С. В. Знаменский

Приближение длины наибольшей общей подпоследовательности пары случайных строк

Аннотация. Математическое ожидание E длины длиннейшей общей подпоследовательности букв двух случайных слов рассматривается как функция от длин m и n этих слов и мощности алфавита $\alpha = A$. При этом предполагается, что любая буква независимо и с равной вероятностью оказывается в любой позиции слова.

Указан вид приближённой формулы для $E(m, n, \alpha)$, позволяющий вычислять $E(m, n, \alpha)$ с погрешностью в 0.3 процента для $64 \leq m + n \leq 65536$ и $1 < \alpha < 129$. Коэффициенты подобраны вручную и могут быть уточнены. Ожидается, что формула справедлива для всех больших значений аргументов с той же относительной погрешностью.

Ключевые слова и фразы: сходство строк, выравнивание последовательностей, случайные общие подпоследовательности, LCS, метрика Левенштейна.

Введение

Два случайных слова длин m и n из алфавита α часто иначе называют случайными последовательностями символов. Будем считать появления букв в различных позициях равновероятными и независимыми событиями. Тогда математическое ожидание длины наибольшей общей подпоследовательности этих случайных последовательностей является функцией $E(m, n, \alpha)$, характеризующей близость исходных слов. Эта функция тесно связана с эффективностью разнообразных алгоритмов нечёткого поиска и выделения различий, поэтому её поведение с 70-х годов прошлого века привлекало внимание исследователей [1], выявивших линейную асимптотику при фиксированном α при больших равных длинах m и n и приближённо вычисливших коэффициенты пропорциональности

$$\gamma_\alpha = \lim_{n \rightarrow \infty} \frac{E(n, n, \alpha)}{n}$$

Рис. 1. Зависимость $f_3(n) = E(n, 2048 - n, 3)$

для некоторых фиксированных α . Для уточнения численного значения этих констант последователи использовали как сложный математический аппарат, так и численное моделирование с применением особых алгоритмов [2, 3]. Для случая $m = n$ в [4] эмпирически получена формула

$$(1) \quad \frac{E(n, n, \alpha)}{n} = \gamma_\alpha + \frac{A_\alpha}{\sqrt{n} \ln n} + \epsilon(n, \alpha),$$

точность которой подтверждается и на использованных автором данных. Однако зависимость от α так и осталась не вполне ясной. Известна [5, 6] лишь оценка асимптотики

$$\lim_{\alpha \rightarrow \infty} \gamma_\alpha \sqrt{\alpha} = 2.$$

Не исследован и общий случай $m \neq n$, для которого характер асимптотической зависимости от $\frac{m}{n}$ при $\alpha = 2$ начал проявляться лишь в [7], хотя, к сожалению и без подробного обоснования. Компьютерные вычисления E при небольших m, n ([8]) выявили похожую связь при $\alpha = 4$, подтвердившуюся с сопоставимой с точностью экспериментальных данных погрешностью порядка 0.2% в диапазоне $1 < \alpha < 15$, $63 < n < 65537$ [9]. Оказалось, что при фиксированных α и больших $m + n = s$ график функции $f_{s,\alpha}(n) = E(n, s - n, \alpha)$ практически точно состоит из отрезков прямых $y = x$ и $y = s - x$ и соединяющей их дуги эллипса. Левая половина этого симметричного графика показана на рис. 1. Аналитическое выражение для функции

стандартно выписывается через дуги полуосей r_x и r_y этого эллипса:

$$(2) \quad E(m, n, \alpha) = \begin{cases} n, & 1 < r\delta \\ \frac{m+n}{2} - r + r_y \sqrt{1 - \delta^2}, & -1 < r\delta < 1 \\ m, & r\delta < -1 \end{cases},$$

где $\delta = \frac{m-n}{2r_x}$, $r = \sqrt{r_x^2 + r_y^2}$.

1. Общая постановка задачи поиска формулы для $E(m, n, \alpha)$

Общая задача исследования — найти практичную приближённую формулу для $E(m, n, \alpha)$. Выделим целевые признаки практичности искомой формулы для приближения экспериментальных данных:

- точность аппроксимации, дающая возможность строить визуально точные графики функции $E(m, n, \alpha)$ на любых прямых и плоскостях;
- лаконичность и выразительность формулы;
- рост точности с ростом значений переменных;
- единственность экстраполяции при правдоподобных априорных предположениях;
- правдоподобность поведения при больших значениях аргументов.

