УДК 519.688

АЛГОРИТМ ОПРЕДЕЛЕНИЯ РЕПРЕЗЕНТАТИВНОСТИ НЕДЕТЕРМИНИРОВАННОГО КОНЕЧНОГО АВТОМАТА

С.В. Пивнева, О. А.Рогова
Тольяттинский государственный университет
Тел. 89272093131, e-mail: tlt.swetlana@rambler.ru

Аннотация. Предлагается алгоритм случайной генерации недетерминированного конечного автомата, более подходящего для выбранной предметной области. Проверяется случайность генерируемой числовой последовательности. К сгенерированным структурам применяются конкретные характеристики данной предметной области. Сравниваются результаты, полученные при программной реализации рассматриваемых методов случайной генерации недетерминированного конечного автомата.

Ключевые понятия. Алгоритм, репрезентативность, случайная генерация, недетерминированный конечный автомат.

Репрезентативность (от франц. representatif – представляющий собой что-либо, показательный) в статистике - это основное требование к выборочной совокупности, заключающееся в соответствии её характеристик к соответствующим характеристикам генеральной совокупности – из которой с соблюдением определённых правил и отобрана выборочная. Суждение о степени репрезентативности выносится на основании рассмотрения выборочной совокупности в двух направлениях. Во-первых, она сравнивается с генеральной совокупностью в отношении всех признаков, зафиксированных как в той, так и в другой. Во-вторых, суждение о степени репрезентативности может быть вынесено на основании колеблемости исследуемых характеристик в выборочной совокупности.

Репрезентативность – величина измеряемая, которая может быть определена «ошибкой репрезентативности» – т.е. разностью между специально выбираемыми характеристиками выборочной и генеральной совокупностей. Однако фактическая (действительная) величина указанной разности остаётся неизвестной – вследствие чего мерой репрезентативности обычно служит определяемая по правилам математической статистики её вероятная величина (или среднее квадратичное её возможных значений).

Качество результатов выборочного наблюдения зависит от того, насколько выборка репрезентативна (представительна). Это обеспечивается следующими условиями: случайностью выбора объектов (результатов измерений) из генеральной совокупности, когда каждому из них обеспечивается одинаковая возможность быть отобранным (включенным в выборку); независимостью результатов наблюдений в выборке; правильным определением объема выборки

с учетом всех конкретных условий.

Различают различные способы формирования выборочной совокупности, такие как повторные и бесповторные случайные выборки, а также механический, стратифицированный случайный, серийный, гнездовой отбор.

Случайная генерация комбинаторных структур позволяет проверять алгоритмы, основанные на этой структуре, и исследовать поведение этих структур.

Сгенерированные объекты адекватны тем потребностям, которые возникают в реальных задачах. Например, при описании контекстно-свободных языков с помощью конечных автоматов. В реальных задачах требуется достаточно большое количество состояний конечного автомата, поэтому необходимо генерировать недетерминированные конечные автоматы с целью применения к ним конкретных характеристик, которые в последствии будут применены, например, к генерации LR-анализа.

Рассмотрим равновероятную случайную генерацию недетерминированного конечного автомата, имеющего n-состояний. Случайный метод генерации НКА основан на случайных битовых потоках. Ван Зиджл использовал метод с равновероятными битовыми потоками, чтобы успешно сравнивать различные представления регулярных языков.

Метод, используемый Ван Зиджлом для случайной генерации недетерминированного автомата:

- задан алфавит $\sum = \{1,...,m\}$ и набор состояний $Q = \{1,...,n\}$,
- произведены равновероятные битовые потоки размера mn^2; они описывают функцию перехода ∂; возникновение бита отличного от нуля в положении $(l-1)n^2 + (i-1)n + j$ обозначает существование перехода от состояния i в состояние j, помеченного *l*,
- есть единственное начальное состояние (вершина 1),
- набор финальных состояний случайно выбран, каждое состояние имеет равный шанс на то, чтобы быть финальным.

Этим методом строится недетерминированный конечный автомат $A = < Q, \sum, \delta, I, F >$, который не обязательно допустимый или ко-допустимый. Автомат допустимый тогда и только тогда, когда в любое состояние q из множества Q есть путь от одного из начальных состояний. Автомат ко-допустимый тогда и только тогда, когда из любого состояния q из Q есть путь в одно из финальных состояний.

