Web data mining部分：

第四章讨论部分：

读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：

使用k-means算法的时候，全局最小值对于大规模数据集来说在计算上是不可行的，那么有没有方法在保证计算相对较小的情况下，获得较高的相对最小值呢？

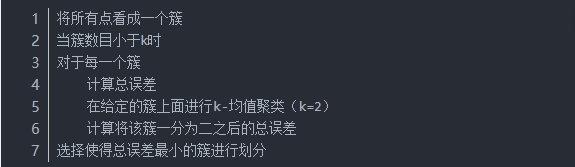
讨论后的理解：

有两种方法，第一种想法比较简单，用时间换较高的相对最小值，多试几次，多几次不同的选种，看到一个博客用实际数据跑出的log n次有比较好的表现，那么时间复杂度o（nlogn）换较高的相对最小值。

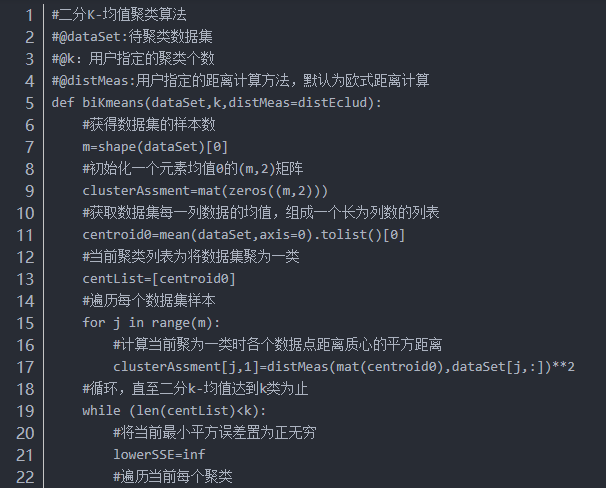
第二种想法：基于二分法的k-means算法改进

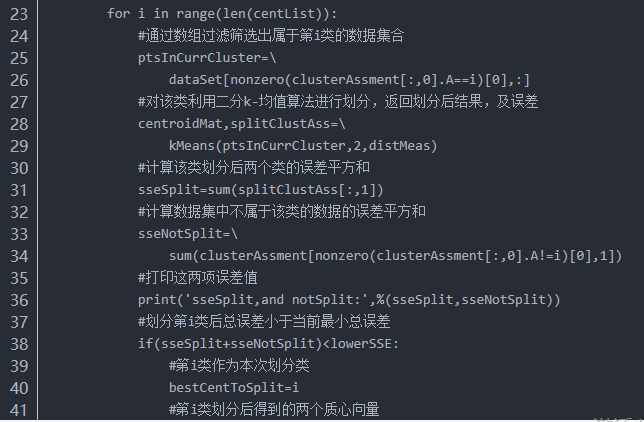
二分K-means算法首先将所有点作为一个簇，然后将簇一分为二。之后选择其中一个簇继续进行划分，选择哪一个簇取决于对其进行划分是否能够最大程度的降低SSE的值。上述划分过程不断重复，直至划分的簇的数目达到用户指定的值为止。

伪代码如下：

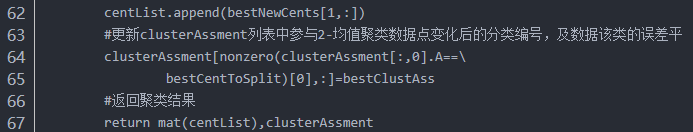


也可以选择sse最大的族进行划分直到到达k为止，伪代码如下：







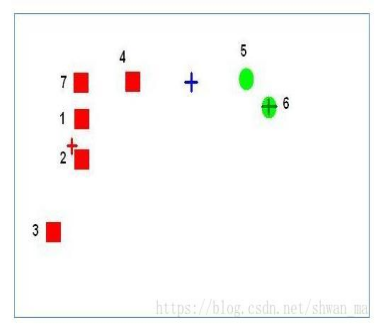
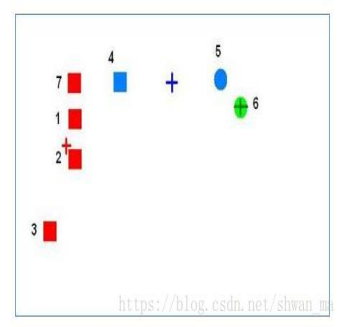
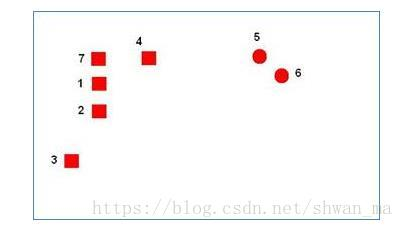