Поиск такой формулы можно разделить на два этапа. Задачами первого этапа являются:

- (1) определение требований к постановке эксперимента и подготовке данных;
- (2) определение априорных предположений, обеспечивающих единственность продолжения и общего искомого вида исследуемых зависимостей;
- (3) предварительная проверка на частичных данных и оценка достижимости поставленной цели.

Задачи второго этапа — сбор недостающих данных, оптимизация коэффициентов формулы и уточнение её вида и итоговые оценки оптимизированной формулы — выходят за рамки настоящей статьи.

Рис. 2. Графики остатков формул, аппроксимирующих $E(n, n, \alpha)$ при $2 \leq \alpha \leq 128$ и $64 \leq s \leq 65536$

2. Особенности подготовки данных

Задача обеспечения заданной точности для исходных данных оказалась непростой. Время на получение значения с заданной точностью неожиданно оказалось существенно зависящим не только от s и α . Дисперсия при независимых изменениях длин меняется резко и этот эффект не объясняется доказанной в [10] асимптотикой $D \sim \sqrt{n}$. Например, если при $\alpha = 2$ для $(m, n) = (401, 623)$ экспериментальная дисперсия получилась равна 0.0222, то для $(m, n) = (329, 695)$ на порядки меньше: 0.0002.

Получается, что вычисления $E(401, 623, 2)$ и $E(329, 695, 2)$ с одинаковой точностью требуют примерно в сто раз отличающихся количеств тестов и машинного времени. Поэтому собранные для старых значений данные оказались сильно разнородны. Они были дополнены, что привело к двукратному росту точности формулы (2) на суммарных длинах 32768, 65536. Сложность ещё в том, что при близких к n значениях $E(n, m, \alpha)$ отклонения с превышением n невозможны, и вызванная этим сильная асимметрия делает сомнительной возможность правильной оценки по (выборочной) дисперсии без учёта неисследованных моментов третьего и высших порядков.

Хорошо известно, что генерация исходных данных требует значимых ресурсов, поэтому важен выбор исходной сетки значений независимых переменных. Сетка (декартово произведение множеств) полезна тем, что упрощает визуальный контроль за выделением членов асимптотики. Переменные (m, n) при этом симметричны, и с их ростом задача более существенно усложняется, чем с ростом α . Уже при начальном $\alpha = 2$ задача сложна и несколько десятилетий исследований

Рис. 3. Графики зависимости полуосей эллипса от s и α

не привели к её полному решению, а при малых $m = 1$, $n = 1$ для произвольных α она превращается в простое стандартное упражнение по теории вероятностей. Отсюда разный ожидаемый характер зависимости по этим переменным и разные требования к сетке.

Асимптотика по (m, n) сложна по своей природе и плохо описывается суммой степенных функций. Для обоснованного выделения членов асимптотики необходим достаточно широкий диапазон изменения переменных: каждый член асимптотической суммы имеет свою зону влияния. Хотя эти зоны могут перекрываться, у главного члена суммы есть часть, где он доминирует почти безраздельно. Логарифмическая сетка облегчает поиск вида очередного члена аппроксимации сложной зависимости. Рассмотрим это на примере уточнения формулы де Монвеля (1) с использованием описанных в [9] и несколько дополненных данных.

На рис. 2 слева отчётливо видна систематическая ошибка формулы. Сетка, близкая к логарифмической, позволила увидеть логарифмическое и постоянное слагаемое, на порядок уменьшающее среднюю ошибку в диапазоне $64 \div 1024$ и существенно не изменяющую её при больших n , когда ошибка формулы теряется за неточностью экспериментальных данных. На графиках видна и проблематичность дальнейшего повышения точности на имеющихся исходных данных.

Неравномерная сетка мешает увидеть зависимость от двух переменных. Разумный компромисс состоит в использовании двух или нескольких серий равномерных (для α) или логарифмических (для m, n) отсчётов.

3. Искомый общий вид и единственность продолжения

Графики искомых зависимостей r_x и r_y показаны на рис. 3.

