартфоне доказывает серверу, что та новая открытая сессия должна быть ассоциирована с таким-то пользователем. Информация о сессии попадает в приложение через QR-код, который необходимо сканировать со страницы браузера. После этого пользователь выбирает аккаунт для входа именно на этот сайт.

3. Описание токена

Как было описано выше, через QR-код передаётся некоторая информация об открытой сессии. Предоставим полную структуру токена, который создаёт сервер. Поле «type» используется для того, чтобы приложение автоматически предлагало пользователю либо зарегистрировать новый аккаунт, либо выбрать существующий для входа. Может принимать значение «LOGIN» или «SIGNUP». Поле «domainName» представляет собой имя сервиса. Предполагается, что оно уникально и не меняется. В приложении поиск аккаунтов происходит по этому ключу. Поле «path» содержит в себе путь, по которому нужно обратиться на сервис для осуществления входа или регистрации. Так как в рамках своего приложения разработчики имеют право использовать любую адресацию, то было решено вынести адрес конечной точки в токен. Поле «token» содержит случайное 16-байтное число в шестнадцатеричном формате UUID-v4. Оно используется для поиска сессии в БД на сервере для ассоциации сессии браузера с авторизованным через приложение пользователем. Поле «expiresAt» — дата истечения жизни токена в формате числа — количества миллисекунд со времени UTC. В данной реализации время жизни любого токена 30 секунд. В течение этого времени необходимо осуществить вход или регистрацию в систему. После этого времени токен больше не принимается сервером и удаляется из БД. Поле «algorithm» используется для выбора нужного алгоритма для аутентификации и регистрации в приложении. Так как предполагается, что приложение работает со многими протоколами, поэтому имя протокола было вынесено в это поле. В дополнение к вышеописанным полям было добавлено ещё одно поле «requestInfo», содержащее информацию о том, откуда был вызван запрос на вход или регистрацию. Это было добавлено с целью защиты от фишинга и «обмана, выполненного мафией». Подробнее об этом написано далее, где будут рассмотрены возможные атаки на протокол. В этом поле содержится информация об IP адресе хоста и о клиентском приложении, с которого был выполнен запрос. После заполнения всех полей структуры данные конвертируются в формат JSON и отображаются в браузере как QR-код или как intent-link, специальная ссылка, при переходе по которой открывается приложение. Это удобно если вы хотите зайти на сайт с самого смартфона.

4. Сценарий использования приложения

Опишем процесс взаимодействия пользователя с системой. Для начала сценарий регистрации.

1. Пользователь заходит на сайт и нажимает на кнопку «зарегистрироваться». Сервер в этот момент создаёт специальный token и отдаёт его пользователю в двух форматах — как QR-код и как intent-link.
2. Пользователь с помощью приложения аутентификатора сканирует QR-код или переходит по intent-link.
3. Приложение разбирает токен и приглашает пользователя ввести новый login.

4. Приложение генерирует секрет, соответствующий указанному в токене протоколу, и сохраняет его в локальной шифрованной базе данных. Из секрета создаётся публичная часть ключа и отправляется на сервер.
5. Сервер обрабатывает полученный публичный ключ, сохраняет в БД и автоматически аутентифицирует сессию, в которой была запрошена регистрация.
6. Дополнительные данные, такие как e-mail, имя и прочее, пользователь будет вводить уже на самом сайте, так как эти данные уже не участвуют в протоколе.

Сценарий входа в систему:

1. Пользователь запрашивает token для входа в систему.
2. Пользователь сканирует код. Далее приложение ищет в БД аккаунт для данного сервиса и протокола по доменному имени. Если пользователь имеет несколько аккаунтов, то можно будет выбрать нужный. Кроме того, пользователь будет видеть мета-информацию о системе, сессию браузера которой мы хотим аутентифицировать. В метаинформации могут содержаться сведения о координатах, IP-адресе, операционной системе, т.е. все, что удалось получить из http запроса на вход.
3. Начинается фаза доказательства аутентичности. Если она проходит успешно, то страница в браузере автоматически обновится, а сессия будет авторизована.

Процесс восстановления утерянного секрета не входит в рамки исследования, так как не является частью протокола. Реализовать восстановление кода можно, как обычно, через e-mail или sms.

5. Анализ угроз

Рассмотрим основные угрозы в сети интернет.

1. Атака «человек по середине». Будем рассматривать только пассивную атаку, т.е. прослушивание канала связи. В этом случае злоумышленник может находиться между сервером и клиентом, между устройством ввода и программой или просто подсмотреть через плечо ввод данных. Если от перехвата через сам канал связи нас может защитить HTTPS, то от кейлоггеров построить защиту сложнее, так как выдвигаются требования к конечному устройству. Как описывалось в самом начале статьи, это нельзя гарантировать на публичных компьютерах.
2. Обычно атаку MITM провести сложно, поэтому злоумышленники прибегают к социальной инженерии и фишингу. Суть этих атак заключается в заманивании жертвы на ресурс, очень похожий на оригинальный, где жертва, ничего не подозревая, оставит свои секретные сведения. Этот тип атак направлен прямо на человека, поэтому противостоять им очень сложно.
3. Не стоит недооценивать и вероятность утечки базы с паролями. Обычно для защиты используют хэширование паролей с модификатором, такой подход сильно усложняет криптоанализ, но тем не менее, не делает его

невозможным. И в дополнение, криптоанализ тем легче, чем проще придуманные секреты, так как они с большой вероятностью уже будут в заготовленных таблицах у злоумышленников.

Исходя из построения системы и выдвинутых ранее критериев, система защищена от перечисленных угроз. Но как оказалось, существует одна интересная атака, которой подвержены все протоколы с нулевым разглашением — это «атака, выполненная мафией» или «проблема гроссмейстера».

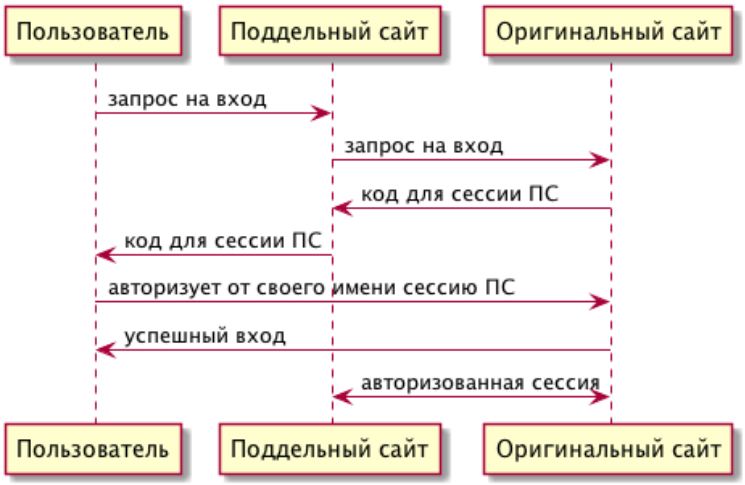


Рис. 2. Атака, выполненная мафией

Суть её заключается в следующем (см. рис. 2):

1. Пользователь заходит на фишинговый сайт. Запрашивает токен для входа.
2. Сайт злоумышленника проксирует этот запрос на оригинальный сайт, тем самым получая токен для своей сессии, и отдаёт его пользователю неизменным.
3. Жертва аутентифицирует сессию злоумышленника, ничего не заметив.

