Артём Рипатти @ripatti
Математик-программист
92,0 карма 36,1 рейтинг

сегодня в 07:14

Разработка → Судоку: так сколько же их? Часть 2/2 перевод

Программирование*, Математика*, Алгоритмы*

Привет Хабр! Это вторая часть перевода статьи про подсчет различных судоку.

В этой части мы погрузимся в теорию групп, начиная с самых основ, но затрагивая только то, что нам пригодится для ответа на вопрос: а сколько же есть действительно различных судоку — без всяких повторов в виде поворотов, отражений и т.п. Те, кто довольно хорошо знаком с теорией групп — вероятно, найдут тут мало что интересного. Для остальных же почитать очень даже полезно. На всякий случай: я себя специалистом по теории групп не считаю, при переводе статьи я сам по сути изучал ее почти с нуля. То есть, вполне могут быть косяки — пишите мне о них в личку. С другой стороны — я для большинства определений лазил в википедию, а все численные результаты подтвердил собственноручно написанной программой. Так что, по идее, количество косяков должно стремиться к нулю. Но мало ли.

Как обычно, мои комментарии выделены курсивом или спрятаны под спойлеры. Под спойлерами можно найти самое интересное — куски кода, которые верифицируют все числа, полученные в повествовании.

Симметрии

Итак, в предыдущей части статьи мы посчитали, что количество различных сеток судоку равно

$N=6670903752021072936960\approx 6.671\times 10^{21}$. Но довольно часто мы можем получить одну сетку судоку из другой, применяя простые преобразования. Например, если у нас есть корректно заполненная сетку судоку, то, поворачивая ее на 90°, мы получаем другую сетку, которая отличается от исходной, но все еще остается корректной. На самом деле, мы можем рассматривать эти сетки как одинаковые, поскольку в результате преобразования наша сетка — все еще одно из решений судоку. Аналогично, если мы заменим все 3-ки в сетке на 4-ки, а все 4-ки на 3-ки, то мы получим другую корректную сетку. Мы также могли бы взять какую-нибудь сетку судоку, поменять местами пятую и шестую строки, и, в конце концов, все равно получить корректную сетку. Когда мы делаем такие преобразования, мы сохраняет такое свойство сетки, как ее корректность (*В предыдущей части статьи мы делали что-то очень похожее, но только с частью сетки — первой полосой. Теперь же мы играемся со всей таблицей 9×9. Важное отличие в том, что сейчас мы рассматриваем только такие преобразования, которые можно применить к любой сетке (с сохранением корректности), т.е. рассмотренная ранее операция свапания углов прямоугольника внутри сетки судоку не подходит, так как к некоторым сеткам она применима, а к некоторым — нет.*).

Операции такого вида называются симметриями. Симметрия объекта — это операция, которая сохраняет некоторое свойство объекта. Интересно отметить, что если мы применяем к объекту одну симметрию, и сразу после этого — другую симметрию, то итоговое преобразование будет еще одной симметрией. Операция, которая оставляет объект неизменным — это тоже одна из симметрий. Для любой симметрии существует другая симметрия, которая откатывает назад все изменения, сделанные первой. И наконец, если нам нужно применить одну симметрию, затем вторую, а затем третью, то мы можем сгруппировать либо первую и вторую симметрию вместе, либо вторую и третье — и в итоге получить одно и то же преобразование.

Все эти свойства говорят нам о том, что множество симметрий объекта формируют **группу**. Группа — это множество G, вместе с операцией ·, для которого выполняются следующие свойства:

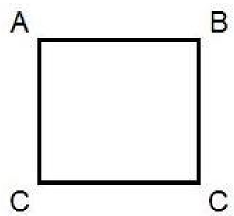
1. Если g и h — элементы G, то и g·h — тоже элемент G (Математики говорят, что G **замкнуто** относительно операции ·).
2. Если g, h и k — элементы G, то g·(h·k)=(g·h)·k (Это свойство операции группы называется **ассоциативностью**).
3. Существует элемент e в G такой, что g·e=e·g=g для всех g из G (Этот элемент e называется **нейтральным элементом** группы G, так как он оставляет каждый элемент G неизменным относительно операции).
4. Для каждого элемента g из G существует другой элемент h из G такой, что g·h=h·g=e, где e — нейтральный элемент (Элемент h называется **обратным элементом** к g, обозначается как g^{-1}).