Рис. 4. Зависимости $g_1(\alpha)$, $g_2(\alpha)$ и $g_3(\alpha)$ для $2 \leq \alpha \leq 128$

Даже для зависимости $r_x(s, \alpha)$ (которая проще, чем $r_y(s, \alpha)$) операций суммирования произведений степенных и логарифмических функций недостаточно для получения конечной аппроксимирующей формулы, и существование простой убедительно точно аппроксимирующей формулы крайне сомнительно.

Представляется разумным использовать одномерные рациональные аппроксимации с полюсами в левой полуплоскости, поскольку для них можно строго гарантировать однозначность в следующем смысле: любая голоморфная и ограниченная в правой полуплоскости и вне круга небольшого радиуса аппроксимирующая функция будет давать приближение на всём положительном луче с практически той же точностью, с какой она приближает в исследованных точках. Строгие формулировка и доказательство этого утверждения могут быть получены стандартной техникой. Отсюда особое понимание точности найденной аппроксимации на множестве всех натуральных чисел: если существует какая-нибудь аппроксимирующая на всём луче аналитическая функция и её особенности далеки от положительного луча вещественной оси, то её значения от любых натуральных значений аргумента незначительно отличаются от нами найденной.

Сказанное означает, что как правило при возможности для подобных аппроксимаций должны использоваться рациональные функции. Хотя последние можно представлять частным полиномов, но разложение на сумму простейших дробей гораздо выразительнее передаёт особенности поведения.

4. Приближённые формулы для $r_x(s, \alpha)$ и $r_y(s, \alpha)$

Основной результат статьи — это возможность аналитического представление $r_x(s, \alpha)$ в виде

$$(3) \quad r_x(s, \alpha) = g_1(\alpha) \cdot s + g_2(\alpha) \cdot \ln s + g_3(\alpha) + \epsilon(s, \alpha)$$

Рис. 5. Погрешности аппроксимации

с относительно малым ϵ и рациональными функциями

$$(4) \quad g_1(\alpha) = 0.5002 - \frac{0.02\alpha + 0.442}{(\alpha - 1.01)^2 + 1.115} - \frac{0.0192\alpha - 0.0356}{(\alpha - 1.85)^2 + 1.39}$$

$$(5) \quad g_2(\alpha) = 1.2 + \frac{73\alpha}{\alpha^2 + 10}$$

$$(6) \quad g_3(\alpha) = \frac{129\alpha}{\alpha^2 + 16}$$

с полюсами в треугольнике $-4 + \text{Im } \alpha \leq \text{Re } \alpha \leq 0$.

Поскольку $E(n, n, \alpha) = n - \sqrt{r_x^2 + r_y^2} + r_y$, то r_y несложно выразить через r_x и формулу де Монвеля и получить, что

$$(7) \quad r_y(s, \alpha) = h_1(\alpha) \cdot s + h_2(\alpha) \sqrt{s} \cdot \ln s + h_3(\alpha) \cdot \ln s + h_4(\alpha) + \tilde{\epsilon}(s, \alpha)$$

для рациональных $h_1(\alpha)$, $h_2(\alpha)$, $h_3(\alpha)$ и $h_4(\alpha)$ и ограниченного остатка.

Коэффициенты функций $h_i(\alpha)$ могут быть независимо от формулы де Монвеля найдены минимизацией суммы квадратов значений остаточного члена.

5. Оценки и перспективы

Как видно из графиков на рис. 5, при больших s ошибка ϵ не превышает десятой доли процента длины диапазона и визуально исчезает. Абсолютное значение ошибки растёт с увеличением s , а относительное — убывает.

Нерегулярность со стороны длинных последовательностей может быть вызвана недостаточным качеством экспериментальных данных

и, пока коэффициенты подобраны вручную, не даёт оснований сомневаться в применимости формулы для больших значений s .

Ясно просматривается регулярность ошибки при малых значениях, очевидно указывающая на возможность уточнения функций g_i . Каждую из них разумно искать в виде суммы постоянного слагаемого и минимального количества дробей с постоянными числителями и квадратичным или линейным знаменателем. Неточность исходных данных побуждает вычислить коэффициенты дробей, минимизирующие среднеквадратическую ошибку формулы. Учитывая, что производительность таких алгоритмов оптимизации резко падает при росте количества переменных, быстрее получится оптимизировать каждую из функций g_i при фиксированных остальных и повторить это несколько раз.

