МАТЕМАТИЧЕСКИЕ МЕТОДЫ СИСТЕМНОГО АНАЛИЗА И УПРАВЛЕНИЯ

УДК 519.688

МУЛЬТИЭВРИСТИЧЕСКИЙ ПОДХОД К ПРОБЛЕМЕ ЗВЕЗДНО-ВЫСОТНОЙ МИНИМИЗАЦИИ НЕДЕТЕРМИНИРОВАННЫХ КОНЕЧНЫХ АВТОМАТОВ

С. В. Баумгертнер, Б. Ф. Мельников

Тольяттинский государственный университет

Поступила в редакцию 06.04.2010 г.

Аннотация. В данной статье рассматривается задача построения регулярного выражения, оптимального с точки зрения звездной высоты, для заданного конечного автомата. Предлагается anytime-алгоритм, позволяющий получить псевдо-оптимальное решение за определенный промежуток времени.
Ключевые слова: Проблема звездное высоты, недетерминированный конечный автомат, регулярное выражение, незавершённый метод ветвей и границ, мультиэвристический подход.

Abstract. In this paper, we consider the problem of searching regular expression with minimum star height by finite automata. We formulate an anytime-algorithm, that finds pseudo-optimum solve during the given time period.
Keywords: Star-height problem, non-deterministic finite automata, regular expression, multi-heurist method.

ВВЕДЕНИЕ

В теории формальных языков звёздной высотой регулярного выражения называется мера сложности, которая равна максимальной глубине вложенности операции *. Более формально, звёздная высота регулярного выражения (обозначим её $\mathrm{sh}(r)$) определяется по индукции следующим образом [1]:

1. $\mathrm{sh}(\varnothing) = \mathrm{sh}(\varnothing^*) = \mathrm{sh}(a) = 0$ для всех $a \in \sum$.

2. Пусть r и s — произвольные регулярные выражения. Тогда $\mathrm{sh}((r+s)) = sh((r \cdot s)) = \max(\mathrm{sh}(r), \mathrm{sh}(s))$.

3. Для любого регулярного выражения r (где $r \neq \varnothing$) : $sh((r^*)) = sh(r)+1$.

Звёздной высотой регулярного языка называется минимальная из звёздных высот регулярных выражений, определяющих этот язык.

Несмотря на то, что звёздную высоту регулярного выражения вычислить достаточно просто, вычисление звёздной высоты регулярного языка обычно представляет собой очень сложную проблему. Например, звёздная высота регулярного выражения $(b + aa^*b)^*aa^*$ над алфавитом $A = \{a,b\}$ равна 2. Однако легко заметить, что данный язык является множеством всех слов над алфавитом A, заканчивающихся на a. Следовательно, данный язык может быть представлен другим регулярным выражением, $(a + b)^*a$, звёздная высота которого равна 1. Таким образом, звёздная высота этого регулярного языка равна 1.

Проблема звёздной высоты — одна из важных задач в теории формальных языков. Она оставалась нерешённой в течение примерно 25 лет, пока Хашигучи в 1988 не опубликовал алгоритм поиска звёздной высоты для любого регулярного языка [5]. Однако алгоритм был совершенно неприменим на практике. Процедура, описанная Хашигучи, ведёт к вычислениям, которые невозможны даже для маленьких примеров.

Намного эффективнее оказался алгоритм, который был разработан Кирстеном в 2005 [6]. На входе этой процедуры — недетерминированный конечный автомат. Но объёмы ресурсов, необходимые для этого алгоритма по-прежнему выходили за рамки, обоснованные практическими возможностями.

В данной статье будет рассмотрен алгоритм поиска псевдо-оптимального регулярного вы-

ражения для регулярного языка, заданного с помощью недетерминированного конечного автомата. Алгоритм основан на применении нескольких эвристик.

