稀疏矩阵的存储演变路径

叶萌萌
浙江经济职业技术学院　浙江　杭州　310018

摘　要：为提高稀疏矩阵存储空间利用率和计算效率，稀疏矩阵的存储方式必须有别于普通矩阵的存储方式。本文从数学的角度，讨论稀疏矩阵的存储如何一步一步优化和改进，并以具体的实例分析各种存储方式的优点和缺点，计算各种存储方式的空间利用率或优化率。

关键词：稀疏矩阵；存储空间；优化率

引言

矩阵在现代计算中有着不可缺少的地位，可以说整个计算都是建立在矩阵的基础之上。矩阵是线性代数的最基本的概念，而稀疏矩阵是一种特殊的矩阵形式。对于刚接触线性代数和矩阵的大学生，学习完矩阵及其运算，它们在计算机世界里如何使用呢？首先要面对的就是如何把矩阵‘放入’计算机里，即矩阵的存储。

1 稀疏矩阵

稀疏矩阵是指矩阵中有非常多的0元素，在大规模矩阵里面，通常认为非零元素小于5%的矩阵为稀疏矩阵。稀疏矩阵在现实生活中有各种应用，比如在图像处理中，图像的数据通常是以稀疏矩形的形式呈现出来的，要对图像进行分析和处理，通常会转换到对稀疏矩阵进行奇异值分解[1]；在信号处理中，我们可以借助稀疏矩阵实现对信号的鲁棒分类[2]；在机器学习中，我们可以借助稀疏矩阵，做特征选择和模型压缩，以优化算法的设计[3]。因此稀疏矩阵的存储研究和优化一直是科学计算和高性能计算领域的热点问题。为更好地理解稀疏矩阵的存储，我们将以一个具体的例子来分析各种存储方法的实现和空间优化率。稀疏矩阵A如下

$$\begin{pmatrix} 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 7 & 0 & 0 & 8 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (1)$$

2 稀疏矩阵的存储方法

2.1 二维数组存储

常见的矩阵存储是用一个二维数组来存储这个矩阵，因此它所需要的存储空间就是两个维度的乘积。比如上面稀疏矩阵是一个5行10列的矩阵，它就需要用5×10 = 50个位置来存放这50个数据。这种方式直接、简单明了。但是如果有大量的零元素，则会浪费许多内存空间。

试想一下，我要对一篇文章使用文档频次法进行自动分类[4]，需要考虑被分类文章词语与已有类别文章词语之间的频次，据商务印书馆出版的现代汉语常用词表可知，中文里经常使用的词共有56,008[5]个，这相当于构建了一个56008×N的矩阵，其中N表示待匹配文章的类型个数。它的存储和计算将占用很大的内存，更重要的是这，各矩阵里面的绝大多数元素都是0，我们完全没有必要把这些0保存下来，因此我们需要考虑稀疏矩阵如何更有效的存储。

2.2 三元组存储

三元组存储是指存储矩阵当中的非零元素，比如上面二维矩阵中50个元素只有10个元素是非零的，我们只需记录这些非零元素在矩阵当中的行号、列号以及元素的值即可。由于矩阵行、列地位同等，所以有按行搜索非零元素和按列搜索非零元素之分，其三元存储形式分别如下表1、表2所示。

表1 矩阵A非零元素的按行优先三元存储

行号	列号	元素的值
1	1	1
1	4	2
1	9	3
2	2	4
2	7	5
4	2	6
4	5	7
4	8	8
5	1	9
5	6	1

作者简介：叶萌萌（1988年—），女，湖南常德人，讲师，硕士，主要从事高等数学、线性代数、数学建模教学及研究。

表2 矩阵*A*非零元素的按列优先三元存储

列号	行号	元素的值
1	1	1
1	5	9
2	2	4
2	4	6
4	1	2
5	4	7
6	5	1
7	2	5
8	4	8
9	1	3

三元数组存储，它旨在筛选出非零元素的信息，对于0元素不需要记录。本例当中原矩阵采用二维数组存储需要50个位置，而采用三元数组法存储只需要记录30个位置。相比较节约了20个位置空间，既节省了约40%的存储空间。

试想上面的文章自动分类，或者是更大的矩阵，比如网页排名，假设全球有10亿网页（2014年数据），计算网页排名用到的将是有100亿个元素，采用三元存储法节省下的空间将是巨大的。无论是按行存储优先还是按列存储优先，我们会发现行号或列号有很多重复的地方，因此我们进一步思考是否这个地方还能有更高效的存储办法。

