УДК 681.3

Э.И. Ватутин, канд. техн. наук, доцент, кафедра вычислительной техники, ЮЗГУ (e-mail: evatutin@rambler.ru)
В.С. Титов, докт. техн. наук, доцент, зав. кафедрой вычислительной техники, ЮЗГУ (e-mail: titov-kstu@rambler.ru)

СТРАТЕГИИ ПРОВЕРКИ КОРРЕКТНОСТИ МЕТОДОВ ВЫЯВЛЕНИЯ ИЗОМОРФИЗМА ГРАФОВ С ИСПОЛЬЗОВАНИЕМ ГРИД-СИСТЕМ

Приведено описание стратегий проверки методов выявления изоморфизма графов путем организации попарного сравнения матриц смежности графов выбранной размерности и с использованием построения разбиения на классы изоморфизма, что существенно снижает требуемые вычислительные затраты. Показано, что точная проверка методов возможна только для графов небольшой размерности $(N = 10)$.

Ключевые слова: теория графов, изоморфизм графов, грид-системы

***

Одной из задач в области теории графов является задача проверки изоморфизма графов $G_1 = \langle A_1, V_1 \rangle$ и $G_2 = \langle A_2, V_2 \rangle$, где A – множество вершин, V – множество дуг, с целью выяснения факта наличия/отсутствия изоморфизма либо отыскания подстановки изоморфизма F, являющейся биекцией между множествами вершин A_1 и A_2. Для проверки изоморфизма графов частного вида известен ряд быстрых полиномиальных алгоритмов (например, [1]), однако в то же время неизвестен полиномиальный алгоритм для проверки изоморфизма графов общего вида, но и не доказана невозможность его построения. На протяжении последних десятилетий предпринимались неоднократные попытки его отыскания, однако для известных подходов либо известны контрпримеры, либо не существует общепризнанного аналитического доказательства [2, 3].

Проверку корректности интересующего алгоритма можно провести в ходе вычислительного эксперимента, при котором инвариант проверяемого метода сопоставляется с предположительно полным инвариантом, образованным совокупностью известных быстро вычислимых инвариантов. При несовпадении результатов сравнения предположительно полных инвариантов и проверяемого инварианта может быть найден искомый контрпример, опровергающий корректность проверяемого метода. Если же подобных несовпадений не выявлено, например, при проверке всех графов с числом вершин $N = |A|$ меньше заданного $N_{\max}$, можно утверждать, что метод корректно работает в области $N < N_{\max}$. Ввиду того, что взятый за эталон для сравнения предположительно полный инвариант не является на самом деле полным, возможны ситуации, когда несовпадение результатов проверки на изоморфизм вызвано низкой дифференцирующей способностью выбранного предположительно полного инварианта, а не отысканием контрпримера для тестируемого метода. В таком случае необходимо либо доопределение предположительно полного инварианта дополнительными частными инвариантами (что невозможно в общем случае, т.к. иначе данный предположительно полный инвариант был бы по настоящему полным и быстровычислимым, а таковой на данный момент неизвестен), либо автоматизированная ручная проверка полученных результатов.

В качестве компонентов предположительно полного инварианта были взяты следующие быстро вычислимые инварианты:

- число вершин графа N (на самом деле оно совпадает в рамках проводимых вычислительных экспериментов);
- число дуг графа $M = |V|$;
- отсортированный вектор степеней вершин $P = sort\left[d(a_1), d(a_2), ..., d(a_N)\right]$, где $d(a_i)$ – степень вершины, т.е. число инцидентных ей дуг;
- детерминант матрицы смежности графа $D = \det M$, где $m_{ij} = 0$, если вершины a_i и a_j не связаны дугой, и $m_{ij} = 1$ в противном случае, $i, j = \overline{1, N}$;
- индекс Рандича [4] $r = \sum\limits_{\forall(a_i, a_j)\in V} \frac{1}{\sqrt{d(a_i)d(a_j)}}$ и две его модификации [5], повышающие дифференциальную способность:

$$r' = \sum_{\substack{\forall((a_i, a_j)\in V)\vee \\ \vee((a_i, a_k)\in V)\vee \\ \vee((a_j, a_k)\in V)}} \frac{1}{\sqrt{d(a_i)d(a_j)d(a_k)}},$$

$$r'' = \sum_{\substack{\forall(a_i, a_j, a_k), \\ i\neq j, i\neq k}} \frac{1}{\sqrt{d(a_i)d(a_j)m_{ij} + d(a_i)d(a_k)m_{ik} + d(a_j)d(a_k)m_{jk}}};$$

- число компонент связности графа N_C;
- вектор мощностей компонент связности графа $V = [v_1, v_2, ..., v_N]$, в котором значение v_i показывает, что компонента связности из i вершин встретилась v_i раз;
- диаметр графа, определяемый как минимальное число дуг $l(a_i, a_j)$ между максимально удаленной парой вершин: $d(G) = \max\limits_{i, j=\overline{1, N}} l(a_i, a_j)$.