Заметим, что в определениях выше мы используем мультипликативную нотацию для операции группы. Мы могли бы использовать аддитивную нотацию, и в этом случае обратный элемент для g обозначался бы -g. (*Мультипликативная нотация — это как будто наша операция «умножить», а аддитивная — «прибавить»*).

Упражнение Проверьте, что множество целых чисел Z с операцией сложения — группа. Для этого нужно проверить выполнение всех четырех свойств.

Группа целых чисел имеет еще одно очень интересное свойство: порядок сложения чисел совершенно не важен. То есть, для любых двух целых чисел a и b, a+b — это то же самое, что и b+a. Когда операция группы удовлетворяет этому свойству (это свойство называется **коммутативность**), группа называется **абелевой**.

Давайте рассмотрим пример группы симметрий, например, квадрат с пронумерованными углами:



Для этого геометрического объекта есть восемь преобразований, которые сохраняют тот факт, что это — квадрат. Другими словами, у квадрата восемь симметрий. Все они — различные повороты и отражения вместе с операцией композиции преобразований:

1. Поворот на 0 градусов (нейтральный элемент).
2. Поворот на часовой стрелке на 90 градусов.
3. Поворот на часовой стрелке на 180 градусов.
4. Поворот на часовой стрелке на 270 градусов.
5. Отражение относительно горизонтальной оси (которая проходит через центр квадрата).
6. Отражение относительно вертикальной оси (которая тоже проходит через центр квадрата).
7. Отражение по-диагонали из левого нижнего угла в правый верхний.
8. Отражение по-диагонали из левого верхнего угла в правый нижний.

Упражнение Выберите какие-нибудь два преобразования из списка выше и проверьте, что применение одного из них, а затем другого — это тоже преобразование, которое уже есть в этом списке. Можете ли вы найти два таких преобразования, что приводят к различным результатам в зависимости от порядка их применения?

В отличие от группы целых чисел, группа симметрий квадрата неабелева.

Группа симметрий G корректной сетки судоку содержит все преобразования квадрата и, помимо этого, некоторые другие преобразования вроде перемещения блоков, строк и столбцов, а также композиция всех этих преобразований. Таким образом, группа симметрий **порождается** преобразованиями следующих видов:

1. Переназначение девяти цифр.
2. Перестановка трех стеков.
3. Перестановка трех полос.
4. Перестановка трех столбцов в каком-либо стеке.
5. Перестановки трех строк в какой-либо полосе.
6. Всякие отражения и повороты (из списка симметрий квадрата).

Важное замечание: это не список всех элементов G! Это список различных симметрий группы, и комбинируя их всеми различными способами мы можем получить все остальные элементы группы. И конкретные преобразования всех описанных выше типов, и любая их композиция, которая отличается от этих конкретных «базовых» преобразований — все они включаются в группу симметрий G. Например, один из элементов G — это обмен первой и второй строк, конкретное преобразование типа (5). Пусть X — множество всех корректных сеток судоку. Мы знаем, что это множество из конечного числа элементов. Из того, что каждый элемент G — это некоторое соответствие, которое отображает одну из сеток на другую (*на самом деле оно отображает все сетки на какие-то другие, т.е. это некоторая перестановка всех элементов из X*), мы можем заключить, что G тоже включает в себя только конечное число элементов-симметрий.

Мы называем две сетки судоку **эквивалентными**, если мы можем преобразовать одну из них в другую применив одну или более симметрий из G. Если ни одна из последовательностей симметрий не преобразовывает одну из сеток в другую, то мы такие сетки называем **существенно различными**.

Это отношение действительно является отношением эквивалентности в формальном математическом смысле, поскольку оно удовлетворяет следующим трем свойствам:

1. Сетка A эквивалентна самой себе (это свойство называется **рефлексивностью**).
2. Если A эквивалентна B, то B эквивалента A (это свойство называется **симметричностью**).
3. Если A эквивалентна B, а B эквивалентна C, то A эквивалентна C (это свойство называется **транзитивностью**).

