

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 $\psi_i = \{a_{i_1}, a_{i_2}, \dots, a_{i_K}\}$ 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 ($s \geq 0$) 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] $d_H(A_1, A_2)$:

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

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

¹1 — полное сходство, 0 — полное несходство.

расстояния Хэмминга по следующему соотношению:

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

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:

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

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

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

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 -грамм². m -граммами являются сочетания из m смежных символов³. В качестве меры сходства строк X и Y используется величина μ_m :

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

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.

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

³Например, 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 однограмм, 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 «однограммных» совпадений n_1 , «двуграммных» n_2 , ..., количество « m -граммных» совпадений n_m .

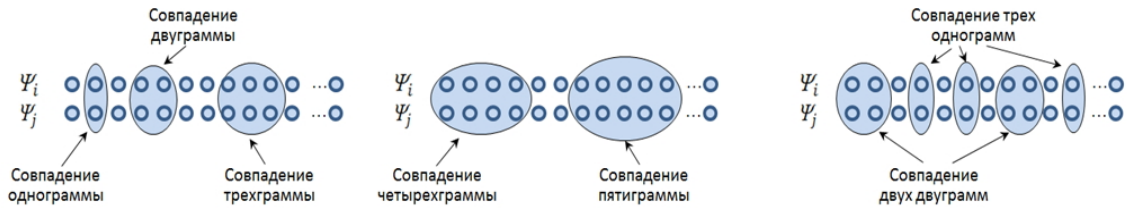


Рис. 1. Примеры сходства последовательностей по совпадению однограмм, двуграмм и т.д.

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).

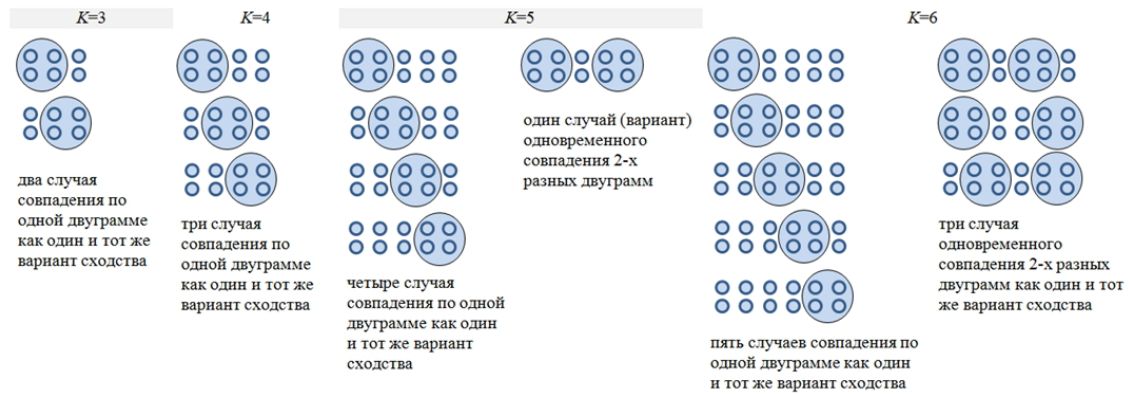


Рис. 2. Варианты совпадения двуграмм при $K = 3$, $K = 4$, $K = 5$ и $K = 6$

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 $(t + 1)$ «разбросанных» по разным позициям однограмм (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) — $\sum_{m=1}^K m n_m$. Каждое слагаемое $m n_m$ 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 $m n_m$ в зависимости от размерности m -грамм.

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

$$\mu = \frac{1}{K} \sum_{m=1}^K m n_m c_m, \quad (6)$$

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

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 $m n_m c_m$ ($m = 1, 2, \dots, K$).

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

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

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

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 $c_m = 1$ ($m = 1, 2, \dots, K$) 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 $c_m = 1$ ($m = 1, 2, \dots, K$) 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 однограммам, двуграммам, триграммам и т.д.