Ван Зиджл использует этот метод, чтобы измерить сжатость недетерминированного конечного автомата n-состояний. Последовательность недетерминированных конечных автоматов построена, и для каждого недетерминированного конечного автомата выполнены следующие операции:

- проверяют, подходит ли этот автомат (неподходящие автоматы отклоняются),
- вычисляют эквивалентный детерминированный минимальный автомат.

Благодаря этим операциям размер распределения минимальных автоматов, эквивалентных случайному НКА данного размера, приводит к

результатам, подобным таковым из иллюстрации 1.

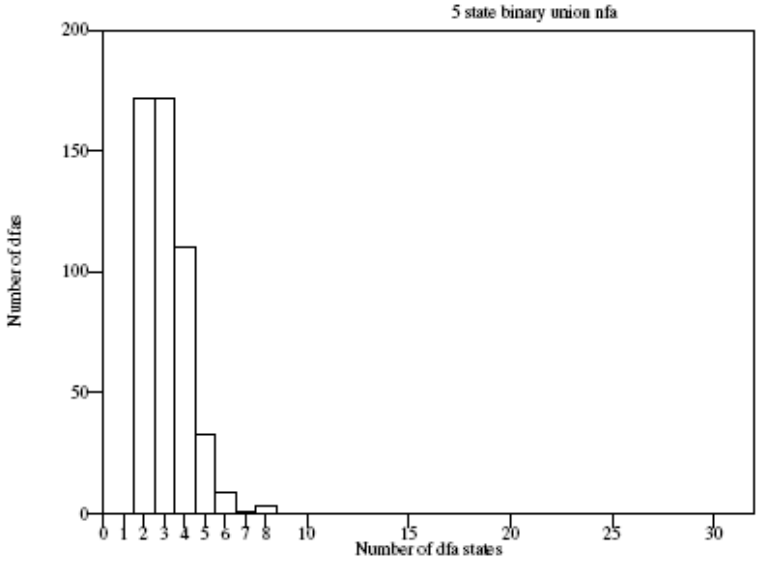

Иллюстрация 1. Число минимальных ДКА n-состояния для двойного НКА с 5 состояниями.

Генерируется подмножество, основанное на достижимости в случае равновероятных битовых потоков. Отображение данного подмножества размера *i* данным символом имеет размер k. Выводится вероятность, что отображение начального состояния, установленного словом длины t имеет размер k.

Генерируется подмножество, основанное на недостижимости. Определяется вероятность того, что состояние является достижимым от равновероятно выбранного состояния ДКА. Следовательно, распределение подмножеств, происходящее в течение строительства подмножества - равновероятное в случае битовых потоков, произведенных с вероятностью $2-2^{\frac{n-1}{n}}$. С одной стороны, отклоняющийся алгоритм может использоваться, чтобы сгенерировать допустимый НКА, и, с другой стороны, связанные ДКА имеют асимптотический размер m + 2. Асимптотическое поведение получено, как только автомат имеет больше чем 30 состояний. Генерация НКА, связанная с битовыми потоками с вероятностью $\frac{1}{x(n)} = 2-2^{\frac{n-1}{n}}$ приводит к оптимальному диапазону.

К сгенерированным этим методом автоматам были применены определенные характеристики, соответствующие рассматриваемой предметной области, такие как: разряженность автомата (количество выходящих из вершины дуг), вложенность циклов, минимальная длина пути от стартового до финального состояния, деленная на количество вершин.

Рассмотренный алгоритм генерации недетерминированного конечного автомата не устраивает, например, разреженностью автомата, поэтому был разработан другой алгоритм генерации недетерминированного конечного автомата, результатом работы которого являются автоматы, более

соответствующие рассматриваемым характеристикам, чем недетерминированные конечные автоматы, сгенерированные первым алгоритмом. Новый алгоритм был разработан при помощи эвристик.

Новый метод, используемый для случайной генерации недетерминированного автомата в выбранной предметной области:

- задан алфавит $\sum=\{1,...,m\}$ и набор состояний $Q=\{1,...,n\}$,
- произведены равновероятные битовые потоки размера mn^2; они описывают функцию перехода ∂; возникновение бита отличного от нуля в положении $2*((l-1)*n^2+(i-1)+j$ обозначает существование перехода от состояния i в состояние j, помеченного *l*;
- строится трехмерная матрица переходов вида:

	q_1	q_2	…	q_n
q_1				
q_2				
…				
q_n				

В ячейках матрицы записываются символы, которые можно прочитать при переходе из состояния q_i в состояние q_j;

- есть единственное начальное состояние (вершина 1),
- набор финальных состояний случайно выбран, каждое состояние имеет равный шанс на то, чтобы быть финальным.