Для любой корректной сетки судоку A мы можем рассматривать все сетки, эквивалентные A, как действительно точно такие же, как A. Если мы сгруппируем вместе все сетки, которые эквивалентны друг другу, то мы, на самом деле, разобьем множество всех сеток на непересекающиеся части: в самом деле, X будет разбито на такие подмножества, что никакие два из них не имеют общих элементов. Математики называют такие подмножества **классами эквивалентности**. Любые два элемента из одного класса эквивалентности эквивалентны друг другу по какой-нибудь симметрии из G. Множество классов эквивалентности обозначается как X/G и читается как «X по модулую G» или «X mod G» (*еще математики называют X/G* ***фактор-множеством***).

В предыдущей части статьи мы задавались вопросом о количестве различных сеток судоку без всяких симметрий, и было бы интересно найти число существенно различных сеток. Согласно рассуждениям, изложенным выше, общее количество классов эквивалентности, или число элементов в X/G — это и есть ни что иное, как количество существенно различных сеток судоку. Далее мы рассмотрим метод, который в начале 2006 года использовали Эд Расселл (Ed Russell) и Фразер Джарвис (Frazer Jarvis) для вычисления этого числа.

Сначала мы немного обделим вниманием операцию переназначения чисел и рассмотрим только те симметрии, которые что-то делают именно с сеткой — со всей сеткой, с блоками или с отдельными ячейками. Рассмотрим эти симметрии — их типы (2)-(6) в списке выше — и их композиции. Эти симметрии дают нас группу H, в которой Расселл и Джарвис насчитали ровно 3359232 различных симметрий. Другими словами, H — это группа, которая порождается симметриями типов (2)-(6).

Комментарий

Комментарий

Теперь наше понятие эквивалентности двух сеток можно определить следующим образом: сетка A **эквивалентна** сетке B если мы можем сетку A преобразовать симметриями из группы H в некоторую сетку C, такую, что в C можно переназначить числа так, что в итоге получится сетка B. Мы можем сказать, что A **H-эквивалентна** C, а C эквиватентна B по переназначению. Заметим, что H-эквивалентность и эквивалентность по переназначению — это действительно отношения эквивалентности (они обе удовлетворяют трем свойствам из списка выше).

Мы считаем, что H **действует** на множество корректных сеток судоку X следующим образом: каждый элемент h из H представляет собой отображение X на X, то есть переводит каждую из сеток из X в другую (возможно ту же самую) сетку из X. Другими словами, h дает нам способ преобразования любой корректной сетки в другую корректную сетку, и каждый элемент h из H дает нам такое отображение.

Для любой симметрии из h из H мы можем рассмотреть те сетки, которые h **оставляет на месте** с точностью до переназначения. Мы имеем в виду все такие сетки A, что если мы применим к A симметрию h и получим сетку B, то B будет эквивалентна A по переназначению (*Такие объекты еще называются* ***неподвижными точками*** *относительно элемента h*). Так мы учитываем тот факт, что мы не учли переназначение в группе симметрий H.

Комментарий

Для того, чтобы посчитать количество существенно различных сеток судоку, нам нужна теорема из теории групп, под названием лемма Бернсайда.

Лемма Бернсайда Пусть G — конечная группа, которая действует на множестве X. Для каждого элемента g из G, пусть X^g обозначает множество элементов из X, которые элемент g оставляет на месте. Тогда количество элементов в множестве X/G равно $|X/G|=1/|G|\Sigma_{g \text{ in } G}|X^g|$, где |-| — количество элементов в множестве.

Упражнение Представьте квадрат, в котором каждое из ребер раскрашено в один из двух цветов — синий или зеленый. Две раскраски назовем существенно одинаковыми, если существует такая симметрия квадрата, что переводит первую из этих раскрасок во вторую. Используйте лемму Бернсайда для того, чтобы определить количество существенно различных раскрасок. (Для каждой из восьми симметрий квадрата посчитайте количество раскрасок, которые не меняются при применении симметрии. Для лучшего понимания происходящего можно нарисовать все $2^4=16$ раскрасок квадрата и представить себе как они переходят друг в друга. Лемма Бернсайда говорит нам, что для получения ответа нужно сложить количество не изменяющихся раскрасок для каждой симметрии и затем разделить полученную сумму на количество симметрий.)