Чтобы избежать этой уязвимости, каждый токен должен содержать в себе информацию о запросе токена. Из всех возможных данных подделать нельзя только IP-сокет источника запроса, из которого можно определить приблизительные координаты клиента. Тогда перед выбором аккаунта в приложении будет показана эта метаинформация, и пользователь сможет понять, подставляют ли его или нет. Кроме того, в приложение можно встроить чёрный список фишинговых сайтов и постоянно его пополнять. Токен в данном случае придётся подписывать сертификатом сервера, чтобы избежать модификации метаинформации.

6. Заключение

Предложенная в статье модель аутентификации с использованием протоколов нулевого разглашения призвана избавиться от недостатков существующих схем и решить проблемы уязвимости к атаке, направленной на самого человека.

Исходя из самих критериев, озвученных в начале, следует, что такие проблемы как пассивное подслушивание трафика, не дадут злоумышленнику ничего. Пользователь не вводит свои секретные данные на каких-то устройствах ввода в незнакомых местах, он всегда использует только свой защищённый смартфон, поэтому вирусы и кейлогеры не представляют опасности. Пользователь не обязан придумывать сложные пароли и помнить их все, таким образом решается проблема слабых паролей и забывания сложных. Так как серверы не должны хранить пароли, то у злоумышленников больше нет вектора атаки на базы данных с пользовательскими секретами. Вместо этого серверы хранят публичные ключи пользователей, и предполагается, что по ним создать закрытую часть ключа очень сложно, по крайней мере намного сложнее, чем перебрать пароли по словарю. Вместе со всем этим следует и удобство в использовании.

Исходные коды сервера аутентификации доступны по адресу: https://github.com/blan4/ZeroKnowledgeProofServer, а приложение клиент — https://github.com/blan4/ZeroKnowledgeProofClient. Тестовый сервер развернут по адресу: https://zkpauth.herokuapp.com/.

Литература

1. Шнайер Б. Прикладная криптография. М. : Триумф, 2002. 541 с.
2. Шнайер Б. Протоколы, алгоритмы, исходные тексты на языке Си. М. : Триумф, 2002. 816 с.
3. Варновский Н.П. Криптография и теория сложности // Математическое просвещение. 1998. Сер. 3. Вып. 2. С. 71–86.
4. Доказательства с нулевым разглашением. URL: `http://citforum.ru/security/cryptography/yaschenko/16.html` (дата обращения: 15.02.2016)

ZERO KNOWLEDGE PROOF AUTHENTICATION ON WEB APPLICATIONS

I.D. Siganov
Postgraduate Student, e-mail: ilya.siganov@gmail.com

Dostoevsky Omsk State University

Abstract. In the article we overview how to implement zero knowledge proof authentication protocol in the web. The proposed system consists of two parts: a server side and Android application. Despite the classical password-based approach, users have to install special application and use it as authentication manager. We use QR-codes to send necessary data from the server to the application, so login experience is just scanning this code. Finally, because of zero knowledge proof features the system is resistant to interception, secret-cracking and fishing, so supposed even to use on insecure public devices.

Keywords: zero knowledge Proof, web authentication, passwordless.

Дата поступления в редакцию: 23.06.2016

Математические структуры и моделирование
2016. № 4(40). С. 151–163

УДК 004.056.53

О СВЯЗИ МЕЖДУ ОБЪЕКТНО-ОРИЕНТИРОВАННОЙ ДИСКРЕЦИОННОЙ И СУБЪЕКТНО-ОБЪЕКТНОЙ МАНДАТНОЙ МОДЕЛЯМИ БЕЗОПАСНОСТИ

С.В. Усов
к.т.н., e-mail: raintower@mail.ru

Омский государственный университет им. Ф.М. Достоевского

Аннотация. В статье рассмотрены объектно-ориентированная модель Харрисона-Руззо-Ульмана и субъектно-объектные модели с мандатным разграничением доступа. Показано, что модель Белла-Лападулы и классические мандатные модели могут быть реализованы с помощью объектно-ориентированной модели HRU.

Ключевые слова: дискреционные модели безопасности, мандатные модели безопасности, разграничение доступа, HRU, модель Белла-Лападулы.

Введение

Как дискреционные, так и мандатные политики безопасности известны еще с 70-х годов прошлого столетия, и традиционно базируются на субъектно-объектной парадигме компьютерной системы. Однако в связи с возрастающей актуальностью объектно-ориентированного подхода к построению компьютерных систем, возникает необходимость в пересмотре классических политик безопасности. Так, например, в [1] была предложена объектно-ориентированная модель разграничения доступа, базирующаяся на модели HRU (Харрисона-Руззо-Ульмана) [2], однако, обладающая более широкими возможностями, в частности, в рамках охвата компьютерных систем, которые можно описать с помощью этой модели.

Получают более широкое применение и мандатные политики безопасности [3], [4]. В частности, мандатная модель используется как семейством операционных систем Windows (начиная с Vista применяется для контроля целостности), так и семейством операционных систем Linux (доступна в качестве расширений). Цель данной работы — установить взаимосвязь между объектно-ориентированными дискреционными системами безопасности и мандатными системами безопасности, эксплуатирующими субъектно-объектный подход.

Прежде всего необходимо ответить на вопрос, позволяет ли инструментарий объектно-ориентированной модели HRU реализовать мандатную политику безопасности, и если позволяет, то с какими ограничениями.

В работе [1] была предложена иерархическая модель OOHRU, устройство которой подразумевает, что объект o, находящийся на более низком уровне

иерархии, чем объект o', обладает меньшим набором прав (как в отношении доступа к другим объектам, так и в отношении ограничения доступа других объектов по отношению к себе) по сравнению с объектом o'. Такая структура в точности повторяет решётку ценностей мандатной политики безопасности, что позволяет сделать предположение о структурной близости данных моделей.

1. Объектно-ориентированная модель безопасности с дискреционным разграничением доступа (OOHRU)

Компьютерная система в OOHRU рассматривается в виде множества объектов $\mathbf{O}$, разбитых по множеству классов $\mathbf{K}$ (все объекты одного класса имеют одинаковый набор полей и методов), обладающих открытыми полями $f \in \mathbf{F}$ и скрытыми полями $p \in \mathbf{P}$, а также методами обработки полей $s \in \mathbf{S}$. Здесь $F = \bigcup_{k \in \mathbf{K}} k.\mathbf{F}$ — множество всевозможных открытых полей всех объектов и классов, $k.\mathbf{F}$ — множество открытых полей класса k (каждый объект класса k обладает тем же набором $k.\mathbf{F}$ открытых полей), аналогично определяются $\mathbf{P}$ и $\mathbf{S}$. Причём если поле $k.f$ наследуется классом k у класса k', то соответствующее поле класса k' мы будем для удобства обозначать именно $k'.f$, подчёркивая тем самым их взаимосвязь (таким образом, $f \in k.\mathbf{F}$ и $f \in k'.\mathbf{F}$). Пусть $\mathbf{O}^k \in \mathbf{O}$ — множество объектов класса $k \in \mathbf{K}$. В случае, если требуется уточнить класс объекта, поле f объекта $o^k \in \mathbf{O}^k$ будем обозначать $o^k.f$, поле f класса k — $k.f$. Для скрытых полей класса будем использовать аналогичные обозначения.