Для проверки случайности сгенерированной последовательности чисел применяются статистические критерии. Если критерии $T_1,T_2,...,T_n$ подтверждают, что последовательность ведет себя случайным образом, это не означает, что проверка с помощью T_{n+1}-го критерия будет успешной. Но каждая успешная проверка дает все больше уверенности в случайности последовательности. Обычно к последовательности применяется около шести статистических критериев, и если она удовлетворяет этим критериям, то последовательность считается случайной.

Критерий «хи - квадрат». Проводим n независимых наблюдений, каждое наблюдение может принадлежать к одной из k категорий. Пусть p_s - вероятность того, что каждое наблюдение относится к категории s, пусть Y_s - число наблюдений, которые действительно относятся к категории s. Образуем статистику

$$V=\frac{1}{n}\sum_{s=1}^{k}(\frac{Y_s^2}{p_s})-n$$

Приемлемое значение статистики V можно определить по таблице 1 (Приложение 1), которая дает значения «χ^2- распределения с v степенями свободы» для различных значений v. Используется строка таблицы с $v=k-1$, так как число «степеней свободы» равно $k-1$, что на единицу меньше, чем число категорий. Если V меньше 1%-й точки или больше 99%-й точки, эти числа отбрасываются как недостаточно случайные. Если V лежит между 1%- и 5%-й точками или между 95%- и 99%-й точками, то эти числа

«подозрительны»; если V лежит между 5%- и 10%-й точками или 90%- и 95%-й точками, числа можно считать «почти подозрительными». Проверка по χ^2-критерию проводится три раза (и более) с разными данными. Если по крайней мере два из трех результатов оказываются подозрительными, то числа рассматриваются как недостаточно случайные.

Критерий равномерности. Проверяется равномерность распределения чисел. Выбирается число d. Для каждого r, $0 \le r < d$, подсчитывается число случаев, когда $Y_j = r$ для $0 \le j < n$, а затем применяется χ^2-критерий, принимая $k = d$ и вероятности $p_s = 1/d$ для каждой категории.

Критерий серий. Проверяется требование к последовательности, состоящее в том, чтобы пары последовательных чисел были равномерно распределены независимым образом. Подсчитываем число случаев, когда пара $(Y_{2j}, Y_{2j+1}) = (q, r)$ для $0 \le j < n$. Такая операция осуществляется для каждой пары целых чисел (q, r), таких, что $0 \le q, r < d$. Затем применяется χ^2-критерий к этим $k = d^2$ категориям, где $1/d^2$ - вероятность отнесения пары чисел к каждой из категорий.

Критерий интервалов. Этот критерий используется для проверки длины «интервалов» между появлением U_j на определенном отрезке. Если α и β - два действительных числа, таких, что $0 \le \alpha < \beta \le 1$, то рассмотрим длины подпоследовательностей $U_j, U_{j+1}, ..., U_{j+r}$, в которых U_{j+r} лежит между α и β, а другие U_s не лежат между этими числами. (Эту подпоследовательность, состоящую из $r+1$ числа, будем называть интервалом длиной r.)

χ^2-критерий применяется при $k = t+1$ к значениям $count[0], count[1], ..., count[t]$ ($count[r]$ - число интервалов длиной r) с использованием следующий вероятностей:

$$p_r = p(1-p)^r \text{ для } 0 \le r \le t-1; \quad p_t = (1-p)^t$$

Здесь $p = \beta - \alpha$ - вероятность того, что $\alpha \le U_j < \beta$. Значения n и t выбираются так, чтобы ожидаемое значение $count[r]$ равнялось 5 или больше.

Покер – критерий (критерий разбиений). Покер-критерий рассматривает n групп по пять последовательных целых чисел $\{Y_{5j}, Y_{5j+1}, Y_{5j+2}, Y_{5j+3}, Y_{5j+4}\}$ для $0 \le j < n$ и проверяет, какие из следующих пяти категорий соответствуют таким пятеркам чисел:

5 значений = все разные;
4 значения = одна пара;
3 значения = две пары или три числа одного вида;
2 значения = полный набор или четыре числа одного вида;
1 значение = пять чисел одного вида.