В нашем же случае, конечная группа H действует на множество X всевозможных корректных сеток судоку. Для каждого h из H, мы хотим найти количество таких элементов X, что они не изменяются при применении h с точностью до переназначения чисел. Затем нам нужно сложить все эти числа и разделить их на число элементов в H. И в итоге получится ответ — количество существенно различных суток судоку. То есть, нам нужно взять среднее арифметическое от количеств неизменных сеток (с точностью до переназначения) по всем элементам H.

Например, следующая корректная сетка судоку является неподвижной точкой для поворота на 90° по часовой стрелке (опять же, с точностью до переназначения чисел):

1	2	4	5	6	7	8	9	3
3	7	8	2	9	4	5	1	6
6	5	9	8	3	1	7	4	2
9	8	7	1	2	3	4	6	5
2	3	1	4	5	6	9	7	8
5	4	6	7	8	9	3	2	1
8	6	3	9	7	2	1	5	2
4	9	5	6	1	8	2	3	7
7	1	2	3	4	5	6	8	9

Упражнение Определите как переназначить числа в повернутой сетке выше так, чтобы она совпала с изначальной. Выпишите все числа от 1 до 9 и посмотрите куда переходит число 1, затем куда переходит число 2, затем 3, ну и так далее. В общем, удостоверьтесь, что данная сетка — неподвижная точка для данной симметрии поворота.

Для того, чтобы применить лемму Бернсайда (чтобы вычислить количество существенно различных сеток), мы могли бы вычислить количество неподвижных точек для каждого из 3359232 элементов H. Но, оказывается, некоторые из этих 3359232 преобразований имеют одно и то же число неизменяемых сеток! Расселл и Джарвис использовали специальную программу GAP и определили, что все элементы H распадаются на 275 классов симметрий таких, что любые две симметрии из одного класса имеют одно и то же число неизменных сеток, и каждая симметрия из H имеет столько же неподвижных точек, сколько элемент одного из этих классов (*знакомый процесс, не правда ли?*). Поэтому нам нужно просто посчитать количество фиксированных точек только для одной симметрии из каждого из 275 классов. Зная сколько симметрий находится в каждом из классов, мы потом сможем узнать среднее, как в лемме Бернсайда.

Комментарий

Упражнение Рассмотрите отражение сетки судоку относительно горизонтальной оси, проходящей через центр сетки, то бишь через пятую строку. Заметим, что пятая строка после отражения останется неизменной. Можете ли вы переназначить числа в отраженной сетке так, чтобы она совпала с изначальной? И что можно сказать о неподвижных точках для отражения отосительно горизонтальной оси? Тут вообще есть сетки, которые не изменяются? (Вспомните тот факт, что мы сейчас рассматриваем какую-то корректную сетку судоку, и в пятой строке находятся все девять чисел).

Как можно видеть из упражнения выше, в H бывают такие симметрии, для которых нет неподвижных точек. Расселл и Джарвис с удивлением обнаружили, что на самом деле аж 248 классов из 275 содержат симметрии, для которых нет ни одной фиксированной точки. Так что им осталось только посчитать количество неизменных сеток для симметрий из 27 оставшихся классов и найти сколько симметрий содержится в каждом из этих классов перед тем, как использовать лемму Бернсайда. Они написали программу, которая делает последние вычисления и в результате получили число 5472730538≈5.473×10^9 — количество существенно различных сеток судоку.

Комментарий

Случай 4x4

Давайте посмотрим на судоку порядка 2, или сетку судоку размера 4×4, для того, чтобы лучше понять идеи, которые были рассмотрены выше (*и в предыдущей части тоже*).

Для того, чтобы посчитать количество сеток судоку 4×4, мы начнем — как и в случае с случаем 9×9 — с заполнения левого верхнего блока стандартным способом. Это делит количество различных сеток судоку на 4!, поскольку у нас есть ровно 4! переназначить числа в этом блоке.