Для построения модели дискреционного разделения доступов для каждого объекта и для каждого класса вводится дополнительное скрытое поле M, содержащее локальную матрицу доступов, и методы работы с матрицей доступов. Модификация матриц доступа производится посредством выполнения команд системы безопасности, о которых будет сказано ниже.

Модель безопасности OOHRU называется иерархической (или моделью с иерархией), если на множестве объектов $\mathbf{O}$ задан частичный порядок-иерархия, и в любой момент работы системы для любых двух объектов $o, o' \in \mathbf{O}$ таких, что $o' \leqslant o$, для любого поля или метода $x \in \mathbf{X}$, общего для объектов o и o', и для любого поля или метода $x' \in \mathbf{X}$ объекта $o'' \in \mathbf{O}$ верно следующее: $o''.M[o, x'] \subset o''.M[o', x']$ и $o'.M[o'', x] \subset o.M[o'', x]$. Здесь и далее $\mathbf{X}$ — множество всевозможных полей и методов всех существующих в системе на данный момент времени объектов, «$\leqslant$» — отношение частичного порядка.

Состояние системы в модели HRU изменяется под действием команд, которые состоят из условной части и последовательности элементарных операторов [2], которая выполняется, только если истинна условная часть. Список элементарных операторов в OOHRU включает [1]:

1. $Create(o^k, k)$ — создаёт объект o^k класса $k \in \mathbf{K}$, если $o^k \in \mathbf{O}$.
2. $Destroy(o^k)$ — уничтожает объект $o^k \in \mathbf{O}$.
3. $Enter(r, o^k, o'^{k'}.f)$ — вносит право доступа r в $o'^{k'}.M[o^k, o'^{k'}.f]$, где o^k – объект класса k, $o'^{k'}$ — объект класса k'.
4. $Delete(r, o^k, o'^{k'}.f)$ — удаляет право доступа r из $o'^{k'}.M[o^k, o'^{k'}.f]$.
5. $Grant(r, o^k, o'^{k'}.s)$ — разрешает вызов объектом o^k метода $o'^{k'}.s$.

6. $Deprive(r, o^k, o'^{k'}.s)$ — запрещает вызов объектом o^k метода $o'^{k'}.s$.

Изменения, производимые операторами, отражаются в матрицах доступа объектов системы. Подробное описание модели OOHRU можно найти в [1].

2. Мандатные политики безопасности

Мандатные политики безопасности оперируют понятиями уровня секретности информации и уровня доверия к пользователю. На множестве уровней секретности (уровней доверия) задано отношение нестрогого порядка. Таким образом, получаем частично упорядоченное множество L, в отдельных случаях являющееся решёткой (например, если L линейно упорядочено). Такую решётку будем называть решёткой ценностей.

Типичным примером мандатной политики безопасности является общепринятая для секретного документооборота в большинстве стран модель MLS, основанная на решётке ценностей. На множестве объектов $\mathbf{O}$ системы определяется функция ценности C, сопоставляющая каждому объекту один из уровней решётки ценностей L. Поток информации от объекта o к объекту o' допускается, только если $C(o)$ не превосходит $C(o')$.

Другим примером может служить модель Белла-ЛаПадулы [3], [4], в которой, однако, мандатная политика безопасности совмещается с дискреционной. Опишем эту модель подробнее.

Система представляется совокупностью множества объектов $\mathbf{O}$ доступа, множества субъектов $\mathbf{S}$ доступа, множества видов доступа $A = \{read, write, append, execute\}$ и матрицы доступов M, аналогичной используемой в модели HRU. Кроме того, заданы решётка ценностей L (обычно это линейно-упорядоченное множество, содержащее четыре классических уровня секретности: Unclassified, Confidencial, Secret, TopSecret, перечислены в порядке возрастания секретности) и тройка отображений $f = (f_S, f_C, f_O)$. Множество всех таких троек f обозначим через $\mathbf{F}$. $f_S : \mathbf{S} \to L$ определяет максимальный уровень допуска субъекта, $f_C : \mathbf{S} \to L$ — текущий уровень допуска субъекта, а $f_O : \mathbf{O} \to L$ — уровень секретности объекта. Дополнительно может определяться иерархия H объектов системы (например, на основе отношения вложенности папок [4]), для любой пары «родитель-потомок» этой иерархии уровень секретности родителя не может превосходить уровень секретности потомка (но может совпадать с ним). Отображение $f_H : \mathbf{O} \to L$ сопоставляет каждому объекту системы его место в иерархии, причём все объекты o, находящиеся в иерархии на одной позиции $h = f_H(o)$, имеют один и тот же уровень секретности $f_O(o)$. Поэтому в дальнейшем под уровнем секретности объекта мы будем понимать именно его позицию h в иерархии.

Множество текущих доступов в системе можно записать как $B \subset \mathbf{S} \times \mathbf{O} \times A$. В каждый момент времени система находится в определённом состоянии, являющемся декартовым произведением $d = \mathbf{B} \times \mathbf{M} \times f \times H$, переход в другое состояние осуществляется посредством выполнения одной из системных команд из множества Γ, а также исполнения запросов на доступ из множества Q, также являющегося подмножеством $\mathbf{S} \times \mathbf{O} \times A$. В отличие от дискреционной

модели Харрисона-Руззо-Ульмана, вид команд из множества Γ не специфицирован, однако, приводится список их возможностей [3]:

1. Изменить положение объекта в иерархии H, изменить текущий уровень доверенности субъекта или уровень секретности объекта, то есть изменить функции f_H, f_C и f_O.

2. Добавить или удалить право доступа субъекту на объект, то есть изменить содержание матрицы доступов M.

3. Создать новый объект или удалить группу объектов, то есть изменить иерархию H.

В то же время запросы из множества Q служат для оперирования (создания или прекращения) текущими потоками между субъектами и объектами. Подобные запросы не специфицированы в HRU, поэтому мы не будем останавливаться на них подробно.

Безопасность системы определяется с помощью трех свойств: ss-свойства, *-свойства и ds-свойства.

Доступ $b = (s, o, r) \in B$ обладает ss-свойством относительно $f = (f_S, f_C, f_O)$, если

1. $r = read$ или $write$, и $f_O(o) \leqslant f_S(s)$.

2. $r = execute$ или $append$.

Доступ $b = (s, o, r) \in B$ обладает *-свойством относительно $f = (f_S, f_C, f_O)$, если

1. $r = read$ и $f_O(o) \leqslant f_S(s)$.