В общем случае можно рассматривать n групп k последовательных чисел и подсчитывать число групп из k чисел с r различными числами. Затем применяется χ^2-критерий, в котором используются вероятности того, что в группе r различных чисел

$$p_r = \frac{d(d-1)...(d-r+1)}{d^k}\left\{ {k \atop r} \right\}, \quad \left\{ {r \atop k} \right\} = \prod_{j=1}^{k} \frac{r+1-j}{j}$$

Критерий собирания купонов. Используется последовательность $Y_0, Y_1, ...$ и находятся длины отрезков $Y_{j+1}, Y_{j+2}, ..., Y_{j+r}$, содержащие «полный набор» целых чисел от 0 до $d-1$.

Если дана последовательность целых чисел $Y_0, Y_1, ...$, таких, что $0 \le Y_j < d$, то подсчитываются длины n последовательных «собравших купоны» отрезков. $count[r]$ - это число отрезков длиной r для $d \le r < t$, а $count[t]$ - это число отрезков длиной $\ge t$.

После того как вычислено n длин, нужно применить χ^2-критерий к $count[d], count[d+1], ..., count[t]$ с $k = t - d + 1$. Соответствующие вероятности равны

$$p_r = \frac{d!}{d^r}\left\{ {r-1 \atop d-1} \right\}, \quad d \le r < t; \qquad p_t = 1 - \frac{d!}{d^{t-1}}\left\{ {t-1 \atop d} \right\}$$

Критерий сериальной корреляции. Если задано n величин $U_0, U_1, ..., U_{n-1}$ и n других величин $V_0, V_1, ..., V_{n-1}$, то коэффициент корреляции между ними определяется следующим образом:

$$C = \frac{n\sum(U_j V_j) - (\sum U_j)(\sum V_j)}{\sqrt{(n\sum U_j^2 - (\sum U_j)^2)(n\sum V_j^2 - (\sum V_j)^2)}}, \quad 0 \le j < n$$

Коэффициент корреляции всегда лежит между -1 и +1. Когда он равен 0 или очень мал, значит, величины U_j и V_j независимы одна от другой (между ними нет линейной зависимости); если же значение коэффициента корреляции равно +1 или -1, это означает полную линейную зависимость.

Описанные критерии были применены к сгенерированной последовательности чисел. Практическая реализация данных критериев показала, что сгенерированная последовательность чисел достаточно случайна, требования всех критериев были выполнены.

К сгенерированным недетерминированным конечным автоматам были применены рассматриваемые характеристики и получены следующие результаты.

1 характеристика. Разреженность автомата - разработанный метод должен генерировать автоматы с уменьшенным средним числом дуг из вершины.

2 характеристика. Вложенность циклов - уровень вложенности циклов должен быть увеличен.

3 характеристика. Минимальная длина пути от стартовой вершины до финальной, деленная на количество вершин должна уменьшиться.

Программная реализация разработанного метода случайной генерации недетерминированного конечного автомата показала, что сгенерированные структуры удовлетворяют требуемым характеристикам.

Результаты программной реализации обоих методов продемонстрированы на представленных диаграммах.