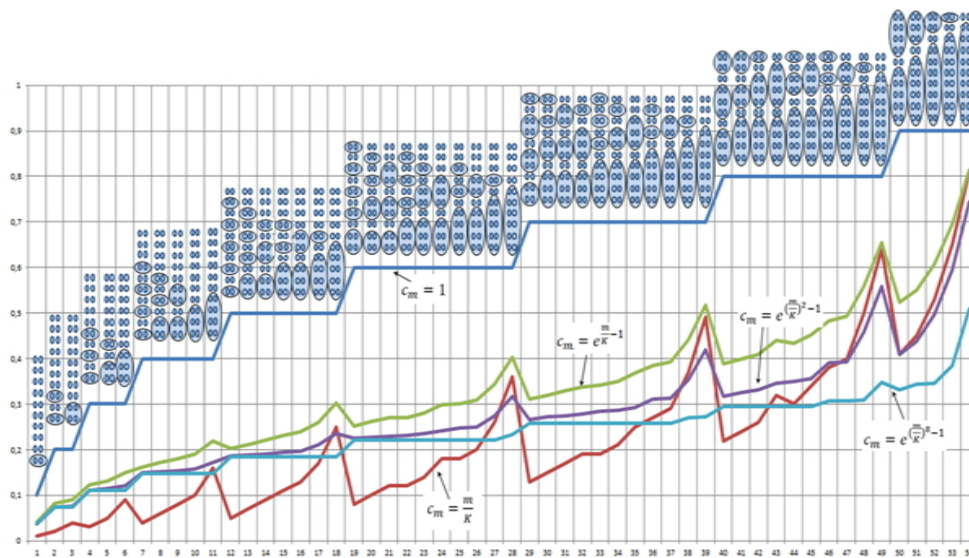


Рис. 4. Зависимость коэффициента сходства последовательностей при $K = 10$ от количества и качества совпадений при различных видах функции $c_m = f\left(\frac{m}{K}\right)$

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» переходах количества-качества совпадений (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 двуграмм⁴ 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

⁴Напомним, 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 $(m + 1)$:

$$\max_m(K) = \left\lfloor \frac{K + 1}{m + 1} \right\rfloor, \quad (7)$$

где $\lfloor \cdot \rfloor$ 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,

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

■

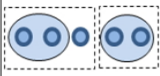
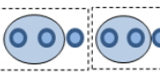
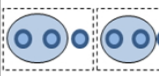
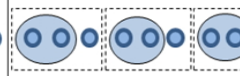
$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
					
Одна двуграмма и одна совокупность из 3-х позиций	Одна двуграмма, одна совокупность из 3-х позиций и одна «свободная» позиция	Две несливающихся двуграммы, одна полная и одна неполная на одну позицию совокупности из 3-х позиций	Две несливающихся двуграммы и две полных совокупности из 3-х позиций	Две несливающихся двуграммы, две полных совокупности из 3-х позиций и одна «свободная» позиция	Три несливающихся двуграммы, две полных и одна неполная на одну позицию совокупности из 3-х позиций
$\max_2(3) = 1$	$\max_2(4) = 1$	$\max_2(5) = 2$	$\max_2(6) = 2$	$\max_2(7) = 2$	$\max_2(8) = 3$

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 $m = 1, 2, \dots, 6$ и $K = 2, 3, \dots, 12$

	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	$K=11$	$K=12$
$\max_1(K)$	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 $\mu = \frac{1}{K} \sum_{m=1}^K mn_m c_m$ в «критичных» точках — при переходе от варианта совпадения по одной m -грамме к варианту с $(m + 1)$ совпадений однограмм.