Когда мы сделаем это, 1 и 2 окажутся на первых двух позициях первой строки, а 3 и 4 — на следующих двух позициях. Для каждой из 4! расстановок, у нас есть два способа вставить 3 и 4, и это влияет на общее число различных сеток. С другой стороны, если мы ходим посчитать лишь количество существенно различных сеток, порядок следования 3 и 4 не имеет значения, поскольку мы можем поменять местами третий и четвертый столбцы и получить все еще корректную сетку. Так что мы можем полагать, что 3 находится в третьей ячейке первой строки, а 4 — в четвертой.

Так как 1 и 3 уже имеются в первом столбце, 2 и 4 должны быть вставлены в третью и четвертую ячейку в этом столбце. Так как у нас два способа сделать это, наше количество различных сеток теперь считается сразу пачками по 4!×2×2 штук. Для подсчета количества существенно различных сеток, мы отмечаем, что обмен местами третьей и четвертой строк — это одна из симметрий сетки, так что мы можем считать, что 2 находится в третьей ячейке первого столбца, а 4 — в четвертой. К текущему моменту наша сетка имеет следующий вид:

1	2	3	4
3	4		
2			
4			

Поскольку в третьей строке есть 2, а в третьем столбце есть 3, в ячейку (3,3) мы можем вставить либо 1, либо 4.

Упражнение Покажите, что если в позицию (3,3) вставить 1, то это в конечном итоге приведет к нарушению Главного Правила.

Таким образом, в позиции (3,3) обязано находиться число 4. В четвертой строке, четвертом столбце, да еще и в правом нижнем блоке — во всех них у нас есть только 4, поэтому в ячейку (4,4) мы можем вставить любое из чисел 1, 2 и 3. Это дает нам следующие три сетки, то есть для каждой из 4!×2×2, которые мы насчитали ранее, есть ровно 3 способа поставить следующее число в позицию (4,4) (*а дальше сетки заполняются однозначным образом*). Таким образом, число различных сеток судоку 4×4 равно 4!×2×2×3=288.

1	2	3	4
3	4		
2		4	
4			1

1	2	3	4
3	4		
2		4	
4			2

1	2	3	4
3	4		
2		4	
4			3

Упражнение Заполните до конца все три сетки. После этого проверьте, что вторая и третья сетки на самом деле эквивалентны — можно отразить третью сетку по диагонали (какой из двух?) после чего переназначить числа 2 и 3.

То есть, в конечном итоге, у нас остается всего две существенно различные сетки судоку 4×4.

Еще немного интересных фактов

Правильно составленная головоломку судоку должна иметь единственное решение. В общем случае, судоку может иметь несколько решений, но в этом случае логических приемов, которые мы обсуждали в главе про подходы к решению, может оказаться недостаточно. Известен пример судоку порядка 3 с 17 числами, данными изначально, который правильно составлен. Минимальное же число подсказок для правильно составленного судоку порядка 3 — неизвестно (*на самом деле, уже* известно).

Упражнение Можете ли вы привести пример судоку, который не является правильно составленным?

Другой интересны вопрос (который мог прийти вам в голову, когда вы решали предыдущее упражнение): как много различных чисел может быть использовано для того, чтобы головоломка была правильно составленной? Оказывается, что для судоку порядка n, для получения единственного решения, необходимо использовать минимум n^2-1 различных чисел. Если мы для головоломки порядка n используем только n^2-2 различных чисел, то перестановка тех чисел, которых нет в подсказках, в одном из решений даст нам другое решение. В частности, для обычной судоку порядка 3, нужно использовать не менее 3^2-1=8 различных чисел, чтобы получить правильно составленную головоломку. Иначе пазл будет иметь более одного решения (*или ни одного?*).

Факт, описанный в предыдущем абзаце, можно переформулировать следующим образом: если судоку порядка n правильно составлена, то в ней должно содержаться как минимум n^2-1 различных чисел в самом начале. Важно отметить, что это не то же самое, что если судоку порядка n имеет n^2-1 различных чисел, то она правильно составлена. Вспомните, что если из A следует B, то из B не обязательно следует A. Следующее упражнение иллюстрирует этот момент специальным примером.

Упражнение Следующая судоку порядка 2 имеет 2^2-1=3 различных чисел. Найдите два различных решения.