1. 提出的问题2：

空聚类是如何产生的？为什么解决空聚类的时候，选择离一个含有大量数据的聚类的聚类中心最远的数据点？

讨论后的理解：

针对这个问题，我们进行了激烈的讨论，最后结果是，举例如图所示：



一开始选择了3，5，6为种子，后来第二个聚类本来划分的4和5会被它两遍的两个聚类抢走从而形成了空聚类，形成了图三所示的格局。

因为空聚类因为没有数据点，不选择干预的话，聚类中心一直不会变，所以要手动指定一个大聚类中最远的数据点让他们重新聚类，形成新的聚类。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：如何找到最佳的簇数？

自己的理解：

主流的确定最佳的k的方法有两种，一种是手肘法，一种是轮廓系数法

手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。

轮廓系数法的核心指标是轮廓系数（Silhouette Coefficient），求出所有样本的轮廓系数后再求平均值就得到了平均轮廓系数。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的k便是最佳聚类数。

1. 问题4：如何保证聚类算法一定收敛呢？换言之，在算法实现时，如何判断迭代过程收敛呢？因为如果按照P100(中文版第二版)提到的三个收敛条件终止迭代，但是一个条件都无法满足，那么这时候要怎么去判断终止（收敛）条件呢？如果通过设置迭代次数终止，阈值该如何考虑？如果通过判断两次迭代结果的差别，又该从哪些角度考虑呢？

自己的理解：

这个问题和潘晓梅进行了讨论，最后想法是这样的，首先聚类的算法一定是可以收敛的，这可以进行数学证明，但是具体迭代多少次可以完成收敛，我个人认为是，没有一个确定的阈值，因为种子选取是随机的，无论是什么数学模型都没办法考虑到种子随机的影响，倒是k可以考虑进去。但是却是可以有别的方法可以作为迭代终止条件，比如说，对聚类进行评估，如果评估的结果可以满足需求，那么就算可以进行更好的聚类，也没有必要。评估聚类的方法就有很多，比如说，评估聚类的分离度和压缩密度（sse），或者间接评估聚类等方法。

1. 问题5：在混合属性的处理中，为什么将一个具有多余两个状态的名词性属性或者非对称布尔属性转换为区间度量函数是没有意义的？

自己的理解：

因为，多余两个状态的名词属性转换区间度量是没有意义的，比如说水果，现在有苹果，橘子，梨子。如果把苹果设置为1，橘子设为0，梨子0.5.这其实是没有任何意义的，因为如果算出来结果是0.7，这样的结果是没有任何意义的，所以只能设置成向量苹果（0，0，1），梨子（0，1，0），橘子（1，0，0）。

问题6：如果数据点很适合被分成2个聚类，但用户给的K值为3怎么办？或者适合分成3类，k值为2，没有很理想的分法怎么办？

自己的理解：

如果客户给了k值，那必然是按照客户的需求来的，说到底不过是再讲大的聚类 再进行划分，其实也就相当于衣服的尺码，s码，l码两码和s码，l码，xl码这种问题。对于客户来说是否需要再将l码划分成l码和xl码。其实实际上，书中所说的局部最小值就是对于指定的k，然而又根据选择的种子决定当前选中的相对最小值，所以k-means算法是已知的k中来求相对解的，对于不同的k有不同聚类。更理想的办法当然是在算法开始之前用手肘法确定较为合适的k再进行计算。

