2008 Прикладная теория графов № 1(1)

DOI 10.17223/20710410/1/16

УДК 512.1, 519.6, 519.7

ИНСТРУМЕНТАРИЙ ГРАФИЧЕСКОГО ИССЛЕДОВАНИЯ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ[1]

А.А. Евдокимов, А.А. Левин

Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск

E-mail: evdok@math.nsc.ru, levin@math.nsc.ru

Разработан пакет «BruijnViz» для исследования свойств символьных последовательностей, или слов большой длины. Все подслова длины *n* отображаются на граф перекрытия слов (граф де Брёйна), образуя граф-портреты в процессе роста длины последовательности. Реализованы различные способы изображения графа на плоскости экрана. Приводятся примеры граф-портретов последовательностей, возникающих в приложениях, и анализируются их свойства.

Ключевые слова: *граф де Брёйна*, *граф подслов*, *символьная последовательность*, *сложность*, *визуализация графа.*

Вначале об идее подхода к задаче визуализации символьных последовательностей.

При исследовании символьных последовательностей эффективным оказывается анализ связи их свойств со свойствами структур, образуемых множеством фрагментов (подслов) этой последовательности. Выявление таких структур и исследование динамики их изменения при увеличении длины фрагментов и самой последовательности дает ценную информацию о ее свойствах. Возможности изучения свойств последовательностей зависят от «хорошего» изображения их графов перекрытия подслов на плоскости (на экране компьютера).

Графы перекрытия слов, введенные де Брёйном в 1946 г. и теперь называемые его именем [1], оказываются удобными для изображения последовательностей и изучения их структурных свойств. Изображение графа подслов последовательности на графе де Брёйна мы называем граф-портретом этой последовательности. Рассмотрение динамики последовательности граф-портретов помогает исследовать свойства как отдельных последовательностей, так и их классов.

Прикладная направленность исследования граф-портретов последовательностей состоит в расширении методов и инструментария для анализа структуры последовательностей как естественного происхождения, например генетических [3, 4], так и математических – порождаемых рекурсивными процедурами различного типа. Анализ граф-портретов символьных последовательностей выявил ряд их интересных свойств, связанных с особенностями структуры множества подслов. В частности, с помощью пакета «BruijnViz» были исследованы последовательность, введенная Евдокимовым для решения известной проблемы «змея в ящике», и последовательность непрерывного кодирования, близкая к последовательности, известной в англоязычной литературе как последовательность «Look and say» [5].

Наиболее интересные полученные граф-портреты демонстрируются в работе, в частности последовательность непрерывного кодирования имеет сложные граф-портреты и большую комбинаторную сложность (количество различных её подслов).

Задача построения граф-портретов последовательностей приводит к задачам поиска вложений графов, сохраняющих определенные структурные свойства вкладываемых объектов: метрические, алгебраические или комбинаторные [6 – 8], в частности построения таких вложений графов на плоскость, которые сохраняют отношение близости между вершинами, а расстояния между далекими вершинами оставляют больше некоторого заданного порога [7]. Идея такого типа вложений для задач визуализации граф-портретов символьных последовательностей реализована в работе [9].

1. Определения

Теперь определим основные понятия. Вершинами графа де Брёйна B_m^n размерности *n* являются всевозможные слова длины *n* в алфавите из *m* букв. Две вершины $\tilde{\alpha} = (\alpha_1, ..., \alpha_n)$ и $\tilde{\beta} = (\beta_1, ..., \beta_n)$ соединены дугой,

[1] Исследование выполнено при финансовой поддержке РФФИ (проект 08-01-00671) и программы Отделения математических наук РАН «Алгебраические и комбинаторные методы математической кибернетики» (проект «Новые методы дискретного анализа и комбинаторной оптимизации»).

ориентированной от $\tilde{\alpha}$ к $\tilde{\beta}$, тогда и только тогда, когда $\alpha_2 = \beta_1$, $\alpha_3 = \beta_2$, ... , $\alpha_n = \beta_{n-1}$, т.е. когда слова $\tilde{\alpha}$ и $\tilde{\beta}$ перекрываются по $n - 1$ буквам.

Граф B_m^n имеет m петель в вершинах, соответствующих словам – константам, состоящим из одной буквы алфавита. Он связен, однороден, полустепень входа и выхода каждой его вершины равна m. При изображении последовательности графов B_m^n на плоскости для $n = 1, 2, 3, \ldots$ можно использовать процедуру их построения индукцией по размерности n, основанную на том, что граф B_m^{n+1} является рёберным графом для B_m^n.

Произвольной (бесконечной или конечной длины $\geq n$) последовательности $\tilde{x} = x_1, x_2, x_3, \ldots$ букв m-алфавита сопоставляется путь в графе B_m^n, который начинается в вершине $(x_1, \ldots, x_n)$ и последовательно проходит вершины $(x_i, \ldots, x_{i+n-1})$ при $i = 2, 3, \ldots$ Заметаемый этим путем подграф графа B_m^n называется графом n-подслов последовательности $\tilde{x}$ или граф-портретом размерности n и обозначается $G^n(\tilde{x})$. Таким образом, множеством вершин $V^n(\tilde{x})$ графа $G^n(\tilde{x})$ является множество всех подслов длины n в $\tilde{x}$, а множеством дуг $E^n(\tilde{x})$ – множество всех подслов длины $n + 1$ в $\tilde{x}$ [6]. Изображение графа-дополнения $B_m^n \setminus G^n(\tilde{x})$ позволяет наблюдать структуру множества отсутствующих n-подслов в последовательности $\tilde{x}$. Для построения последовательности граф-портретов $\{G^i(\tilde{x})\}$, $i = 1, 2, 3, \ldots$, при росте их размерности $i \to i + 1$ используется операция построения рёберного графа, поскольку $G^{i+1}(\tilde{x})$ является подграфом графа рёберного для $G^i(\tilde{x})$.