Количество дробей-слагаемых в g_i разумно увеличивать только если имеющихся дробей недостаточно. Это потребует лишь в случае, если регулярность графика остатка, видимая на рис. 5 слева, не изменит скорость роста после оптимизации коэффициентов.

Возможно расширение диапазона длин в формуле в сторону коротких на основе вычисления точных статистик как в [8]. Добавление в формулу новых слагаемых с отрицательными степенями s может привести к её неоправданному усложнению. Поэтому когда будут оптимизированы g_1 , g_2 и g_3 , интересно было бы найти рациональную убывающую на произведении положительных лучей функцию двух переменных небольшой степени, хорошо приближающую ошибку приведённой выше формулы. К сожалению, вместо простейших дробей, на которые рациональные функции двух переменных разлагаются лишь в исключительных случаях, для начала придётся использовать общий вид частного двух полиномов второй степени от двух неизвестных с 17 неизвестными коэффициентами.

Имеются однако и основания сомневаться в асимптотической точности формулы (2). Хотя её погрешность сравнима с погрешностями экспериментально определённых значений $E(m, n, \alpha)$, но формула де Монвеля в разы точнее и это настораживает. Причиной может являться скачок второй производной аппроксимирующего сплайна (2), которого нет причин ожидать у асимптотической предельной функции $f_\alpha(\frac{p}{q}) = \lim_{n \rightarrow \infty} E(pn, qn, \alpha)$. График на рис. 6 экспериментальных данных в окрестности места, где расхождение с формулой наиболее велико (а это стык отрезка прямой и дуги эллипса) выглядит более похожим на фрагмент гиперболы, чем на стык прямой и дуги

Рис. 6. График $E(n, 65536 - n, 4)$ в окрестности стыка отрезка и дуги

(кривизна дуга эллипса максимальна в месте стыка). Это наглядно свидетельствует об отсутствии скачка второй производной.

Выводы

- Предложен вид приближённой формулы для длины длиннейшей общей подпоследовательности двух случайных последовательностей не обязательно одинаковых длин.
- Подобраны вручную коэффициенты, с которыми относительная погрешность составила доли процента в диапазонах длин от тысячи до ста тысяч символов и алфавитов до 128 букв.
- Указан способ уточнения формулы с расширением диапазона применимости.
- Выявлена невозможность существенного повышения асимптотической точности формул, которые могут быть получены на этом пути.

Список литературы

- [1] V. Chvátal, D. Sankoff. “Longest Common subsequences of two random sequences”, *J. Appl. Probability*, **12**:2 (1975), pp. 306–315. ³⁴⁷
- [2] R. Bundschuh. “High precision simulations of the longest common subsequence problem”, *The European Physical Journal B-Condensed Matter and Complex Systems*, **22**:4 (2001), pp. 533–541. ³⁴⁸

- [3] R. Baeza-Yates, G. Navarro, R. Gavaldá, R. Schehing. “Bounding the expected length of the longest common subsequences and forests”, *Theory of Computing Systems*, **32**:4 (1999), pp. 435–452. ↑ ³⁴⁸
- [4] J. Boutet de Monvel. “Extensive simulations for longest common subsequences”, *The European Physical Journal B - Condensed Matter and Complex Systems*, **7**:2 (1999), pp. 293–308. ↑ ³⁴⁸
- [5] G. S. Lueker. “Improved bounds on the average length of longest common subsequences”, *Journal of the ACM*, **56**:3 (2009), pp. 17. ↑ ³⁴⁸
- [6] M. A. Kiwi, M. Loebli, J. Matousek. “Expected length of the longest common subsequence for large alphabets”, *Advances in Mathematics*, **197**:2 (2005), pp. 480–498. ↑ ³⁴⁸
- [7] J. D. Dixon. *Longest common subsequences in binary sequences*, 2013, arXiv: [1307.2796](https://arxiv.org/abs/1307.2796). ↑ ³⁴⁸
- [8] С. В. Знаменский. «Картина наибольшей длины общих подпоследовательностей пары случайных строк 4-буквенного алфавита», *Программные системы: теория и приложения*, **7**:1(28) (2016), с. 201–208 (англ.), URL: http://psta.psisras.ru/read/psta2016_1_201-208.pdf ↑ ^{348,354}
- [9] S. V. Znamenskij. “A Formula for the Mean Length of the Longest Common Subsequence”, *Journal of Siberian Federal University. Mathematics & Physics*, **10**:1 (2017), pp. 71–74. ↑ ^{348,351}
- [10] J. Lember, H. Matzinger, *The Annals of Probability*, **37**:3 (2009), pp. 1192–1235. ↑ ³⁵⁰

Рекомендовал к публикации

к.т.н. Е. П. Куршев

Пример ссылки на эту публикацию:

С. В. Знаменский. «Приближение длины наибольшей общей подпоследовательности пары случайных строк», *Программные системы: теория и приложения*, 2016, **7**:4(31), с. 347–358.