2. $r = append$ и $f_S(s) \leqslant f_O(o)$.

3. $r = write$ и $f_O(o) = f_S(s)$.

4. $r = execute$.

Доступ $b = (s, o, r) \in B$ обладает ds-свойством относительно $f = (f_S, f_C, f_O)$, если $r \in M[s, o]$.

Состояние системы обладает ss-свойством (*-свойством, ds-свойством) относительно $f = (f_S, f_C, f_O)$, если каждый доступ b в этом состоянии обладает тем же свойством.

Состояние системы называется безопасным, если оно обладает всеми тремя свойствами. Реализация системы называется безопасной, если каждое состояние её безопасно. Ограничения на команды в безопасной реализации системы описаны в так называемой Basic Security Theorem [3]. Перечислим эти ограничения для перехода системы из состояния (B, M, f, f_H) в состояние (B', M', f', f'_H):

1) любой доступ $(s, o, r) \in B' \setminus B$ обладает ss-свойством относительно f';

2) если $(s, o, r) \in B$ и не обладает ss-свойством относительно f', то $(s, o, r) \notin B'$;

3) любой доступ $(s, o, r) \in B' \setminus B$ обладает *-свойством относительно f';

4) если $(s, o, r) \in B$ и не обладает *-свойством относительно f', то $(s, o, r) \notin B'$;

5) для любого доступа $(s, o, r) \in B' \setminus B$ верно, что $r \in M[s, o]$;

6) если $(s, o, r) \in B$ и $r \notin M[s, o]$, то $(s, o, r) \notin B'$.

Basic Security Theorem утверждает, что система безопасна тогда и только тогда, когда начальное состояние системы безопасно, и для всех переходов системы в последующие состояния выполнены шесть вышеперечисленных условий.

Заметим, что ss-свойство следует из *-свойства, что вызвано историческими причинами [4]. В первоначальной работе Белла и ЛаПадулы *-свойство отсутствовало, и было введено позднее, чтобы избавить систему, защищённую по БЛП, от уязвимости к атакам вида «троянский конь».

Исходя из тех соображений, что право доступа $write$ является комбинированным, по сути совмещая в себе права как на чтение объекта, так и на запись в объект, в то время как $read$ подразумевает доступ только на чтение, $append$ — только на запись, а $execute$ вообще не подразумевает прямого доступа субъекта к данным, мы можем заменить четыре права доступа четырьмя комбинациями всего двух пар доступа, $read$ (только чтение) и $write$ (только запись). В этом случае *-свойство можно сформулировать заметно проще:

Доступ $b = (s, o, r) \in B$ обладает *-свойством относительно $f = (f_S, f_C, f_O)$, если

1. $r = read$ и $f_O(o) \leqslant f_S(s)$,
2. $r = write$ и $f_S(s) \leqslant f_O(o)$.

Что касается иерархии H, то будем следовать замечанию Белла, изложенному в [4], что естественно полагать, что субъект, имеющий доступ к подпапке, имеет доступ и к документам самой папки, а субъект, не имеющий доступа к папке, не сможет просматривать и вложенные в неё подпапки. Это означает, что если $read \in M[s, o]$ и $f_H(o) \geqslant f_H(o')$, то $read \in M[s, o']$, и наоборот, если $write \in M[s, o]$ и $f_H(o) \leqslant f_H(o')$, то $write \in M[s, o']$.

3. Связь между моделями Белла-ЛаПадулы (МБЛ) и OOHRU

Основной результат данной работы заключается в том, что субъектно-объектная модель Белла-ЛаПадулы (МБЛ) может быть реализована объектно-ориентированной моделью OOHRU.

Рассмотрим два принципиально отличающихся случая, для каждого сформулируем и докажем отдельную теорему.

Будем называть МБЛ ds-свободной, если матрица доступов не накладывает дополнительных (относительно *-свойства) ограничений на доступ субъектов к объектам.

МБЛ однозначно определяется множеством своих состояний (включая начальное) и способов перехода из одного состояния в другое, то есть набором команд из множеств Q и Γ, описанных выше. Поэтому в каждый момент времени t допустимо рассматривать такую модель Σ как набор $(D(t), \Gamma)$, где D — множество состояний системы, Γ — набор команд системы, для которых выполнены условия Basic Security Theorem. В худшем случае множество $\mathrm{B}(t)$ текущих доступов может содержать всевозможные доступы, не противоречащие безопасности состояния $d(t)$ системы в момент времени t. Будем считать,

что $\mathrm{B}(t)$ именно таково, что избавляет нас от рассмотрения запросов на доступ из множества Q.

С другой стороны, связанную с МБЛ объектно-ориентированную модель Σ' будем рассматривать как набор $(D'(t), \Gamma')$, где $D'(t) = (\mathbf{O'}(\mathbf{t}), M'(t), \mathbf{K}, F', R)$ — множество состояний системы, $\mathbf{O'}$ — множество объектов системы, M' — множество прав доступа, оформленное в виде матрицы доступов, $\mathbf{K}$ — множество классов системы (возможно, в виде иерархии; зависит от L и H), F' — отображение классов системы на решётку ценностей, R — множество видов доступа, наконец, Γ' — набор команд системы (зависит от L, f_S, H).

Будем говорить, что объектно-ориентированная модель Σ' реализует субъектно-объектную МБЛ Σ, если существует взаимно-однозначное отображение ϕ, определённое на каждом из элементов модели Σ, устанавливающее соответствие между состояниями и командами систем Σ и Σ', такое что любому переходу системы Σ из состояния d в состояние d', совершаемому в результате выполнения команды γ, соответствует переход системы Σ' из $\phi(d)$ в состояние $\phi(d')$ в результате выполнения команды $\phi(\gamma)$.

Теорема 1. *Для любой безопасной ds-свободной МБЛ существует реализующая её иерархическая модель OOHRU.*

Доказательство. Построим искомую систему OOHRU и одновременно — требуемое отображение ϕ.

Во-первых, в МБЛ будем рассматривать только два вида доступа, $read$ и $write$. Причина тому была приведена в конце предыдущего параграфа. Соответственно, в OOHRU сохранятся те же виды доступа.

Во-вторых, иерархию классов в OOHRU будем строить на основе решётки L ценностей и иерархии H уровней секретности объектов.

Введём множество служебных классов $K' = \{k^0, k^{read}, k^{write}, k^{system}\}$. Класс k^0 — корневой и не обладает никакими правами доступа, в то время как объекты других классов обладают полным множеством прав доступа к нему. Класс k^{read} обладает полным набором прав чтения, но полностью лишён прав записи. Класс k^{write}, напротив, обладает полным набором прав записи, но полностью лишён прав чтения. Эти три класса могут не содержать объектов, либо методы данных классов лишены функциональности, а поля — подлежащей защите информации. Последний класс k^{system} обладает полным набором прав как по записи, так и по чтению, и содержит доверенный объект — объект администратора системы.

