Пишите компараторы правильно

В Java для введения порядка среди определённых объектов можно написать компаратор — класс, содержащий функцию compare, которая сравнивает два объекта. Альтернативой компаратору является естественный порядок объектов: объект реализует интерфейс Comparable, который содержит метод compareTo, позволяющий сравнить этот объект с другим. Сравнивающая функция должна вернуть 0, если объекты равны, отрицательное число (обычно -1), если первый объект меньше второго, и положительное число (обычно 1), если первый больше. Обычно реализация такой функции не представляет сложностей, но имеется один случай, о котором многие забывают.  
  
Сравнение используется различными алгоритмами от сортировки и двоичного поиска до поддержания порядка в сортированных коллекциях вроде TreeMap. Эти алгоритмы завязаны на три важных свойства сравнивающей функции: рефлексивность (сравнение элемента с самим собой всегда даёт 0), антисимметричность (сравнение A с B и B с A должны дать разный знак) и транзитивность (если сравнение A с B и B с C выдаёт одинаковый знак, то и сравнение A с C должно выдать такой же). Если сравнивающая функция не удовлетворяет этим свойствам, алгоритм может выдать совершенно непредсказуемый результат. Причём скорее всего вы не получите никакого исключения, просто результат будет неверный.  
  
Как обнаружилось, несоблюдение этих свойств — не такая уж редкая ситуация. Проблема возникает при сравнении вещественных чисел — float или double.  
  
Предположим, у нас имеется класс с полем типа double, и мы хотим упорядочивать объекты этого класса в зависимости от значения поля. Нередко можно встретить такой код:

public class DoubleHolder implements Comparable<DoubleHolder> {

double d;

public DoubleHolder(double d) {

this.d = d;

}

@Override

public int compareTo(DoubleHolder o) {

return d > o.d ? 1 : d == o.d ? 0 : -1;

}

@Override

public String toString() {

return String.valueOf(d);

}

}

У приведённого метода compareTo есть две проблемы. Первая — незначительная — он не различает +0.0 и -0.0: new DoubleHolder(-0.0).compareTo(new DoubleHolder(+0.0)) вернёт 0. Иногда это нестрашно, но в случае сортировки элементы с +0.0 и -0.0 расположатся в произвольном порядке, что будет смотреться некрасиво. Тем не менее, всё это мелочи по сравнению с NaN. Число NaN (как в типе double, так и во float) — это довольно специфичная вещь. Оно не больше, не меньше и не равно никакому другому числу. В результате мы сразу получаем нарушение свойств сравнения:

DoubleHolder nan = new DoubleHolder(Double.NaN);

DoubleHolder zero = new DoubleHolder(0.0);

System.out.println("nan.compareTo(nan): "+nan.compareTo(nan)); System.out.println("nan.compareTo(zero): "+nan.compareTo(zero)); System.out.println("zero.compareTo(nan): "+zero.compareTo(nan));

nan.compareTo(nan): -1

nan.compareTo(zero): -1

zero.compareTo(nan): -1

Ни рефлексивности, ни антисимметричности не наблюдается. Можно встретить и такую реализацию сравнения:

public int compareTo(DoubleHolder o)

{ return d > o.d ? 1 : d < o.d ? -1 : 0; }

Здесь все три предыдущих сравнения выдадут ноль, то есть как будто бы свойства соблюдаются. Но, конечно, радоваться рано:

DoubleHolder nan = new DoubleHolder(Double.NaN);

DoubleHolder zero = new DoubleHolder(0.0);

DoubleHolder one = new DoubleHolder(1.0);

System.out.println("zero.compareTo(nan): "+zero.compareTo(nan)); System.out.println("nan.compareTo(one): "+nan.compareTo(one)); System.out.println("zero.compareTo(one): "+zero.compareTo(one));

zero.compareTo(nan): 0

nan.compareTo(one): 0

zero.compareTo(one): -1

Здесь нарушается транзитивность: первый объект равен второму, второй равен третьему, но первый третьему не равен.  
  
Чем же это грозит простому обывателю? Чтобы понять это, создадим простой список и попробуем его перемешать и посортировать несколько раз:

List<DoubleHolder> data = new ArrayList<>();

for(int i=1; i<=10; i++) {

data.add(new DoubleHolder(i));

}

data.add(new DoubleHolder(Double.NaN));

for(int i=0; i<10; i++) {

Collections.shuffle(data);

Collections.sort(data);

System.out.println(data);

}

Вывод в каждом запуске отличается и может выглядеть, например, так:

NaN, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[NaN, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[1.0, NaN, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[NaN, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, NaN]

[2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, NaN, 1.0, 9.0, 10.0]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 10.0, NaN, 9.0]

[NaN, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[NaN, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[1.0, NaN, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

Или для второй реализации compareTo:

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0, 9.0, NaN, 7.0, 10.0]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, NaN]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, NaN]

[1.0, 2.0, 9.0, 10.0, NaN, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, NaN]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, NaN]

[2.0, 6.0, NaN, 1.0, 3.0, 4.0, 5.0, 7.0, 8.0, 9.0, 10.0]

[2.0, 4.0, NaN, 1.0, 3.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[1.0, 3.0, NaN, 2.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]

[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, NaN]

Примерно в половине случаев элемент с NaN прибивается к началу или концу сортируемого списка, а в других случаях начинает блуждать, портя порядок окружающих элементов.  
  