Для графов с числом вершин $N \leq 7$ указанный предположительно полный инвариант является полным, т.к. обеспечивает формирование числа классов изоморфизма, в точности равного известному (последовательность A000088 по классификатору целочисленных последовательностей [6]), для $N = 8$ формируемое число классов изоморфизма равно 12319 вместо искомого 12346, что говорит о наличии 27 дополнительных не выявленных классов изоморфизма.

В качестве метода, претендующего на построение полного инварианта, был взят метод, предложенный в работе [2, 3]. С целью поиска контрпримера осуществлялось попарное сравнение инвариантов для выбранной пары матриц смежности графов M_i и M_j, где нижний индекс обозначает код матрицы, полученный путем построчного выписывания двоичных значений верхней треугольной подматрицы и перевода полученной битовой строки в десятичную форму:

$$M_{13} = \begin{pmatrix} - & 1^{(0)} & 0^{(1)} & 1^{(2)} \\ 1 & - & 1^{(3)} & 0^{(4)} \\ 0 & 1 & - & 0^{(5)} \\ 1 & 0 & 0 & - \end{pmatrix} \quad 001101_2 = 13_{10}.$$

Несложно показать, что для графов из N вершин существует $n = 2^{((N-1)+(N-2)+...+1)} = 2^{\frac{N(N-1)}{2}}$ различных матриц смежности, а необходимое число попарных сопоставлений инварианта составляет $x = (n-1)+(n-2)+...+1 = \frac{n(n-1)}{2} = \frac{2^{\frac{N(N-1)}{2}}\left(2^{\frac{N(N-1)}{2}}-1\right)}{2}$. Так для графов из $N = 5$ вершин число различных матриц смежности $n = 1024$, необходимое число сопоставлений $x = 523\,776$. При реализации подобной стратегии «в лоб» [7] время вычислительного эксперимента составило 6 часов на компьютере с процессором Core 2 Duo 1,86 ГГц для однопоточной реализации расчетного модуля. Для $N = 6$ число матриц смежности $n = 32\,768$, необходимое число попарных сопоставлений $x = 536\,854\,528$, на которые потребовалось бы 6 150 часов (256 дней) вычислительного времени. При дальнейшем увеличении N необходимые затраты стремительно растут: при увеличении числа вершин в графе на 1 число различных матриц смежности возрастает в 2^N раз, а необходимое число попарных сопоставлений – приблизительно в 2^{2N} раз (т.е. для $N = 7$ затраты вычислительного времени будут еще в $2^{14} = 16\,384$ раз больше, для $N = 8$ – еще в 65 536 больше и т.д.). Наблюдаемый «комбинаторный взрыв» не позволяет реализовать на практике данную стратегию для графов с большим числом вершин.

Вычислительную сложность проверки можно существенно снизить путем разбиения множества всех матриц смежности на классы эквивалентности [8]. Так, например, для $N = 4$ имеют место $C = 11$ классов изоморфизма:

$$c_1 = \{0\},\ c_2 = \{1, 2, 4, 8, 16, 32\},\ c_3 = \{3, 5, 6, 9, 10, 17, 20, 24, 34, 36, 40, 48\},$$
$$c_4 = \{7, 25, 42, 52\},\ c_5 = \{12, 18, 33\},\ c_6 = \{11, 21, 38, 56\},$$
$$c_7 = \{13, 14, 19, 22, 26, 28, 35, 37, 41, 44, 49, 50\},$$
$$c_8 = \{15, 23, 27, 29, 39, 43, 46, 53, 54, 57, 58, 60\},\ c_9 = \{30, 45, 51\},$$
$$c_{10} = \{31, 47, 55, 59, 61, 62\},\ c_{11} = \{63\},$$

где числа в составе множеств обозначают коды матриц смежности. При наличии подобного множества проверку метода, претендующего на построение полного инварианта, можно организовать в два этапа:

- путем сопоставления на равенство инвариантов для всех графов в рамках каждого класса эквивалентности c_i, что потребует $\sum_{i=1}^{C} \frac{|c_i|(|c_i|-1)}{2} < NC$ сопоставлений;
- путем сопоставления на неравенство инвариантов для одного графа из каждого класса изоморфизма, что потребует $\frac{C(C-1)}{2}$ сопоставлений.