Объекты в конструируемой объектно-ориентированной системе будут принадлежать к одному из четырёх типов: содержащие единственное поле $field$, в которое можно писать, содержащие единственное поле $field$, которое можно читать, содержащие единственный метод $read$ с правом доступа по чтению, содержащие единственный метод $write$ с правом доступа по записи. Данные ограничения необходимы только для доказательства, в реальной системе можно обойтись и без них.

Кроме того, каждому уровню секретности объектов модели Белла-ЛаПадулы сопоставим два класса в объектно-ориентированной модели, а каж-

дому уровню допуска субъектов модели Белла-ЛаПадулы — два семейства классов. В рамках первого семейства представлены классы, позволяющие реализовать возможности субъектов мандатной модели по чтению, в рамках второго — по записи. Данное сопоставление реализуем в виде отображения $F\text{'} : \mathbf{K} \setminus \mathbf{K'} \to L \cup H$, где соблюдаются следующие разбиения:

$$\mathbf{K} = \mathbf{K'} \cup \mathbf{KOR} \cup \mathbf{KOW} \cup \mathbf{KSR} \cup \mathbf{KSW},$$

так что $F\text{'} : \mathbf{KOR} \cup \mathbf{KOW} \to H$ и $F\text{'} : \mathbf{KSR} \cup \mathbf{KSW} \to L$.

$$\mathbf{KOR} = \bigcup_{h \in H} k_h^{OR},$$

$$\mathbf{KOW} = \bigcup_{h \in H} k_h^{OW},$$

$$\mathbf{KSR} = \bigcup_{l \in L} \mathbf{K}_l^{SR},$$

$$\mathbf{KSW} = \bigcup_{l \in L} \mathbf{K}_l^{SW},$$

так что $F\text{'}(k_h^{OR}) = F\text{'}(k_h^{OW}) = h$ и $F\text{'}(\mathbf{K}_l^{RS}) = F\text{'}(\mathbf{K}_l^{SW}) = l$,

$$\mathbf{K}_l^{SR} = \bigcup_{s \in \mathbf{S}} k_{(l,s)}^{read},$$

$$\mathbf{K}_l^{SW} = \bigcup_{s \in \mathbf{S}} k_{(l,s)}^{write},$$

где $\mathbf{S}$ — множество субъектов в модели Белла-ЛаПадулы.

Здесь $k_{(l,s)}^{read}$, например, обозначает класс, находящийся в иерархии на уровне l, в котором может быть создан объект $o_{(l,s)}^{read}$ с единственным методом, реализующий функционал субъекта МБЛ $s \in \mathbf{S}$ по чтению. И такой объект существует в модели OOHRU тогда и только тогда, когда соответствующий субъект в МБЛ находится на уровне безопасности l, т.е. $f_C(s) = l \leqslant f_S(s)$. Одновременно с ним существует и объект $o_{(l,s)}^{write}$ класса $k_{(l,s)}^{write}$, реализующий функционал субъекта $s \in \mathbf{S}$ по записи. Значение $f_S(s)$ для каждого субъекта при этом закладывается на уровне создания системы.

В свою очередь, k_h^{OR} обозначает класс, находящийся в иерархии на уровне h, в котором может быть создан объект $o_{(h,o)}^{OR}$ с единственным полем, содержащим информацию объекта МБЛ $o \in \mathbf{O}$, предназначенную для чтения. И такой объект существует в модели OOHRU тогда и только тогда, когда соответствующий объект в МБЛ находится на уровне иерархии h, т.е. $f_H(o) = h$. Одновременно с ним существует и объект $o_{(h,o)}^{OW}$ класса k_h^{OW}, находящийся на том же уровне h иерархии и содержащий идентичную информацию, однако, к этому объекту могут обращаться только объекты из семейства классов KSW для выполнения операции записи.

Таким образом, $o^{read}_{(l,s)}$ может получить право читать информацию из $o^{OR}_{(h,o)}$, но не из $o^{OW}_{(h,o)}$, в то время как соответствующий тому же субъекту МБЛ s объект $o^{write}_{(l,s)}$ может обладать правом писать в $o^{OW}_{(h,o)}$, но не в $o^{OR}_{(h,o)}$. По завершении операции записи в объект $o^{OW}_{h,o}$, его содержание копируется системным объектом в $o^{OR}_{h,o}$.

Отображение F' таково, что для любых двух уровней безопасности $l, m \in L$, $l \leqslant m$, верно:

$k^0 \leqslant k^{read}_{(l,s)} \leqslant k^{read}_{(m,s)} \leqslant k^{read}$, но

$k^0 \leqslant k^{write}_{(m,s)} \leqslant k^{write}_{(l,s)} \leqslant k^{write}$.

Кроме того, для любых элементов $h \leqslant g$ решётки H выполнено:

$k^0 \leqslant k^{OR}_h \leqslant k^{OR}_g$, но

$k^0 \leqslant k^{OW}_g \leqslant k^{OW}_h$.

Наконец, $k^0 \leqslant k^{system}$.

Матрицы доступа методов этих объектов также индуцируются матрицей доступа субъекта s. Для ds-свободных МБЛ это означает, что если $f_C(s) = l$, то

$read \in o^{OR}_{(h,o)}.M[o^{read}_{(l,s)}, o^{OR}_{(h,o)}.field] \Leftrightarrow f_O(o) \leqslant l$,

$write \in o^{OW}_{(h,o)}.M[o^{write}_{(l,s)}, o^{OW}_{(h,o)}.field] \Leftrightarrow f_O(o) \geqslant l$.

Для завершения доказательства нам достаточно представить реализацию команд МБЛ средствами модели OOHRU. С учётом того, что в ds-свободной МБЛ матрица доступов не претерпевает изменений, если уровни допуска субъектов и уровни секретности объектов не изменяются, достаточно представить только команды, соответствующие изменению функций f_H, f_C и f_O, а также созданию и удалению объектов. Условные части команд отсутствуют, поскольку для соблюдения *-свойства достаточно условий целостности из элементарных операторов модели OOHRU.

1. Присвоение субъекту s с текущим уровнем допуска l нового уровня допуска m.

Команда $ChangeSubjectSecurityLevel[l, m](o^{read}_{(l,s)} : k^{read}_{(l,s)}; o^{write}_{(l,s)} : k^{write}_{(l,s)}; o^{read}_{(m,s)} : k^{read}_{(m,s)}; o^{write}_{(m,s)} : k^{write}_{(m,s)})$

$Create(o^{read}_{(m,s)}, k^{read}_{(m,s)})$,

$Create(o^{write}_{(m,s)}, k^{write}_{(m,s)})$,

$Destroy(o^{read}_{(l,s)})$,

$Destroy(o^{write}_{(l,s)})$.

Данная команда присутствует в системе только для значений $l, m \leqslant f_S(s)$.

2. Присвоение объекту o с текущим уровнем секретности h нового уровня секретности g.