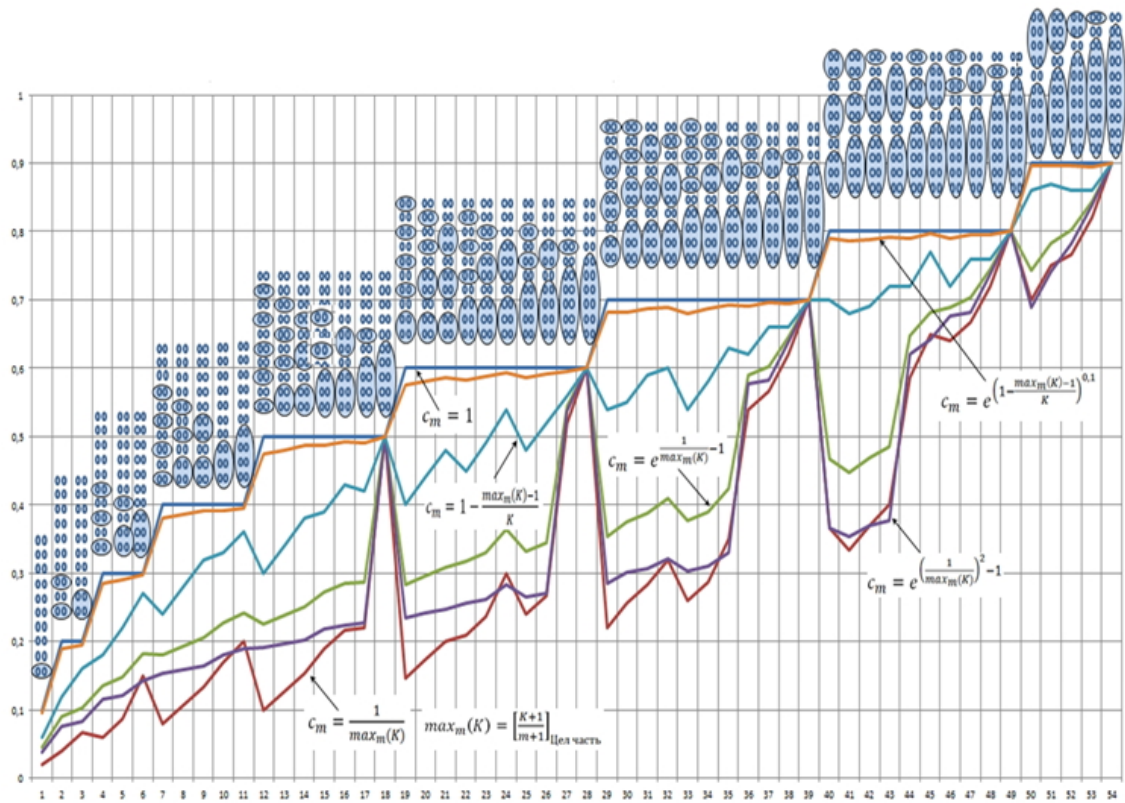


Рис. 7. Зависимость коэффициента сходства последовательностей при $K = 10$ от количества и качества совпадений при различных видах функции $c_m = f\left(\frac{1}{\max_m(K)}\right)$

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, близки 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

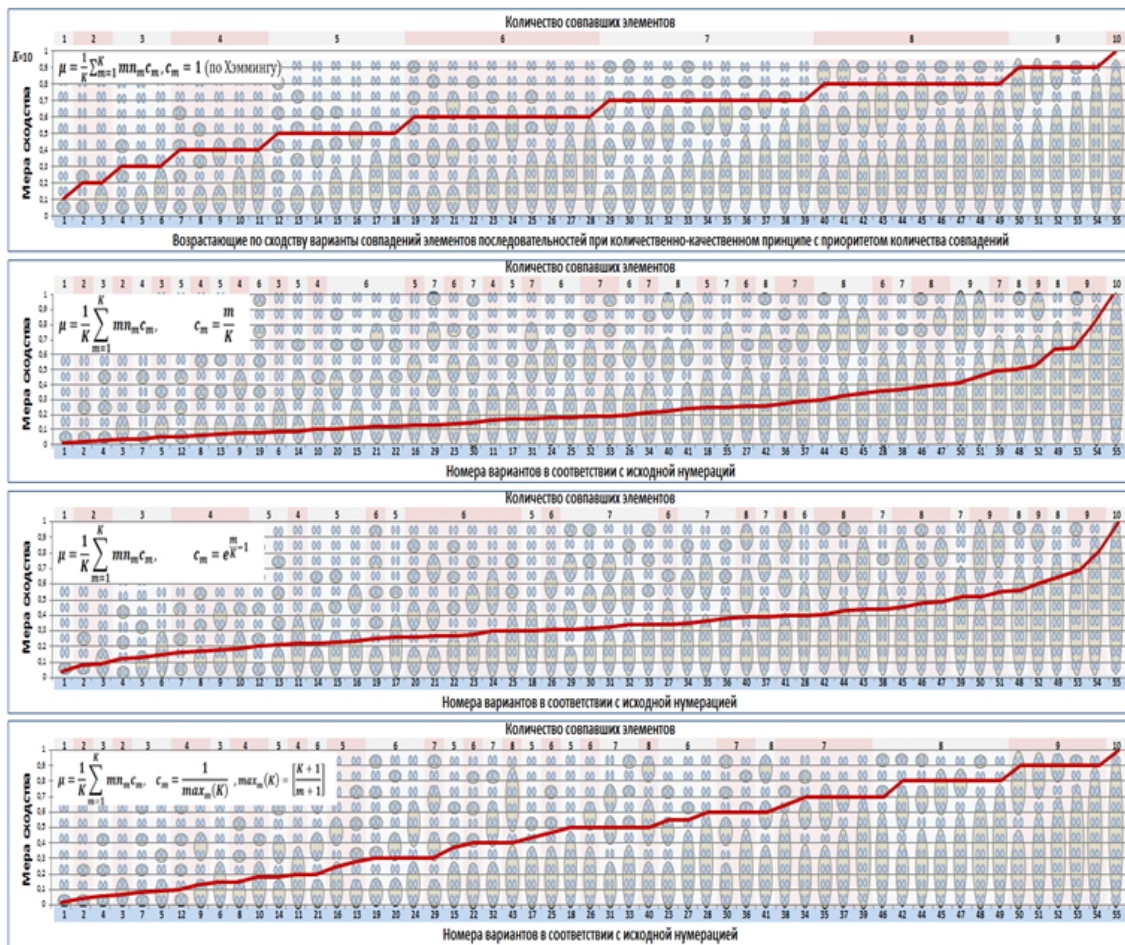


Рис. 8. Возрастающие по сходству ряды вариантов совпадений элементов последовательностей при различных видах функции $c_m = f\left(\frac{m}{K}\right)$

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., Двойченко 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