Общее число сопоставлений в данном случае составляет величину порядка $O(NC)$, что ощутимо лучше полученной выше экспоненциальной зависимости $O\left(2^{N^2}\right)$. С использованием данной стратегии для $N \le 7$ были построены множества классов изоморфизма и совпадения инвариантов для графов из различных классов выявлено не было (при использовании подхода [4] несовпадение инварианта в рамках одного класса изоморфизма представляется невозможным), для чего потребовалось более 2 часов вычислительного времени.

При переходе к графам с большим числом вершин затраты вычислительного времени существенно возрастают. Так при $N = 8$ число классов изоморфизма, построение которого также становится вычислительно сложной задачей, составляет уже $C = 12\,346$, что при попарных сопоставлениях инвариантов потребует приблизительно $7{,}6 \cdot 10^7$ сопоставлений, на что потребуется не менее 12 дней; при $N = 9$ $C = 274\,668$, число сопоставлений $3{,}8 \cdot 10^{10}$, необходимые затраты вычислительного времени – не менее 16 лет вычислительного времени и т.д. Таким образом, данная стратегия позволяет расширить область проверки до $N \le 9 \div 10$ с использованием параллельных вычислительных средств (кластеров, суперкомпьютеров или грид-систем), но не более того.

При рассмотрении графов с числом вершин $N > 10$ точная проверка корректности представляется практически невозможной и она может быть заменена вероятностным сравнением для большого числа граф-схем (по аналогии, например, с тестом Миллера–Рабина [9], используемым для вероятностной проверки чисел на простоту).

Наиболее перспективным классом вычислительных средств для подобного вычислительного эксперимента представляются грид-системы на добровольной основе, крупнейшая из которых построена на платформе BOINC [10] и характеризуется интегральной производительностью 8,5 PFLOPS. При этом на удаленные компьютеры необходима передача данных о множестве кодов матриц смежности для поиска попарных совпадений инварианта (принадлежащих к разным классам эквивалентности при $N \le 10$ или сгенерированных случайно при $N > 10$), искомым результатом ответа клиентской машины может служить факт нахождения пары неизоморфных графов, распознаваемых тестируемым методом как изоморфные (контрпример). Подобные расчеты могут быть реализованы в рамках проекта распределенных вычислений Gerasim@Home [11].

Список литературы

1. Ватутин Э.И., Зотов И.В., Титов В.С. Выявление изоморфных вхождений *R*-выражений при построении множества сечений параллельных алгоритмов логического управления // Информационно-измерительные и управляющие системы. № 11, Т. 7. М.: «Радиотехника», 2009. С. 49–56.
2. Валяев В.В., Ватутин Э.И. Метод определения изоморфизма графов общего вида за полиномиальное время // Известия ЮЗГУ. Серия «Управление, вычислительная техника, информатика. Медицинское приборостроение». № 2. Ч. 1. 2012. С. 200–206.
3. Ватутин Э.И., Валяев В.В. Программа для построения предположительно полного полиномиального инварианта графа и проверки пары графов на изоморфизм // Свидетельство о государственной регистрации программы для ЭВМ № 2013618656 от 13.09.13.
4. Randić, M. Characterization of molecular branching // Journal of the American Chemical Society T. 97 (23): 6609–6615, 1975.
5. Ватутин Э.И. Повышение дифференцирующей способности индекса Рандича // Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации (Распознавание – 2013). Курск, изд-во ЮЗГУ, 2013. С. 237–240.
6. http://oeis.org/A000088
7. Ватутин Э.И. Программа для тестирования корректности методов проверки изоморфизма графов // Свидетельство о государственной регистрации программы для ЭВМ № 2013619188 от 26.09.13.
8. Ватутин Э.И. Программа для построения классов изоморфизма графов и тестирования корректности методов проверки изоморфизма графов на его основе // Свидетельство о государственной регистрации программы для ЭВМ № 2013619186 от 26.09.13.
9. Rabin M.O. Probabilistic algorithm for testing primality // Journal of Number Theory. Elsevier Inc., 1980. B. 1. T. 12. PP. 128–138. ISSN 0022-314X. DOI:10.1016/0022-314X(80)90084-0
10. http://boinc.berkeley.edu
11. http://gerasim.boinc.ru