URL: http://psta.psisras.ru/read/psta2016_4_347-358.pdf

Об авторе:

Foto by A. Yu. Fomenko, CC-BY-SA

Сергей Витальевич Знаменский

Автор критерия разрешимости уравнений свёртки в пространстве функций, голоморфных на множестве, понятий выпуклости в направлении и $\mathbb{C}$ -выпуклости, русификации $\text{T}_{\text{E}}\text{X}$ для журналов Отделения математики РАН, графического пакета `mfpic3d`, ретроспективного подхода к построению информационных систем и критерия NCS близости символьных строк.

e-mail:

svz@latex.pereslavl.ru

Sergej Znamenskij. *Approximation of the longest common subsequence length for two long random strings.*

ABSTRACT. The expected value E of the longest common subsequence of letters in two random words is considered as a function of the $\alpha = A$ of alphabet and of words lengths m and n . It is assumed that each letter independently appears at any position with equal probability.

An approximate analitic expression for $E(\alpha, m, n)$ calculation is presented that allow to calculate the $E(m, n, \alpha)$ with an accuracy of 0.3 percent for $64 \leq m+n \leq 65$, 536 and $1 < \alpha < 129$. The coefficients are selected by hand and can be refined. It is expected that the formula holds for each grater values of the argument with the same relative error.

Key words and phrases: similarity of strings, sequence alignment, edit distance, LCS, Levenshtein metric.

2010 *Mathematics Subject Classification:* 68T37; 68P10, 68W32.

References

- [1] V. Chvátal, D. Sankoff. “Longest Common subsequences of two random sequences”, *J. Appl. Probability*, **12**:2 (1975), pp. 306–315.
- [2] R. Bundschuh. “High precision simulations of the longest common subsequence problem”, *The European Physical Journal B-Condensed Matter and Complex Systems*, **22**:4 (2001), pp. 533–541.
- [3] R. Baeza-Yates, G. Navarro, R. Gavaldá, R. Schehing. “Bounding the expected length of the longest common subsequences and forests”, *Theory of Computing Systems*, **32**:4 (1999), pp. 435–452.
- [4] J. Boutet de Monvel. “Extensive simulations for longest common subsequences”, *The European Physical Journal B - Condensed Matter and Complex Systems*, **7**:2 (1999), pp. 293–308.

- [5] G.S. Lueker. “Improved bounds on the average length of longest common subsequences”, *Journal of the ACM*, **56**:3 (2009), pp. 17.
- [6] M. A. Kiwi, M. Loeb, J. Matousek. “Expected length of the longest common subsequence for large alphabets”, *Advances in Mathematics*, **197**:2 (2005), pp. 480–498.
- [7] J. D. Dixon. *Longest common subsequences in binary sequences*, 2013, arXiv: [1307.2796](https://arxiv.org/abs/1307.2796).
- [8] S. V. Znamenskij. “A picture of common subsequence length for two random strings over an alphabet of 4 symbols”, *Program systems: Theory and applications*, **7**:1(28) (2016), pp. 201–208.
- [9] S. V. Znamenskij. “A Formula for the Mean Length of the Longest Common Subsequence”, *Journal of Siberian Federal University. Mathematics & Physics*, **10**:1 (2017), pp. 71–74.
- [10] J. Lember, H. Matzinger, *The Annals of Probability*, **37**:3 (2009), pp. 1192–1235.

Sample citation of this publication:

Sergej Znamenskij. “Approximation of the longest common subsequence length for two long random strings”, *Program systems: Theory and applications*, 2016, **7**:4(31), pp. 347–358. (In Russian).

URL: http://psta.psiras.ru/read/psta2016_4_347-358.pdf