1. МЕТОД ПОЛНОГО ПЕРЕБОРА

Рассмотрим задачу построения регулярного выражения, оптимального с точки зрения звёздной высоты, по конечному автомату. С помощью алгоритма последовательного удаления вершин [4] или алгоритма теоремы Клини можно для данной задачи получить точный алгоритм решения. Для этого необходимо перебрать все $n!$ (n — количество вершин) перестановок вершин конечного автомата, для всех перестановок найти регулярные выражения и выбрать среди них то, которое имеет наименьшую звёздную высоту. Но на практике алгоритм полного перебора неприемлем даже для автоматов с довольно небольшим количеством вершин — из-за невозможности получить решение за разумное время. Поэтому для решения данной задачи необходимы т.н. anytime-алгоритмы [4], которые позволят получить текущее псевдо-оптимальное регулярное выражение — лучшее решение, найденное за определенный промежуток времени.

2. МУЛЬТИЭВРИСТИЧЕСКИЙ ПОДХОД — ОПИСАНИЕ ЭВРИСТИК

Ясно, что не существует эвристики, которая давала бы возможность находить достаточно хорошие решения для любого заданного автомата. Поэтому каждая из эвристик направлена на решение задачи на определенных классах автоматов, на которых она предположительно должна давать хорошие результаты. Неформально описываемые далее эвристики можно объяснить следующим образом.

Чем больше циклов проходит через данное состояние q, тем больше существует путей на графе переходов автомата, через которые проходят эти циклы. То есть при удалении вершины q ранее других мы с большей вероятностью получаем новые дуги, помеченные регулярными выражениями с большей звёздной высотой.

Эвристика 1. Состояния упорядочиваются в порядке возрастания количества проходящих через них циклов.

Чем больше количество вершин, которые соединяет между собой вершина q, тем больше будет новых пометок дуг с увеличенной звёздной высотой при её удалении.

Эвристика 2. Состояния упорядочиваются в порядке возрастания суммарного количества вершин во всех циклах, проходящих через них.

Эвристика 3. Состояния упорядочиваются в порядке возрастания суммарного количества вершин во всех циклах, проходящих через них, но при этом каждое состояние учитывается только один раз.

Эвристика 4. Оценка состояния получается умножением количества циклов, проходящих через это состояние, на значение эвристики 3.

3. НЕЗАВЕРШЁННЫЙ МЕТОД ВЕТВЕЙ И ГРАНИЦ

Для решения поставленной задачи воспользуемся незавершенным методом ветвей и границ, который получается при внесении некоторых изменений в классический метод [4].

Будем называть *правой задачей* очередного шага метода ветвей и границ задачу, полученную при уменьшении размерности. Другую альтернативу — т. е. когда принимается решение об отсутствии некоторого элемента в оптимальном решении — назовём, соответственно, *левой задачей* очередного шага. С помощью некоторых эвристик мы добиваемся того, чтобы вероятность наличия оптимального решения была больше для правой задачи, чем для левой, т. к. размерность правой задачи на 1 меньше.

Каждый раз при получении очередной правой задачи мы строим *последовательность правых задач*. Также строятся (и включаются в список задач для потенциального решения в последующем) и соответствующие левые задачи. При получении тривиальной задачи (задачи нулевой размерности) мы запоминаем её решение в качестве текущего *на данный момент времени* псевдо-оптимального решения. При получении в какой-либо задаче достаточно большой границы — например, большей, чем имеющееся на данный момент времени псевдо-оптимальное решение — мы исключаем ее из последующего решения.

При этом мы олучаем следующий алгоритм нахождения регулярного выражения:

1. Получаем некоторый вектор оценки состояний автомата, используя мультиэвристический метод и динамические функции риска (подробнее см. ниже).

2. Строим левую и правую подзадачи. Выбираем состояние автомата q с наилучшей суммарной оценкой. Если граница какой-либо задачи не меньше, чем уже существующее решение, то мы прекращаем ветвление этой задачи.

3. Переходим к решению полученной правой подзадачи. Если в правой задаче остается одна вершина, за исключением стартовой и финальной, мы удаляем её и находим решение. Если оно имеет меньшую звёздную высоту, чем уже найденное значение, то оно становится текущим псевдо-оптимальным решением.