С коллекциями, использующими DoubleHolder в качестве ключа, тоже ничего хорошего не произойдёт. Возьмём, к примеру, TreeSet. Со второй реализацией compareTo всё довольно просто: так как элемент, содержащий NaN, равен любому другому, то в непустое множество вставить его не получится, потому что оно решит, что такой элемент уже есть. Но берегитесь, если вы вставили NaN-элемент первым: после этого во множество не выйдет добавить ничего другого.  
  
Первый вариант сравнивающей функции психоделичнее. Напишем, например, такой тест:

Set<DoubleHolder> set = new TreeSet<>();

for(int i=0; i<50; i++) {

set.add(new DoubleHolder( i%10==0 ? Double.NaN : i%10 ));

}

System.out.println(set);

Мы вставили по пять элементов, содержащих NaN, и по пять элементов, содержащих каждую цифру от 1 до 9. В результате имеем следующее:

[NaN, NaN, 1.0, 2.0, 3.0, NaN, NaN, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, NaN]  
  
Вполне ожидаемо увидеть пять раз NaN: ведь они не равны друг другу. Но из-за неправильных сравнений и некоторые другие элементы вставились по нескольку раз. Можете сами посмотреть, что случится с TreeMap. Заметьте, что один случайно попавший NaN может испортить всю коллекцию, причём это не всегда легко отладить: коллекция может долго существовать в некорректном состоянии и делать вид, что всё нормально.  
  
Что любопытно, этой проблемы вообще не должно существовать. Ещё в JDK 1.4 появились специальные статические методы Float.compare и Double.compare, которые сделают всё за вас, корректно обработав специальные случаи. Надо лишь написать:

public int compareTo(DoubleHolder o) {

return Double.compare(d, o.d);

}

Видимо, сказывается обманчивая простота алгоритма сравнения: порой программист считает, что проще написать самому, чем искать в документации стандартный способ сделать это.  
  
Примеры неправильного сравнения double/float в различных открытых проектах: [JTS](http://grepcode.com/file/repo1.maven.org/maven2/com.vividsolutions/jts/1.13/com/vividsolutions/jts/index/intervalrtree/IntervalRTreeNode.java#72), [Batik](http://grepcode.com/file/repo1.maven.org/maven2/fr.avianey.apache-xmlgraphics/batik/1.8pre-20141006/org/apache/batik/gvt/flow/FlowRegions.java" \l "188), [Hadoop](http://grepcode.com/file/repo1.maven.org/maven2/org.apache.hadoop/hadoop-common/2.5.1/org/apache/hadoop/io/DoubleWritable.java" \l "78), [Hudson](http://grepcode.com/file/repo1.maven.org/maven2/org.eclipse.hudson/hudson-core/3.2.1/hudson/ExtensionComponent.java" \l "70), [ICU4J](http://grepcode.com/file/repo1.maven.org/maven2/com.ibm.icu/icu4j/54.1.1/com/ibm/icu/util/ULocale.java#2108), [Lucene](http://grepcode.com/file/repo1.maven.org/maven2/org.apache.lucene/lucene-smartcn/3.6.2/org/apache/lucene/analysis/cn/smart/hhmm/PathNode.java" \l "35). Трудно определить, в каких случаях это может привести к проблемам, но это тот случай, когда я бы исправлял безусловно: правильный и надёжный вариант обычно при этом ещё и короче неправильного.  
  
Чтобы изменить ситуацию, я [написал](https://code.google.com/p/findbugs/source/detail?r=ef65d4ee875d25690a5dcefbf34a1a7c378f1276) маленький детектор для FindBugs, который находит некорректно реализованные функции сравнения и предлагает использовать Float.compare/Double.compare.  
  
Вообще для всех примитивных типов есть подобные методы. Если вам надо сравнить несколько полей по очереди, можно написать так:

public class MyObject implements Comparable<MyObject> {

double d;

int i;

String s;

char c;

@Override

public int compareTo(MyObject o) {

int result;

result = Double.compare(d, o.d);

if(result != 0) return result;

result = Integer.compare(i, o.i);

if(result != 0) return result;

result = s.compareTo(o.s);

if(result != 0) return result;

result = Character.compare(c, o.c);

return result;

}

}

Смотрится лучше, чем куча веток с больше и меньше. А если вы пользуетесь [Guava](http://habrahabr.ru/post/244347/) или чем-то подобным, тогда так:

@Override

public int compareTo(MyObject o) {

return ComparisonChain.start()

.compare(d, o.d)

.compare(i, o.i)

.compare(s, o.s)

.compare(c, o.c)

.result();

}