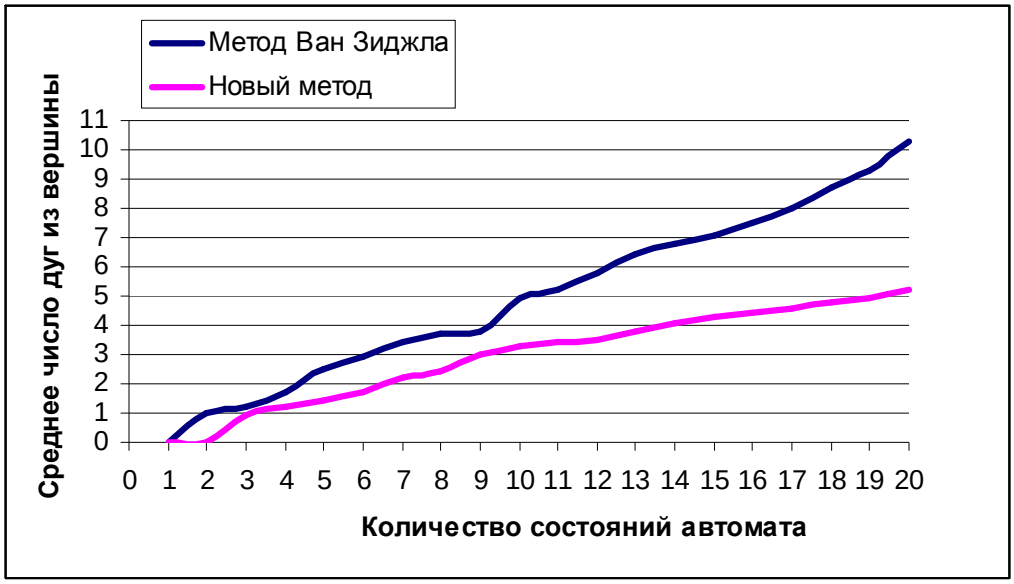

Диаграмма 1. Разреженность автомата. Автоматы, полученные новым методом более разрежены.

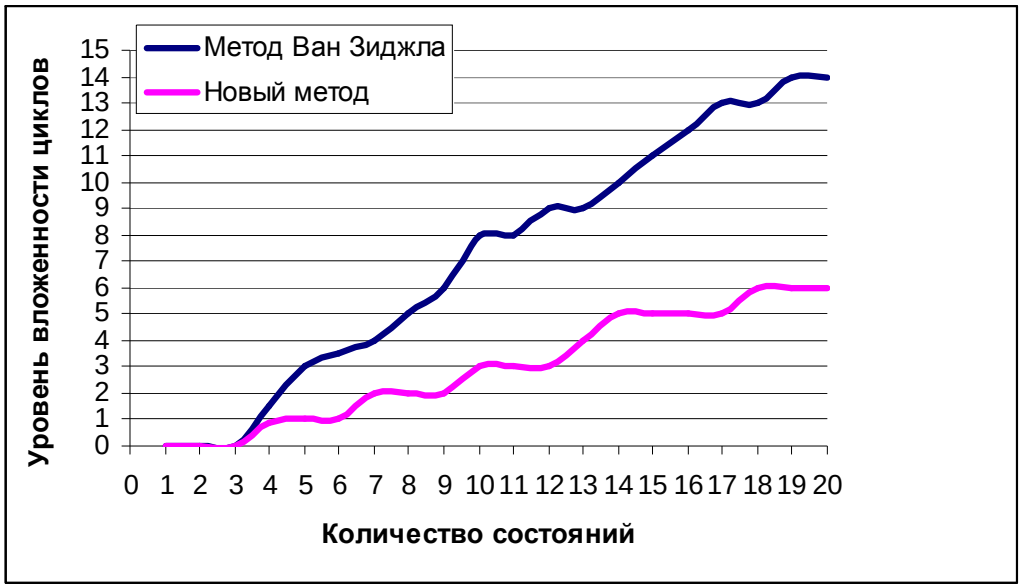

Диаграмма 2. Вложенность циклов. Очевидно, что уровень вложенности циклов стал ниже.

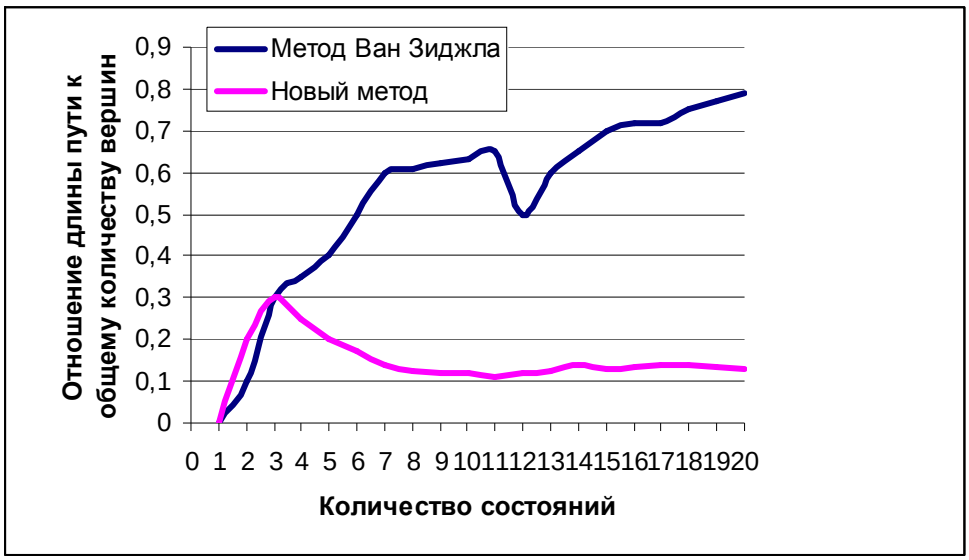

Диаграмма 3. Отношение длины пути к общему числу вершин.