1. （必填）读书计划
2. 本周完成的内容章节：

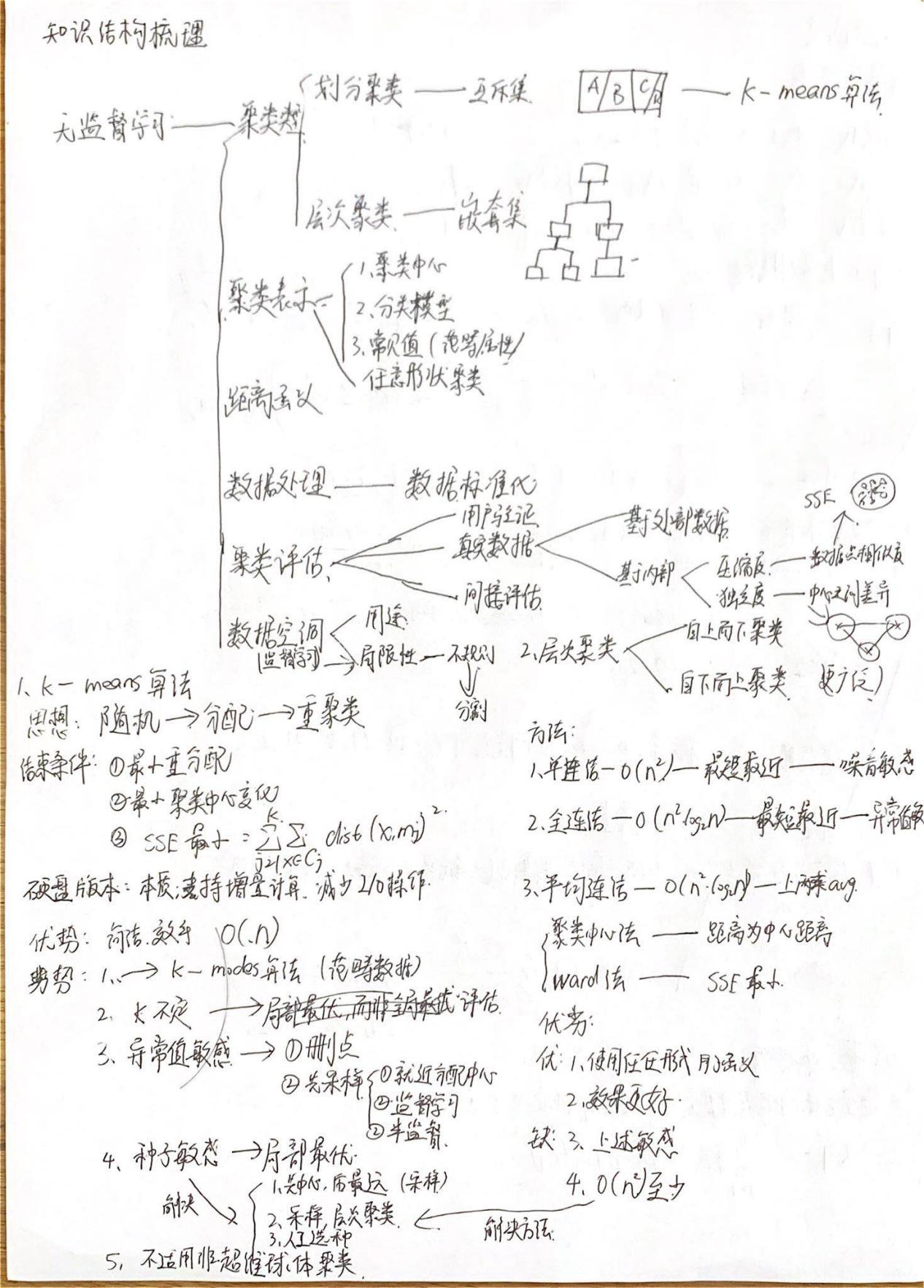
（1）4.1到4.10完成并且梳理知识点，寻找问题，自己思考并且于小组会之前完成了自己的思考。

