Изучение периодических структур в произвольных символьных последовательностях на грид-системах из персональных компьютеров

Колпаков Р. М.[1] (foroman@mail.ru), Посыпкин М. А.[2] (mposypkin@gmail.com), Храпов Н. П.[2] (nkhrapov@gmail.com).

[1] Франко-Российский Институт Информатики и Прикладной Математики.

[2] Институт Проблем Передачи Информации РАН.

Исследование периодических структур в произвольных символьных последовательностях является важным направлением в дискретной математике. Результаты исследований используются при создании алгоритмов сжатия данных и систем поиска по шаблону. В последние годы задача периодических структур получила широкое применение также и в биологии при анализе ДНК-последовательностей. В данной статье изложена постановка одной из задач символьных последовательностей и приведены результаты, полученные к настоящему времени на основе грид-систем из персональных компьютеров.

Пусть $w = w[1]w[2]\ldots w[n]$ - произвольное слово. Длина n слова w обозначается через $|w|$. Фрагмент $w[i]\ldots w[j]$ слова w, где $1 \le i \le j \le n$, называется *фактором* слова w и обозначается через $w[j..j]$. Положительное число p называется *периодом* слова w если $w[i] = w[i+p]$ для любого $i = 1, \ldots, n - p$. Отметим, что любое слово имеет тривиальный период, равный длине слова, но при этом оно может также иметь более короткие периоды. Из всех периодов слова w можно выбрать минимальный период, который обозначается через p(w). Отношение $|w|/p(w)$ называется *порядком* слова w и обозначается через $e(w)$. Слово называется *примитивным*, если его порядок не является целым числом большим, чем 1.

Под периодичностью в слове понимается любой фактор, порядок которого не меньше, чем 2. Периодичности играют фундаментальную роль как в словарной комбинаторике [1], так и в различных приложениях, таких как алгоритмы на строках [2, 3], молекулярная биология [4] или сжатие текстовых данных [5]. Простейшим и наиболее изученным примером периодичностей являются факторы вида *uu*, где u – некоторое непустое слово. Такие периодичности называются *квадратами*. Факторы вида *uuu*, где u – некоторое непустое слово, называются *кубами*. Кубы и квадраты представляют собой частные случаи факторов вида $u^k = uu\ldots.u$ (k подряд идущий непустых одинаковых слов) , где $k > 1$ и u — непустое слово. Такой фактор называется k-ой *степенью* слова u. Если слово u является примитивным, то называется *примитивной* k-ой степенью. Если слово w имеет целый порядок $k > 1$, то оно, очевидно, является примитивной k-ой степенью своего префикса длины $p(w)$. С другой стороны, нетрудно показать (см., например, [6]), что при $k > 1$ порядок примитивной k-ой степени равен k. Таким образом, слово имеет целый порядок $k > 1$ тогда и только тогда, когда оно является примитивной k-ой степенью. Наряду с целыми степенями слов можно также рассмотреть “дробные” периодичности вида $u^k u^j$, где $k \ge 2$, u – примитивное непустое слово и u^j – некоторый префикс слова u, отличный от u. Такие периодичности имеют дробный порядок.

Периодичность в слове называется *максимальной*, если она не содержится внутри некоторой более длинной периодичности с тем же минимальным периодом, т.е. периодичность в слове является максимальной, если она не может быть расширена в этом слове ни на один символ ни вправо, ни влево с сохранением своего минимального периода. Более строго, периодичность $w[i..j]$ с минимальным периодом p в слове w называется максимальной, если она удовлетворяет следующим условиям:

если $i > 1$, то $w[i-1] \neq w[i-1+p]$,

если $j < n$, то $w[j+1-p] \neq w[j+1]$.

Нетрудно убедиться (см., например, [6]), что любая периодичность содержится в некоторой единственной максимальной периодичности с тем же минимальным периодом, поэтому максимальные периодичности естественным образом задают в слове все остальные периодичности (квадраты, кубы, k-е степени и т.д.), что позволяет их использовать в многочисленных приложениях (см., например, [7]). В [8] доказано, что максимальное возможное число максимальных периодичностей в словах длины n равно $O(n)$. Более того, было показано, что максимальная возможная сумма порядков максимальных периодичностей в словах длины n равна $O(n)$.