2.3 三元存储法改进法一

对于按行存储优先的存储法如表1，我们保留第二列列号和第三列元素的值，把第一列的信息改为存储每一行非零元素的起始位置，如下表3所示

表3 按行存储优先的改进法一

列号	元素的值	每行非零元素的起始位置
1	1	1
4	2	4
9	3	6
2	4	6
7	5	9
2	6	
5	7	
8	8	
1	9	
6	1	

三元数组法存储需要记录30个位置，此方法需要25个位置用来记录信息，较上面三元存储方式节省的位置5个，即节省存储空间16.7%。

试想如果原矩阵非常大，比如汉语词典里单词之间的同现频率矩阵，这将是一个的矩阵，若以普遍认为的5%的非零元素来看，其非零元素将有近1.6亿个元素，用三元存储法将要记录亿个位置，采用改进的三元存储法一需要个位置，亿，那节省出来的空间几乎就是一列的空间，因此空间的利用率提高了近33%。

同理，对于表2我们也可以类似处理，进行压缩，得到如下表4：

表4 按列存储优先的改进法一

行号	元素的值	每列非零元素的起始位置
1	1	1
5	9	3
2	4	5
4	6	5
1	2	6
4	7	7
5	1	8
2	5	9
4	8	10
1	3	

同一个矩阵，在计算领域，很可能要多次使用。有时候需要横向使用，有时则需要竖向使用。因此上面按行存储优先和按列存储优先都有需要。那同一矩阵是否的存储两遍呢?

2.4 三元存储法改进法二

根据矩阵乘法的运算定义可知，

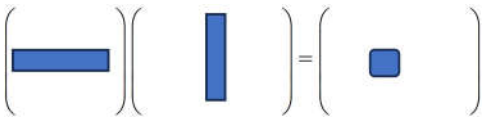

图1 矩阵乘法示意图

如果需要进行矩阵的乘法运算，我们需要对这个矩阵进行按行优先和按列优先两种方式存储，如果按行优先存储已经存储了元素的值，那么按列优先的时候，我们只需要记录元素所在的位置即可，如下表5所示，

表5 按列存储优先的改进法二

行号	元素的值	每列非零元素的起始位置
1	1	1
5	9	3
2	4	5
4	6	5
1	2	6
4	7	7
5	10	8
2	5	9
4	8	10
1	3	

注意：此处元素的位置是指表4中元素在表3中是第几个。

这样做的原因和优点有两个：

●如果同样的矩阵以不同的形式存两份，若工程师要进行元素的修改，则需要进行两次修改。有时会出现工程师修改一个地方，而忘记修改另一个地方，则运算时会造成差错[6]；

●在现实生活中，我们的元素通常不是一个单独的数值，而是一条记录，包含了许多的信息在里面。若是按行优先、按列优先两次存储，那么则浪费了巨大的存储空间，而存储元素的位置的话，则记录只需存储一次，第二次只需用数值存储记录的位置，在超大矩阵（每个元素是一条信息量丰富的记录）存储中，其节省的存储空间接近50%。

稀疏矩阵的存储优化一直在发展中，针对不同行业、特殊形式的稀疏矩阵的存储亦然如此。

3 总结

在信息爆炸的现代社会，每天都有大量的数据产生，而这些信息的数量通常是超过计算机的存储容量的，所以设计好的存储方式，将信息最大限度的存放进有限的空间里，是一代代计算机人的研究的重点。本文以线性代数中普通矩阵的存储为出发点，以刚接触矩阵概念的大学新生视角，看稀疏矩阵的存储是如何一步步优化和改进，希望能让学生感受到线性代数与计算世界的联系，增加高等数学学习的兴趣。

参考文献

[1]卢文锋,侣同光,韩国勇.基于稀疏表示和低秩矩阵逼近的图像去噪算法的研究[J].首都师范大学学报(自然科学版),2022,43(01):19-23.DOI:10.19789/j.1004-9398.2022.01. 004.

[2]黄翔东,杨琳,杨孟凯,等.抗噪鲁棒性可分级的稀疏阵列频率和到达角估计[J].电子学报,2019,47(01):122-128.

[3] X.Li,Y.Wang and R. Ruiz,"A Survey on Sparse Learning Models for Feature Selection," in IEEE Transactions on Cybernetics,vol.52,no.3,pp.1642-1660,March 2022,doi:10.1109/TCYB.2020.2982445.

[4]叶萌萌.线性代数在互联网应用案例分析---以利用余弦定理对文章分类为例[J].中国教工,2023:71-73.

[5]李行健.现代汉语常用词表（第2版）[M].北京:商务印书馆.https://tech.huanqiu.com/ article/9CaKrnJFzKw?imageView2/2/w/228

[6]吴军.计算之魂[M].北京：人民邮电出版社,2022：123-131