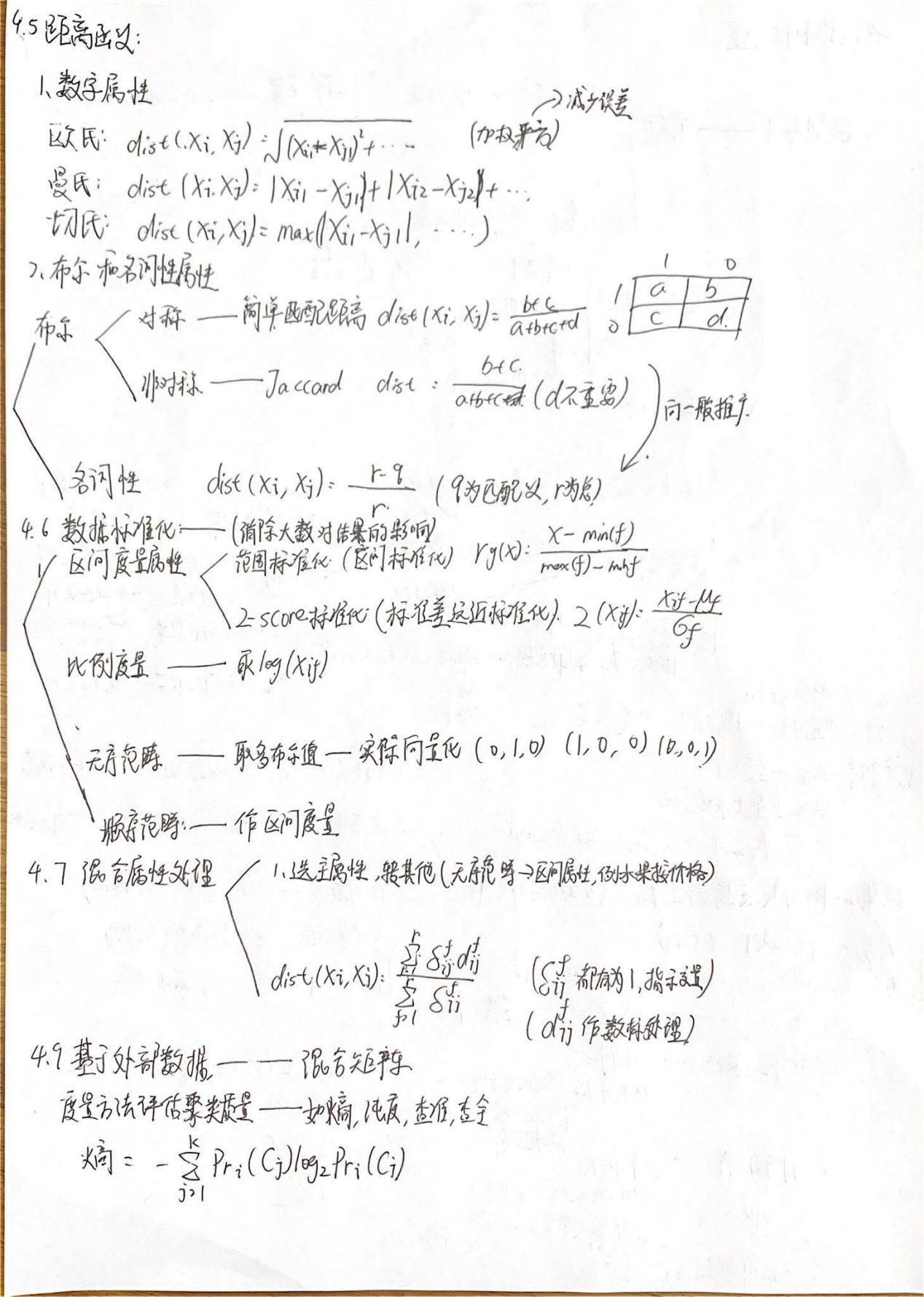
2、下周计划：第三章的阅读和梳理

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

我自己手动梳理了第四章所以的知识框架结构扫描如下：





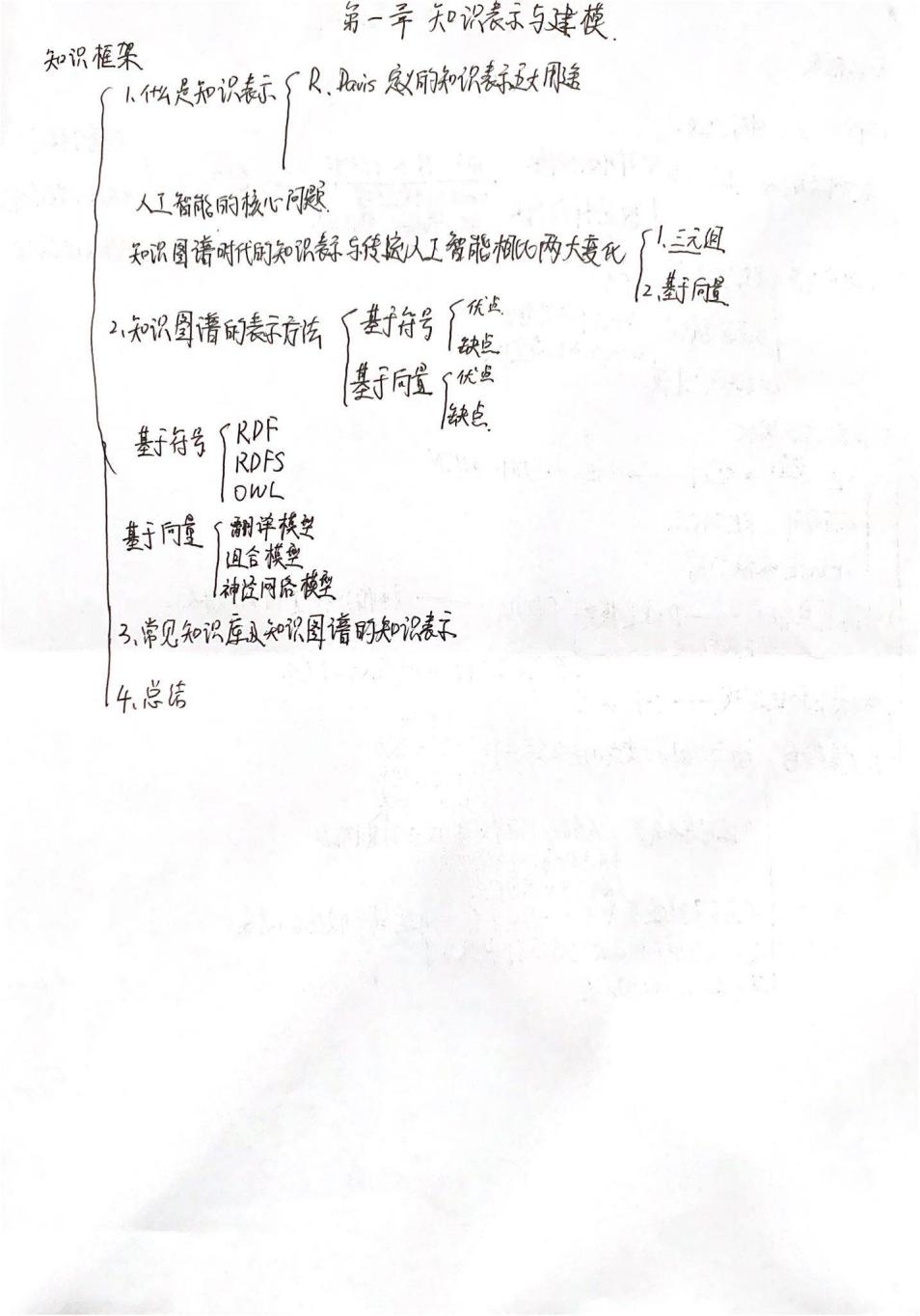
知识图谱部分：

读书报告内容：

分章节知识整理知识框架如下：

第一章：

知识框架：



知识点：

1.知识表示的五大用途或特点：

（1）客观事物的机器标示（A KR is a Surrogate），即知识表示首先需要定义

客观实体的机器指代或指称。

（2）一组本体约定和概念模型（A KR is a Set of Ontological Commitments），

即知识表示还需要定义用于描述客观事物的概念和类别体系。

（3）支持推理的表示基础（A KR is a Theory of Intelligent Reasoning），即知

识表示还需要提供机器推理的模型与方法。

（4）用于高效计算的数据结构（A KR is a medium for Efficient Computation），

即知识表示也是一种用于高效计算的数据结构。

（5）人可理解的机器语言（A KR is a Medium of Human Expression），即知

识表示还必须接近于人认知，是人可理解的机器语言。

2.知识图谱对规模的扩展需求使得知识表示方法逐渐发生了四个方面的变化：

（1）从强逻辑表达转化为轻语义表达；（2）从较为注重 TBox 概念型知识转化为更加注重 ABox 事实型知识；（3）从以推理为主要应用目标转化为综合搜索、

问答、推理、分析等多方面的应用目标；（4）从以离散的符号逻辑表示向以连续

的向量空间表示方向发展。

1. 基于向量的知识表示方法和基于规则的知识表示方法比较：

传统基于逻辑的符号知识表示的优点是基于显性知识表示，因而表示能力强，能处理较为复杂的知识结构，具有可解释性，并支持复杂的推理。基于表示学习的连续向量表示优点是易于捕获隐性知识，并易于与深度学习模型集成，缺点是对复杂知识结构的支持不够，可解释性差，不能支持复杂推理。