Таким образом, на представленных графиках видно, что разработанный метод генерирует структуры, более подходящие к использованию в рассматриваемой предметной области. Очевидно преимущество использования разработанного метода случайной генерации недетерминированного конечного автомата.

Полученные объекты адекватны тем потребностям, которые возникают в реальных задачах. Например, при описании контекстно-свободных языков с помощью конечных автоматов. В реальных задачах требуется достаточно большое количество состояний конечного автомата, поэтому необходимо генерировать КА с целью применения конкретных характеристик к ним. Сгенерированные недетерминированные конечные автоматы могут использоваться для решения задач в этой предметной области.

Практическая часть работы выполнена на языке C++. Реализованы оба метода случайной генерации недетерминированного конечного автомата, реализовано применение конкретных характеристик данной предметной области. Проверена случайность генерируемой числовой последовательности.

Для случайной генерации недетерминированного конечного автомата используется генерация случайных чисел. Для генерации случайных чисел используются стандартные функции языка C rand и srand.

Функция rand стандартной библиотеки C генерирует целое число в диапазоне между 0 и RAND_MAX. Значение RAND_MAX должно быть по меньшей мере равно 32767 – максимальное положительное значение двухбайтового целого числа. Если rand действительно вырабатывает случайные целые числа, то при каждом вызове rand результирующее число имеет равную вероятность оказаться любым целым, лежащим между 0 и RAND_MAX.

Функция rand на самом деле генерирует псевдослучайные числа. Повторный вызов rand производит последовательность чисел, которые кажутся случайными. Но та же самая последовательность повторяется при каждом повторении программ. Когда программа тщательно отлажена, она может быть использована для получения разных последовательностей случайных чисел при

каждом выполнении.

Это называется рандомизацией и реализуется в законченном виде с помощью стандартной библиотечной функции srand. Функция srand получает целый аргумент unsigned и при каждом выполнении программы задает начальное число, которое функция rand использует для генерации последовательности квазислучайных чисел.

Чтобы рандомизировать не вводя каждый раз начальное число, можно использовать оператор srand(time(NULL)). При этом для автоматического получения начального числа компьютер считывает показания своих часов. Функция time (с аргументом NULL)возвращает текущее «календарное время» в секундах. Это значение преобразуется в беззнаковое целое число и используется как начальное значение в генераторе случайных чисел.

Была реализована проверка случайности генерируемой последовательности чисел по шести статистическим критериям, рассмотренным в дипломной работе.

Реализованы оба метода случайной генерации недетерминированного конечного автомата, а также реализованы рассматриваемые характеристики. Получены практические результаты, из которых видны преимущества использования разработанного метода случайной генерации недетерминированного конечного автомата в выбранной предметной области.

В рамках данной работы реализован алгоритм случайной генерации недетерминированного конечного автомата Ван Зиджла. Разработан и реализован алгоритм случайной генерации недетерминированного конечного автомата, более подходящего для выбранной предметной области. Проверена случайность генерируемой числовой последовательности. К сгенерированным структурам применены конкретные характеристики данной предметной области. Практическая реализация характеристик показала, что комбинаторные структуры, полученные разработанным методом случайной генерации, больше соответствует требованиям выбранной предметной области. Сгенерированные недетерминированные автоматы репрезентативны, и в дальнейшем могут использоваться для решения задач, возникающих в рассматриваемой предметной области.

Сведения об авторах

Пивнева Светлана Валентиновна *– кандидат педагогических наук, доцент, Тольяттинский государственный университет, кафедра высшей математики и математического моделирования,* Тел. 89272093131,
e-mail: tlt.swetlana@rambler.ru

Рогова Ольга Алексеевна *– Ассистент, Тольяттинский государственный университет, кафедра прикладной математики и информатики*,
e-mail: rogovaolgatlt@yandex.ru