Возврат к решению левых подзадач происходит после решения всех правых подзадач. Таким образом, отсеиваются большие множества решений, не являющихся лучшими, чем уже найденное псевдо-оптимального решение. Очевидно, что чем лучше первое псевдо-оптимальное решение, тем больше бесперспективных областей можно исключить из последующего решения. Следовательно, чем более эффективны применяемые эвристики, тем быстрее мы получим оптимальное решение.

С помощью такого подхода можно гарантированно получить псевдо-оптимальное решение за указанное время, либо даже точное решение (при наличии времени для решения всех возникающих подзадач).

4. ДИНАМИЧЕСКИЕ ФУНКЦИИ РИСКА

Итак, у нас есть данные, полученные с помощью разных эвристик. Необходимо принять решение о выборе очередной вершины автомата — для последующего её удаления. При этом информация, данная различными эвристиками, может быть противоречива. В связи с этим появляется необходимость в некоторых методах усреднения значений эвристик и получения итоговой перестановки вершин автомата с помощью усредненных оценок. Такие оценки вычисляются с помощью набора коэффициентов [2—4]. Заметим, что данная ситуация аналогична ситуации в недетерминированных играх: есть оценки возможных ходов, учитывая их необходимо принять решение о выборе следующего хода.

Для каждой вершины автомата вычисляется т. н. *динамическая оценка*:

$$z(x_1,\ldots,x_k)=\frac{\sum_{i=1}^{k}x_i\theta(x_i)}{\sum_{i=1}^{k}\theta(x_i)};$$

где $\theta(x_i)$ — т. н. *функция риска*.

$$\theta(x)=\begin{cases}0,8(1-x)^2+0,2, & \text{если } x\approx 1\\ -0,2x^2+1, & \text{если } x\approx 0\\ -0,8(1-x)^2+1, & \text{если } x\approx -1\end{cases}.$$

Усреднение надо делать с весами тем меньшими, чем более хорошей является оценка, данная эвристикой. В итоге получаем оценки вершин, вычисленные с применением сразу нескольких эвристик. На основе проведённых вычислительных экспериментов (которые мы подробно не описываем — в связи с ограничением на объём данной статьи) мы получаем, что в данной задаче они также будут более точными для большинства автоматов.

СПИСОК ЛИТЕРАТУРЫ

1. *Саломаа А.* Жемчужины теории формальных языков. — М.: Мир, 1986. — 159 с.

2. *Мельников Б.* Эвристики в программировании недетерминированных игр. — Известия РАН. Программирование, 2001, № 5. С. 63—80.

3. *Мельников Б., Радионов А. Н.* О выборе стратегии в недетерминированных антагонистических играх. — Известия РАН. Программирование, 1998, № 5, С. 55—62.

4. *Мельников Б.* Мультиэвристический подход к задачам дискретной оптимизации. — Кибернетика и системный анализ (НАН Украины), 2006, № 3. С. 32—42.

5. *Hashiguch K. i*: Algorithms for determining relative star height and star height. — Inform. Comput., 78 (1987) 124—169.6. *Kirsten D.* Distance desert automata and the star height problem. Theoret. Informatics Appl., 39 (2005) 455—509.

Баумгертнер Светлана Викторовна – аспирантка кафедры «Прикладная математика и прикладная информатика», Тольяттинский государственный университет. E-mail: S-Baumgertner@yandex.ru.

Мельников Борис Феликсович – д.ф.-м.н., профессор кафедры «Прикладная математика и прикладная информатика», Тольяттинский государственный университет. Тел. (8482) 53-95-14. E-mail: B.Melnikov@tltsu.ru.

Baumgertner S.V. – post-graduate student, department of Applied Mathematics and Applied Informatics, Togliatti State University. E-mail: S-Baumgertner@yandex.ru.

Melnikov B.F. – professor, department of Applied Mathematics and Applied Informatics, Togliatti State University. Tel. (8482) 53-95-14. E-mail: B.Melnikov@tltsu.ru.