1. 总结：

（1）现代知识图谱由于要满足规模化的扩建需求，大多降低了对强逻辑表达的要求，并以三元组为基础的关系型知识为主。并更多的关注实例层面的知识构建。

（2）由于知识图谱是很多搜索、问答和大数据分析系统的重要数据基础，基于向量的知识图谱表示使得这些数据更加易于与深度学习模型集成，使得基于向量空间的知识图谱表示得到越来越多的重视。

第二章：

自己的问题1：

将大量对象投影到统一空间的过程，能够将高频对象的语义信息用于帮助低频对象的语义表示，提高低频对象的语义表示的精准性。这句话怎么理解？

思考后的回答1：

在低维向量空间中，两个对象的距离越近，说明他们的语义相似度越高。如果有高频对象的和低频对象向量距离较近，那么可以利用高频对象扩充该低频对象的语义表述，提高低频对象的精准性。

自己的问题2：

什么叫知识分布式表示？

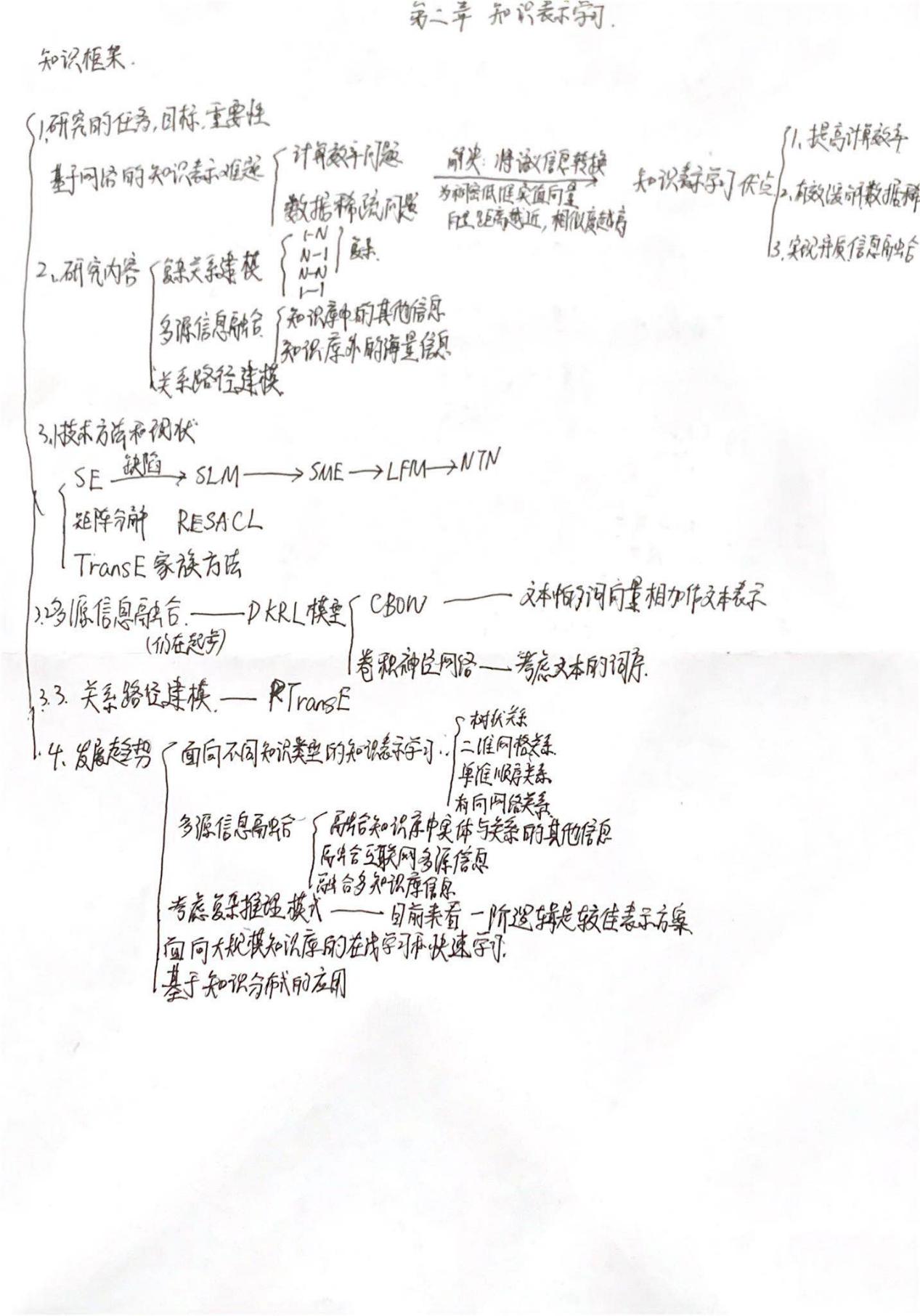
思考后的回答2：

知识图谱的向量表示就是一种知识分布式表示。

形式上，是从高维的离散的空间，转化为低维连续的空间R。

从训练过程上看，空间R是知识之间博弈的结果。实际上就是知识的分布式表示。

知识框架：



知识点整理：

1. SE模型：

对于一个事实三元组，SE 将头实体向量和尾实体向量通过关系的两个矩阵投影到关系的对应空间中，然后在该空间中计算两投影向量的距离。