1			
		2	
	3		

Для судоку 4×4, анализ всех (двух) случаев существенно различных сеток доказывает, что правильно составленная головоломка должна иметь как минимум четыре начальных подсказки.

Наконец, нельзя не отметить, что хотя компьютерные программы в мгновение ока решают судоку порядка 3, используя метод перебора с возвратом, решение судоку произвольного порядка n — гораздо более сложная задача. В то время как порядок судоку изменяется с n на n+1 (всего на единицу), время, нужное для решения, вырастает на порядки. Доказано, что задача решения судоку порядка n входит в класс **NP-полных** задач. Задача называется NP-полной, если для нее выполняются следующие два свойства:

1. Решение любой такой задачи можно проверить относительно быстро, т.е. за полиномиальное время.
2. Если хоть одна из таких задач может быть решена относительно быстро, то тогда относительно быстро можно решить и все задачи, удовлетворяющих свойству (1).

Хотя и проверить правильность решения NP-полных задач — быстро и легко, пока неизвестно алгоритма, который бы быстро находил это решение. Для алгоритмов, известных на текущий момент, время выполнения растет слишком быстро с ростом размера входных данных.

Вместо заключения

И на этой грустной ноте статья подходит к концу, как и ее перевод. Я лично в процессе перевода (а особенно написания комментариев) узнал и понял много нового. Надеюсь, вы тоже.

Список литературы оригинальной статьи можно посмотреть тут.

Спасибо всем, кто осилил до конца.

карма рейтинг
Артём Рипатти @ripatti 92,0 36,1
Математик-программист
Twitter

Самое читаемое

Разработка

Сейчас Сутки Неделя Месяц

+33 Ваш язык программирования — отстой
2,4k 32 6

+17 Визуальный генератор регулярных выражений
4,6k 64 30

+8 Компания Microsoft получила статус платинового участника Linux Foundation
1,1k 4 3

+12 Visual Studio для Mac и другие новости конференции Connect(); //2016
1k 4 0

+24 Сотворение мира Опыт создания разумной жизни своими руками
4,4k 75 8

Комментарии (6)

EndUser 16 ноября 2016 в 08:21 # 0

Меня позабавило, то, что когда минимальное количество подсказок хотели брутфорснуть распределёнными вычислениями, написали BOINC приложение, подключили публику — пришёл один математик и строго доказал, что это число равно 17 — задолго до того, как должен был завершиться полный перебор.

ripatti 16 ноября 2016 в 08:28 # 0

А можно ссылку на пруф этого утверждения? А то мне известна только ссылка на доказательство оптимизированным брутфорсом (она есть в тексте перевода). И считали они там 7 миллионов ядролет (!) на кластере из 300 серверных машин (через MPI, а не BOINC). Причем программа была написана на ассемблере.

EndUser 16 ноября 2016 в 08:48 # 0

http://www.distributedcomputing.info/ap-puzzles.html#sudoku

ripatti 16 ноября 2016 в 09:00 # 0

Интересно. К сожалению, больше информации по этому проекту нагуглить не могу.

Зато я нагуглил информацию по другому проекту. Они другим алгоритмом через BOINC подтвердили результаты, которые я привел в тексте перевода. Вот ссылки: раз два.

ripatti 16 ноября 2016 в 08:38 # 0

Впрочем, я припоминаю подобную историю с проверкой нулей зета-функции Римана. Вроде из-за этого закрыли проект ZetaGrid, поскольку они там всей толпой распределенными вычислениями проверили меньше нулей, чем один математик на своем десктопе…

LoadRunner 16 ноября 2016 в 14:31 # +1

Спасибо большое за перевод. Голова пухнет от попыток понять всё, что написано, особенно код, но в целом стало понятно, как они всё подсчитали. Только пока сам не проверю и не приду к таким же результатам — не поверю :)

Интересные публикации

- Ваш язык программирования — отстой 5
- Visual Studio для Mac и другие новости конференции Connect(); //2016 0
- Компания Microsoft получила статус платинового участника Linux Foundation 3
- Восемь потрясающих игр с искусственным интеллектом от компании Google 1
- Ю. Шмидхубер: «Прекрасно быть частью будущего искусственного интеллекта» 0