2. Пакет BruijnViz

Дадим описание пакета BruijnViz и технологии его использования.

Так как все известные пакеты изображения графов ориентированы для изображения графов определенного типа, то для изображения графов подслов и исследования различных последовательностей нами разработана специальная программа BruijnViz. Программа реализована на языке JAVA, поэтому может функционировать на любой ЭВМ с виртуальной машиной JAVA. Демонстрационная версия начальной конфигурации программы находится на сервере Института математики СО РАН (http://www.math.nsc.ru/LBRT/k3/Graph/Bruijn.htm).

Программа BruijnViz строит граф B_m^n перекрытия слов (граф де Брёйна) для заданных параметров: значности алфавита последовательности m и длины слов n (размерности графа). Затем на графе изображается исследуемая последовательность $\tilde{x}$ и её граф-портрет $G^n(\tilde{x})$. В программе можно варьировать изображение на экране граф-портретов $G^n(\tilde{x})$. Изменение параметров процесса возможно производить непосредственно в ходе наблюдения.

На экране располагаются кнопки управления и меню установки режима работы программы. В верхней части экрана выводится отрезок обрабатываемой последовательности, на котором выделено текущее слово и несколько строк текущей информации. Пользователь может изменять взаимное расположение вершин графа на экране, перемещать весь граф, а также изменять размер изображения графа или его части. Если при запуске программы не задано имя файла с начальным графом, то производится построение полного графа B_m^n со случайным размещением вершин на экране. В противном случае начальный граф считывается из заданного файла и изображается на экране.

После нажатия одной из кнопок управления движением программа считывает очередной символ (в начальной точке последовательности считывается целое слово), формирует очередное слово, приписывая считанный символ в конец предыдущего слова. Вершина, имя которой совпадает с полученным словом, и ребро, соединяющее её с предыдущей вершиной, помечаются и заносятся в пройденную цепочку, а их счетчики увеличиваются на 1. Пройденная цепочка выделяется на изображении и называется «змея». Программа может непрерывно наращивать длину обработанной последовательности в широком диапазоне выбираемых скоростей. В программе реализованы режимы автоматического расталкивания близких вершин и выделения цепей на граф-портрете.

Программа позволяет изменять размерность графа де Брёйна, на котором располагается исследуемая последовательность. Увеличение размерности осуществляется построением реберного графа для всего графа

B_m^n или для части графа подслов $G^i(\tilde{x})$, пройденной последовательностью. При уменьшении размерности происходит возвращение к тому графу, из которого строился реберный. Если комбинаторная сложность последовательности растет медленно, то наблюдать граф-портреты можно для больших значений их размерности *n*, что существенно помогает при анализе свойств.

ЛИТЕРАТУРА

1. *De Bruijn N.G.* A combinatorial problem // Nederl. Akad. Wetensch. Proc. 1946. V. 49. No. 7. P. 758 – 764. (Перевод см. Кибернетический сборник, новая серия, вып. 6. М.: Мир, 1969. С. 33 – 40.)
2. *Математические* методы для анализа последовательностей ДНК: Пер. с англ. / Под ред. М.С. Уотермена. М.: Мир, 1999. 349 с.
3. *Evdokimov A.A.*, *Levin A.A.* Subwords graphs, generated by genetic sequences // Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure – BGRS' 2002. V. 1, IC&G. Novosibirsk, 2002. P. 131 – 133.
4. *Евдокимов А.А.*, *Левин А.А.* Графические модели и комбинаторика генетических и математических символьных последовательностей // Вычислительные технологии. 2002. Т. 7. С. 274 – 278.
5. *Евдокимов А.А.*, *Левин А.А.* Теоретическое и экспериментальное исследование рекурсивно порожденных символьных последовательностей // Вестник ТГУ. Приложение. 2007. № 23. С. 16 – 23.
6. *Евдокимов А.А.* Исследование полноты множеств слов и языков с запретами // Вестник ТГУ. Приложение. 2004. № 9(1). С. 8 – 12.
7. *Евдокимов А.А.* Кодирование структурированной информации и вложения дискретных пространств // Дискрет. анализ и исслед. операций. Сер. 1. 2000. Т. 7. № 4. С. 48 – 58.
8. *Евдокимов А.А.* Анализ, сложность и реконструкция символьных последовательностей // Вестник ТГУ. Приложение. 2005. № 14. С. 4 – 12.
9. *Евдокимов А.А.*, *Левин А.А.* Методы визуализации графов подслов символьных последовательностей // Вычислительные технологии. 2003. Т. 8. С. 5 – 11.