这个距离反映了两个实体在该关系下的语义相关度，它们的距离越小，说明这两

个实体存在这种关系。

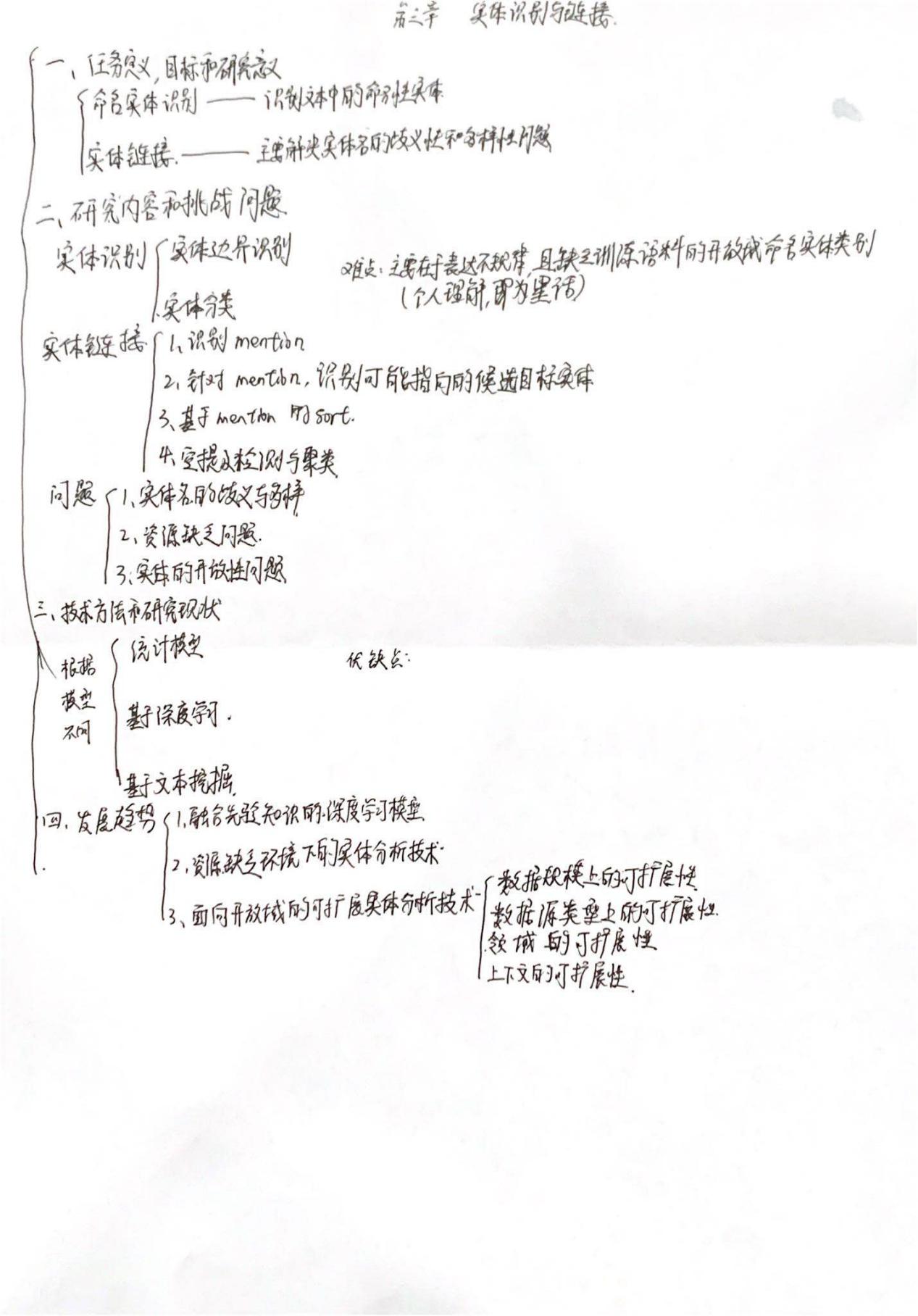
缺陷：

它对头、尾实体使用两个不同的矩阵进行投影，协同性较差，往往无法精确刻画两实体与关系之间的语义联系。

1. SLM模型尝试采用单层神经网络的非线性操作仅提供了实体与关系之间比较微弱的联系，却计算复杂度更高。
2. LFM通过简单有效的方法刻画了实体和关系之间的语义联系，协同性较好，计算复杂度低，后来的DISTMULT进一步简化，NTN又在LFM上引入了张量操作，但计算复杂度较高在大规模稀疏知识上效果较差。
3. 矩阵分解同样是得到低维向量表示的重要途经。

第三章

知识框架：



知识点：

1. 统计模型的方法，优越点：

基于统计模型的方法通常将实体识别任务形式化为从文本输入到特定目标结构的预测，使用统计模型来建模输入与输出之间的关联，并使用机器学习方法来学习模型的参数。

实体链接的核心是计算实体提及（mention）和知识库中实体的相似度。

传统统计模型的主要缺点在于需要大量的标注语料来学习，另外一个缺点是其需要人工构建大量的特征，其训练并非一个端到端的过程。

1. 深度学习的方法，优越点：

实体链接的核心是构建多类型多模态上下文及知识的统一表示，并建模不同信息、不同证据之间的相互交互。通过将不同类型的信息映射到相同的特征空间，并提供高效的端到端训练算法。