Команда $ChangeObjectSecurityLevel[h, g](o^{OR}_{(h,o)} : k^{OR}_{(h,o)}; o^{OR}_{(g,o)} : k^{OR}_{(g,o)}; o^{OW}_{(h,o)} : k^{OW}_{(h,o)}; o^{OW}_{(g,o)} : k^{OW}_{(g,o)}; o^{system} : k^{system})$

$Create(o^{OR}_{(g,o)}, k^{OR}_{(g,o)})$,

$Create(o^{OW}_{(g,o)}, k^{OW}_{(g,o)})$,

$Enter(read, o^{system}, o^{OR}_{(h,o)})$,

$Enter(write, o^{system}, o^{OR}_{(g,o)})$,

$Enter(write, o^{system}, o^{OW}_{(g,o)})$,
$Destroy(o^{OR}_{(h,o)})$,
$Destroy(o^{OW}_{(h,o)})$.

Здесь доступ системного объекта к перемещаемому объекту необходим для того, чтобы можно было скопировать информацию. В МБЛ подобные операции могут быть осуществлены только доверенным субъектом, роль которого в OOHRU отведена объекту системного класса.

3. Создание объекта o уровня секретности h.
Команда $CreateObject[h](o^{OR}_{(h,o)} : k^{OR}_{(h,o)}; o^{OW}_{(h,o)} : k^{OW}_{(h,o)}; o^{system} : k^{system})$
$Create(o^{OR}_{(h,o)}, k^{OR}_{(h,o)})$,
$Create(o^{OW}_{(h,o)}, k^{OW}_{(h,o)})$,
$Enter(write, o^{system}, o^{OR}_{(h,o)})$,
$Enter(write, o^{system}, o^{OW}_{(h,o)})$.

4. Удаление объекта o.
Команда $DestroyObject[h](o^{OR}_{(h,o)} : k^{OR}_{(h,o)}; o^{OW}_{(h,o)} : k^{OW}_{(h,o)})$
$Destroy(o^{OR}_{(h,o)})$,
$Destroy(o^{OW}_{(h,o)})$.

При этом условие Белла [4] «вместе с папкой удаляются и содержащиеся в ней подпапки» можно реализовать цепочкой команд, удаляющих объекты, являющиеся потомками удаляемого, либо представив папку с вложенными подпапками средствами одного объекта. При создании объекта возможности доступов этого объекта к другим объектам, а также других объектов к этому объекту, совпадают с соответствующими возможностями класса этого объекта, то есть согласуются исключительно с *-свойством и никогда не изменяются.

■

Замечание 1. Доказанное утверждение справедливо не только для ds-свободных МБЛ, а вообще для всех, наследующих права доступа, то есть обладающих следующим свойством:
$read \in M[s, o]$и$f_C(s) \leqslant f_C(s') \Rightarrow read \in M[s', o]$,
$read \in M[s, o]$и$f_H(o) \geqslant f_H(o') \Rightarrow read \in M[s, o']$,
$write \in M[s, o]$и$f_C(s) \geqslant f_C(s') \Rightarrow write \in M[s', o]$,
$write \in M[s, o]$и$f_H(o) \leqslant f_H(o') \Rightarrow write \in M[s, o']$.

Замечание 2. Если множество субъектов в системе не является известным заранее, утверждение теоремы остаётся верным с учётом небольшого изменения в доказательстве. А именно, вместо того, чтобы сопоставлять каждому субъекту решётку классов, будем пользоваться единой решёткой классов для всех субъектов. То есть вместо класса $k^{read}_{(l,s)}$, соответствовавшего субъекту s, используем класс k^{read}_{l}, единый для всех субъектов. Сопоставление каждому субъекту классов собственной решётки необходимо для ситуации, в которой один субъект может получить право на активизацию другого субъекта, однако, в МБЛ право execute применяется по отношению к объектам.

Замечание 3. При необходимости выделение индивидуальной иерархии классов для каждого объекта МБЛ также возможно. Например, если объект имеет достаточно сложную структуру, а не ограничивается единственным полем. В этом случае разбиение $\mathbf{KOR} \cup \mathbf{KOW}$ на классы полностью аналогично тому, что было применено в доказательстве теоремы при реализации средствами OOHRU субъектов МБЛ.

Замечание 4. Модель MLS с точки зрения МБЛ является безопасной ds-свободной, поскольку каждое её состояние обладает *-свойством (в его упрощённой формулировке), и лишена дополнительных ограничений на доступ в виде матрицы доступа. А значит, утверждение теоремы 1 справедливо и для MLS.

Замечание 5. Доказательство теоремы 1 возможно провести и другим способом, без выделения отдельных классов, отвечающих за чтение и запись. Например, можно использовать конструкцию, в которой отсутствие права чтения интерпретируется как наличие права записи (с определёнными оговорками).

Теорема 2. *Для любой безопасной МБЛ существует реализующая её модель OOHRU.*

Доказательство. Доказательство этой теоремы в целом повторяет доказательство теоремы 1 за рядом отличий. Ограничимся только перечислением этих отличий.

Во-первых, иначе происходят разбиения классов:

$$\mathbf{K} = \mathbf{K}' \cup \mathbf{KO} \cup \mathbf{KS},$$

так что $F\text{'} : \mathbf{KO} \to H$ и $F\text{'} : \mathbf{KS} \to L$.

$$\mathbf{KO} = \bigcup_{h \in H} \mathbf{K}_h^O,$$

$$\mathbf{KS} = \bigcup_{l \in L} \mathbf{K}_l^S,$$

так что $F\text{'}(\mathbf{K}_h^O) = h$ и $F\text{'}(\mathbf{K}_l^S) = l$,

$$\mathbf{K}_h^O = \bigcup_{o \in \mathbf{O}} k_{(h,o)},$$

$$\mathbf{K}_l^S = \bigcup_{s \in \mathbf{S}} k_{(l,s)},$$

где $\mathbf{O}$ — множество объектов, а $\mathbf{S}$ — множество субъектов в модели Белла-ЛаПадулы.

Здесь $k_{(l,s)}$, например, обозначает класс, находящийся в иерархии на уровне l, в котором может быть создан объект $o_{(l,s)}$ с двумя методами, реализующими функционал субъекта МБЛ $s \in \mathbf{S}$ по чтению и по записи соответственно. И такой объект существует в модели OOHRU тогда и только тогда, когда

соответствующий субъект в МБЛ находится на уровне безопасности l, т.е. $f_C(s) = l \leqslant f_S(s)$.

В свою очередь, $k_{(h,o)}$ обозначает класс, находящийся в иерархии на уровне h, в котором может быть создан объект $o_{(h,o)}$ с единственным полем, содержащим информацию объекта МБЛ $o \in \mathbf{O}$. И такой объект существует в модели OOHRU тогда и только тогда, когда соответствующий объект в МБЛ находится на уровне иерархии h, т.е. $f_H(o) = h$.

Таким образом, для каждого объекта МБЛ строится одно семейство классов в OOHRU, параметризованное уровнями секретности, а не два (отдельно для чтения и записи), как в доказательстве теоремы 1.