Понятие периодичности нетрудно обобщить на случай факторов, порядок которых не меньше, чем $1 + \varepsilon$, где $0 < \varepsilon < 1$. Мы будем называть такие факторы ε-квазипериодичностями. Заметим, что понятие максимальности непосредственным образом переносится на случай ε-квазипериодичностей: определение максимальной ε-квазипериодичности фактически дословно повторяет определение максимальной периодичности. Рассматриваемая нами задача состоит в оценке максимального возможного числа $mrn(n, \varepsilon)$ максимальных ε-квазипериодичностей в словах длины n для произвольных значений n и ε. Можно показать, что $mrn(n, \varepsilon) = O(n/\varepsilon)$ данная оценка для $mrn(n, \varepsilon)$ точной по порядку.

Для получения достаточно точной оценки максимума исследуемой величины требуется большой объем вычислений, выполнение которых заняло бы на одном персональном компьютере неприемлемо длительное время. В таких случаях целесообразно применение методов параллельных и распределенных вычислений[9-11]

Для поиска значений $mrn(n, \varepsilon)$ была разработана программа smallexp, анализирующая символьные последовательности. Последовательности создавались в коде программы специально разработанным генератором случайных чисел. Расчеты проводились на инфраструктуре BOINC в рамках проекта OPTIMA@HOME [9]. Особенность работы приложения smallexp на инфраструктуре BOINC состоит в том, что все вычислительные узлы получали задания с одинаковыми входными данными; в процессе работы на различных узлах генерировались различные случайные символьные последовательности для анализа. В общей сложности было обработано 2.5×10^{13} случайных двоичных слов. В результате предварительных расчётов были получены точные нижние грани $mrn(n,(k+1)/k)$ для случая $n = 120$ и $k = 1, 2, 3, 4, 5$. Полученные результаты позволяют сделать предположение, что $mrn(n, \varepsilon) = O(n/\varepsilon)$. Однако полной уверенности в данной гипотезе на данный момент быть не может, поскольку вызывает сомнения точность полученных значений. Полученный результат является важным приближением к решению поставленной задачи. Сама задача нуждается в дальнейших

исследованиях на основе более эффективных алгоритмов и большего набора исследуемых значений.

Список литературы

1. M. Lothaire, Combinatorics on Words, volume 17 of Encyclopedia of Mathematics and Its Applications, Addison Wesley, 1983.
2. Z. Galil, J. Seiferas, Time-space optimal string matching, J. Comput. System Sci. 26(3) (1983), 280–294.
3. M. Crochemore, W. Rytter, Squares, cubes, and time-space efficient string searching, Algorithmica 13 (1995), 405–425.
4. D. Gusfield, Algorithms on Strings, Trees, and Sequences, Cambridge University Press, 1997.
5. J. Storer, Data compression: methods and theory, Computer Science Press, Rockville, MD, 1988.
6. R. Kolpakov, G. Kucherov, Periodic structures in words, chapter for the 3rd Lothaire volume Applied Combinatorics on Words, Cambridge University Press, 2005.
7. M. Crochemore, C. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, T. Walen, Extracting powers and periods in a string from its runs structure, Lecture Notes in Comput. Sci. 6393 (2010), 258–269.
8. R. Kolpakov, G. Kucherov, On Maximal Repetitions in Words, J. Discrete Algorithms 1(1) (2000), 159–186.
9. О.С. Заикин, М.А. Посыпкин, А.А. Семѐнов, Н.П. Храпов. Опыт организации добровольных вычислений на примере проектов OPTIMA@HOME и SAT@HOME.Вестник Нижегородского университета им. Н.И. Лобачевского, 2012, No5(2), с.340-347.
10. Р. Ловаш, А.П. Афанасьев, В.В. Волошинов, М.А. Посыпкин, О.В. Сухорослов, Н.П. Храпов. О грид-системах из персональных компьютеров и их интеграции с другими видами грид-систем. Сборник избранных трудов V Международной научно-практической конференции "Современные информационные технологии и ИТ-образование". Москва: ИНТУИТ.РУ 2010 г.
11. О.В. Сухорослов. Реализация и композиция проблемно-ориентированных сервисов в среде MathCloud // Вестник ЮУрГУ. Серия «Математическое моделирование и программирование», Вып. 8, № 17(234). – Челябинск: Издательский центр ЮУрГУ, 2011. (с. 101-112)