深度学习方法的主要优点是其训练是一个端到端的过程，无需人工定义相关的特征。另外一个优点是深度学习可以学习任务特定的表示，建立不同模态、不同类型、不同语言之间信息的关联，从而取得更好的实体分析性能。

1. 文本挖掘的方法，优缺点：

传统统计方法和深度学习方法都需要大量训练语料和预先明确定义的目标实体类别，无法处理大数据环境下的开放实体分析任务。

文本挖掘方法的核心是构建从特定结构（如列表、Infobox）构建实体挖掘的特定规则。由于规则本身可能带有不确定性和歧义性，同时目标结构可能会有一定的噪音，文本挖掘方法往往基于特定算法来对语义知识进行评分和过滤。

文本挖掘方法只从容易获取且具有明确结构的语料中抽取知识，因此抽取出来的知识质量往往较高。

（1）绝大部分结构化数据源中的知识都是流行度高的知识，对长尾知识的覆盖不足

（2）人们发现现有结构化数据源只能覆盖有限类别的语义识，相比人类的知识仍远远不够.

读书计划：

**知识图谱自己完成了前三章的阅读。**

学习方式：

第一遍完成阅读，有不理解的地方圈画作标记，圈画自己认为重点做标记。

第二遍再次阅读，完成知识框架的构建。

第三遍回顾解决问题。

下周计划：

1至少完成完成报告第四章的阅读。

2看完孙浩然师兄的《基于百科的新冠知识类型推断方法研究》论文，了解一下工作。