Во-вторых, модель OOHRU не будет обладать иерархией: даже если $l = F'(k_{(l,s)}) \leqslant F'(k_{(m,s)}) = m$, из этого не следует, что $k_{(l,s)} \leqslant k_{(m,s)}$.

В-третьих, матрицы доступа методов этих объектов также индуцируются матрицей доступа субъекта s. Это означает, что если $f_C(s) = l$, то

$$read \in o_{(h,o)}.M[o_{(l,s)}, o_{(h,o)}.field] \Leftrightarrow read \in M[s,o] \text{и} f_O(o) \leqslant l,$$

$$write \in o_{(h,o)}.M[o_{(l,s)}, o_{(h,o)}.field] \Leftrightarrow write \in M[s,o] \text{и} f_O(o) \geqslant l.$$

Наконец, список команд теперь имеет следующий вид:

1. Добавление права на чтение субъектом s уровня допуска l объекта o уровня секретности h.

Команда $EnterRead[l,h](o_{(l,s)} : k_{(l,s)}; o_{(h,o)} : k_{(h,o)})$

$Enter(read, o_{(l,s)}, o_{(h,o)}.field)$.

Такие команды существуют для всех l и h таких, что h не превосходит l, то есть для любого объекта o такого, что $f_H(o) = h$, выполняется $f_O(o) \leqslant l$.

2. Удаление права на чтение субъектом s уровня допуска l объекта o уровня секретности h.

Команда $DeleteRead[l,h](o_{(l,s)} : k_{(l,s)}; o_{(h,o)} : k_{(h,o)})$

$Delete(read, o_{(l,s)}, o_{(h,o)}.field)$.

3. Добавление права на запись субъектом s уровня допуска l в объект o уровня секретности h.

Команда $EnterRead[l,h](o_{(l,s)} : k_{(l,s)}; o_{(h,o)} : k_{(h,o)})$

$Enter(write, o_{(l,s)}, o_{(h,o)}.field)$.

Такие команды существуют для всех l и h таких, что h не меньше l, то есть для любого объекта o такого, что $f_H(o) = h$, выполняется $f_O(o) \geqslant l$.

4. Удаление права на запись субъектом s уровня допуска l в объект o уровня секретности h.

Команда $DeleteRead[l,h](o_{(l,s)} : k_{(l,s)}; o_{(h,o)} : k_{(h,o)})$

$Delete(write, o_{(l,s)}, o_{(h,o)}.field)$.

5. Присвоение субъекту s с текущим уровнем допуска l нового уровня допуска m.

Команда $ChangeSubjectSecurityLevel[l,m,\alpha](o_{(l,s)} : k_{(l,s)}; o_{(m,s)} : k_{(m,s)})$

$if \alpha \in o.M[o_{(l,s)}, o.field]$

$Create(o_{(m,s)}, k_{(m,s)})$,

$Enter(\alpha, o_{(m,s)}, o)$,

$Destroy(o_{(l,s)})$.

Данная команда присутствует в системе только для значений $l, m \leqslant f_S(s)$. α — набор прав, которыми обладает объект $o_{(l,s)}$ на поля других объектов o. Мы должны сохранить этот набор при изменении уровня допуска субъекта s в МБЛ. Для каждого набора прав α существует своя команда, и при переносе объекта с набором прав, в точности совпадающим с α, в другой класс в OOHRU из множества команд выбирается соответствующая.

6. Присвоение объекту o с текущим уровнем секретности h нового уровня секретности g.

Команда $ChangeObjectSecurityLevel[h, g, \alpha](o_{(h,o)} : k_{(h,o)}; o_{(g,o)} : k_{(g,o)}; o^{system} : k^{system})$

$if \alpha \in o_{(h,o)}.M[s, o_{(h,o)}.field]$

$Create(o_{(g,o)}, k_{(g,o)})$,

$Enter(read, osystem, o_{(h,o)})$,

$Enter(write, osystem, o_{(g,o)})$,

$Enter(\alpha, s, o_{(g,o)})$

$Destroy(o_{(h,o)})$.

Здесь α — это множество прав, которыми обладают объекты s на поле объекта $o_{(h,o)}$. Мы должны сохранить этот набор при изменении уровня секретности объекта o в МБЛ.

7. Создание объекта o уровня секретности h.

Команда $CreateObject[h](o_{(h,o)} : k_{(h,o)}; o^{system} : k^{system})$

$Create(o_{(h,o)}, k_{(h,o)})$,

$Enter(write, o^{system}, o_{(h,o)})$.

8. Удаление объекта o.

Команда $DestroyObject[h](o_{(h,o)} : k_{(h,o)})$

$Destroy(o_{(h,o)})$.

■

О безопасности МБЛ.

Важно отметить, что в обоих случаях мы построили дискреционную модель, безопасную с точки зрения МБЛ, однако, возможность проверки её безопасности с точки зрения дискреционных политик безопасности не установлена.

С другой стороны, Basic Security Theorem неоднократно подвергалась критике, в том числе и со стороны МакЛина [5], по ряду причин. Так, фактически Basic Security Theorem лишь утверждает выполнение определённого ряда свойств, но каким образом эти свойства влияют на фактическую безопасность системы, не ясно. Хуже того, Basic Security Theorem допускает деклассификацию всех субъектов и объектов до самого низкого уровня секретности без нарушения определения безопасности, поэтому МакЛин предложил определять безопасность системы не с точки зрения состояний, а с точки зрения переходов между состояниями.

Таким образом, построенные нами модели OOHRU, реализующие безопасные модели Белла-ЛаПадулы, также отнести к безопасным относительно утечки права доступа было бы преждевременно.

Литература

1. Усов С.В. Неоднородные объектно-ориентированные модели с иерархией // Проблемы обработки и защиты информации. Книга 3. Модели разграничения доступа. Коллективная монография / Под общей редакцией С.В. Белима. Омск : ООО «Полиграфический центр КАН», 2013. С. 93–114.
2. Harrison M.A., Ruzzo W.L., Ulman J.D. Protection in Operating Systems // Communications of the ACM. 1975. P. 14–25.
3. Bell, David Elliott and LaPadula, Leonard J. Secure Computer System: Unified Exposition and Multics Interpretation. Technical report 2997, rev. 1. MITRE, 1996.
4. Bell, David Elliott. Looking Back at the Bell-LaPadula Model // 21st Annual Computer Security Applications Conference. Tucson, Arizona, USA, 2005. P. 337–351.
5. McLean J. The Specification and Modeling of Computer Security // Computer. 1990. N. 23(1). P. 9–16.

ON THE RELATION BETWEEN THE OBJECT-ORIENTED DISCRETIONARY SECURITY MODEL AND THE SUBJECT-OBJECT MANDATORY MODEL

S.V. Usov
Ph.D. (Phys.-Math.), e-mail: raintower@mail.ru

Omsk State University n.a. F.M. Dostoevskiy

Abstract. The article deals with object-oriented Harrison-Ruzzo-Ullman access control model and subject-object model with mandatory access control. It is shown that the Bell-LaPadula model and classic mandatory model can be implemented with object-oriented HRU model.

Keywords: access control, discretionary safety models, mandatory security models, HRU, Bell-LaPadula model.

Дата поступления в редакцию: 31.10.2016

Авторам

Редакция журнала «Математические структуры и моделирование» предлагает авторам ознакомиться с данными правилами и придерживаться их при подготовке рукописей, направляемых в журнал.

Общие положения

К публикации принимаются рукописи объёмом не более 16 страниц.

Рукопись сопровождается краткой аннотацией на русском и английском языках (объёмом от 100 до 250 слов).

Авторам необходимо предоставить следующую информацию на русском и английском языках:

- название статьи;
- список авторов с указанием
  - фамилии, имени и отчества,
  - учёного звания,
  - учёной степени,
  - должности,
  - места работы или учёбы,
  - действующего адреса электронной почты;
- аннотация (абстракт);
- список ключевых слов.

Автор также указывает УДК (универсальный десятичный код) статьи. Его можно подобрать по тематике статьи в справочнике `http://msm.univer.omsk.su/udc/`.

Рукопись статьи представляется в редакцию по электронной почте в двух форматах pdf и tex. Статья должна быть набрана с использованием макропакета LATEX2e и стиля msmb.cls, предоставляемого редакцией `http://msm.univer.omsk.su/files/msmb.zip`. Рекомендуется установить компилятор MiKTEX, так как именно им пользуются в редакции.

Отклонения в оформлении рукописи от приведённых правил позволяют редколлегии принять решение о снятии статьи с публикации. Статья может быть отклонена по причинам несоответствия тематике журнала или в связи с низким уровнем качества научного исследования.

Требования к оформлению

При подготовке статьи следует использовать класс (стиль) msmb.cls и шаблон-заготовку для текущего номера. В шаблоне приведены все наиболее типичные примеры оформления формул, рисунков, таблиц, разделов, библиографических ссылок.

В статье запрещается переопределять стандартные команды и окружения.

Нумеруемые формулы необходимо выделять в отдельную строку.

Нумерация только арабскими цифрами в порядке возрастания с единицы. Нумеровать следует только те формулы, на которые в тексте имеются ссылки.

Запрещается использовать в формулах буквы русского алфавита. Если без них никак не обойтись, то следует использовать команду `\mbox{...}`.

Все рисунки и таблицы должны иметь подпись, оформленную с помощью команды `\caption{...}`.

Файлы с рисунками необходимо представить в формате PDF или EPS (использовать редакторы векторной графики типа InkScape, Adobe Illustrator или Corel Draw).

Используйте стандартные команды переключения на готический, каллиграфический и ажурный шрифты: `\mathfrak`, `\mathcal` и `\mathbb`.

Не допускается заканчивать статью рисунком или таблицей.

В списке литературы обязательно указание следующих данных: для книг — фамилии и инициалы авторов, название книги, место издания, издательство, год издания, количество страниц; для статей — фамилии и инициалы авторов, название статьи, название журнала, год издания, том, номер (выпуск), страницы начала и конца статьи (для депонированных статей обязательно указать номер регистрации).

Кавычки в русском тексте («абвгд») должны быть угловыми, в английском прямыми верхними кавычками ("abcdefg" или “abcdefg”).

Обязательна расшифровка сокращений при первом вхождении термина. Например: ... искусственный интеллект (ИИ)...

Порядок рецензирования

Первичная экспертиза проводится главным редактором (заместителем главного редактора). При первичной экспертизе оценивается соответствие статьи тематике журнала, правилам оформления и требованиям, установленным редакцией журнала к научным публикациям.

Все статьи, поступившие в редакцию научного журнала «Математические структуры и моделирование», проходят через институт рецензирования.

Рецензент выбирается главным редактором журнала из числа членов редколлегии или ведущих специалистов по профилю данной работы.

Рецензенты уведомляются о том, что присланные им рукописи являются частной собственностью авторов и относятся к сведениям, не подлежащим разглашению. Рецензентам не разрешается делать копии статей для своих нужд.

Срок для написания рецензии устанавливается по согласованию с рецензентом.

Рецензия должна раскрывать актуальность представленного материала, степень научной новизны исследования, определять соответствие предлагаемого к публикации текста общему профилю издания и стиль изложения.

Рецензент выносит заключение о возможности опубликования статьи: «рекомендуется», «рекомендуется с учётом исправления замечаний, отмеченных рецензентом» или «не рекомендуется». В случае отрицательной рецензии редакция направляет автору мотивированный отказ, заверенный главным редактором или его заместителем.

В случае несогласия с мнением рецензента автор статьи имеет право предоставить аргументированный ответ в редакцию журнала. Статья может быть направлена на повторное рецензирование, либо на согласование в редакционную коллегию.

При наличии в рецензии рекомендаций по исправлению и доработке статьи автору направляется текст рецензии с предложением учесть их при подготовке нового варианта статьи или аргументированно (частично или полностью) их опровергнуть. Доработанная (переработанная) автором статья повторно направляются на рецензирование и рассматривается в общем порядке. В этом случае датой поступления в редакцию считается дата возвращения доработанной статьи.

После принятия редколлегией решения о допуске статьи к публикации автор информируется об этом и указываются сроки публикации.

Оригиналы рецензий хранятся в редакции в течение пяти лет.

Памятка для перевода должностей, учёных степеней и званий на английский язык

Профессор = Professor
Доцент = Associate Professor
Старший преподаватель = Assistant Professor
Преподаватель = Instructor
Ассистент = Instructor
Аспирант = Postgraduate Student или Ph.D. Student
Соискатель = Ph.D. Doctoral Candidate
Магистрант = Master's Degree Student
Студент = Student
д.ф.-м.н. = Dr.Sc. (Phys.-Math.)
к.ф.-м.н. = Ph.D. (Phys.-Math.)
д.т.н. = Dr.Sc. (Eng.)
к.т.н. = Ph.D. (Eng.)
Инженер-программист = Software Engineer
Старший/младший научный сотрудник = Senior/Junior Scientist Researcher

Электронная почта для отправки статей

`lavrov@omsu.ru` — зам. главного редактора, выпускающий редактор Д.Н. Лавров.

Научный журнал

Математические структуры и моделирование

№4(40)

Главный редактор
А.К. Гуц

Выпускающий редактор
Д.Н. Лавров

Технический редактор
Н.Ф. Богаченко

Корректор:
И.Н. Баловнева
Проверка корректности перевода:
Е.А. Илюшечкин
А.Н. Кабанов

Адрес научной редакции
Россия, 644077, Омск, пр. Мира, 55А
Омский государственный университет
E-mail: guts@omsu.ru, lavrov@omsu.ru
Электронная версия журнала:
http://msm.univer.omsk.su
http://msm.omsu.ru

Подписано в печать 06.12.2016. Формат 60 × 84 1/8.
Усл. печ. л. 19,4. Тираж 100 экз. (1-й з-д 1-80) Заказ № 173.

Отпечатано на полиграфической базе издательства ОмГУ им. Ф.М. Достоевского
644077, г. Омск, пр. Мира, 55А

ISSN 2222-8772