本章主要介绍适航领域相关理论、软件配置管理相关理论和技术。首先介绍了适航领域中机载软件相关的标准DO-178B/C，以及配置管理过程与DO178C标准中其他过程的关系， 对变更管理的基本概念、主要活动和主要功能和活动件关系等进行了介绍，然后对信息检索技术的常用技术和模型进行介绍；接下来介绍了本论文中使用的word embedding；最后介绍了学习排序算法的基本概念，以及在软件工程领域中的使用情况和已有的方法。

## 适航领域相关理论

本节首先介绍适航领域软件相关的标准，以及标准中规定的与普通软件开发不同的软件生命周期和相应的软件生命周期数据，

### DO-178B/C标准介绍

在20世纪70年代，美国联邦航空管理局(FAA)、欧洲航空安全局(EASA)和其他全球航空安全机构首先援引了RTCA/DO-178文件。该文件的标题是“机载系统和设备认证中的软件考虑事项”。“该文件的目的是确保飞行中软件的安全，它通过实施结构化的开发过程来做到这一点。自20世纪70年代以来，负责该文件的RTCA委员会每十年就对其进行实质性修订，以跟上技术的发展。2013年7月，FAA通过AC20-115C启用了第四代标准DO-178C(欧洲称为ED-12C)。目前，适航标准RTCA DO-178是商业和军事航空航天项目事实上的软件标准RTCA

DO-178C提供的是一种结构化的机载软件设计保证指南，它是基于目标，面向过程和活动的。SC-205委员编写DO-178C时，故意使其具有“非规定性”。这意味着改标准提供了描述“必须做什么”的一般目标，但没有解释“如何做”的例子。虽然这种方法提供了很大的灵活性，但也带来了很多不确定性。此外，DO-178C还增加了四个新的补充(即单独但相关的文件)，处理软件中使用的特定技术：其中，DO-330用于工具资格认证，DO-331用于基于模型的开发，DO-332用于面向对象技术，DO-333用于形式化方法。

系统安全评估过程中，软件的安全性是根据与软件组件相关的故障条件在系统级确定的。这些安全状况可分为五类:“即灾难性的，危害性的，严重的，不严重的，以及没有影响的[25,26]。然后DO178C指南定义了五个不同的保证级别，它们与上述故障条件的分类相关(级别A到E，其中级别A是最高的，因此需要最严格的流程)。每一层的软件保证都与一组目标相关联，这些目标大多与基本的生命周期过程有关，DO-178C标准针对9个软件过程提出了71个安全目标，以保证航空软件以规范的方式进行开发，并以规范的方式排除软件缺陷，修复代码漏洞，从过程上保证软件最终达到所需的安全性[1]。

软件生命周期是指软件产品从概念提出到产品“死亡”的整个周期。就机载软件而言，DO-178B/C规定的软件生命周期过程包括：软件计划、软件开发以及软件综合过程三个过程。如图4所示，计划阶段指导软件开发和集成过程的四个过程。每个步骤都包含自己的计划来指导它，的目标和活动，自己的检查来验证一切，以及自己的工件来显示遵从性的证据。软件开发阶段包含四个过程：软件需求过程提供了软件设计团队将在程序中实现的软件需求。软件设计开发软件架构和较低层次的需求。软件编码执行实际的软件编码，通过定义的体系结构将SW需求转化为设计。软件集成过程编译源代码，然后链接并将其加载到目标计算机系统。综合过程包含了四个子过程。软件验证过程执行评审、分析和测试的组合，以确保软件设计实现其需求(仅此而已)。配置管理过程提供了适当的数据控制和复制。质量保证确保项目的开发过程符合生命周期过程目标，审定联络 ，审定人员 定期审查计划、过程和工件，以确保遵从性。在整个软件生命周期中，软件计划过程为软件开发和集成过程提供指导，软件开发过程是主线，软件综合过程与软件开发过程同时执行，并且软件综合过程执行在软件开发过程的各个子过程上，具体的关系如图5所示。



图4 DO-178B/C生命周期过程结构图



图5 DO-178B/C软件生命周期方阵图

DO-178C准 出了22项软件生命周期数据（Software Life [cycle](https://cloud.tencent.com/developer/information/cycle) Data）。需要说明的是：这并不是规定要有22份文件，这些文件可以是独立的，也可以合并处理。这22项软件生命周期数据是软件产品的载体。具体见第3章。软件生命周期数据的形式可以多种多样，但要便于高效地检索或走查。软件生命周期数据还应具有确定性。完整性 一致性 可更改性，可追溯性等特征。软件生命周期数据是软件研制整个生命过程中所有数据的记录。它应该完整地、真实地记录了软件的整个生命过程：计划过程的输出、高级别需求的捕获、设计描述以及每一次需求的变更、每一次代码的改动、每一个测试用例、每一次版本升级、每一次同行评审，都要真实地记录下来，纳入构型控制之下。

软件生命周期的过程、数据和目标是DO178C的基本要素，它们的关系如图6所示：目标是在过程中实现的，目标的实现通过软件生命周期数据得到证明；过程使用生命周期数据，也生成生命周期数据。



图6 DO-178B/C中基本要素的相互关系

适航符合性验证是为了保障民用航空活动安全，采取不同的说明和验证方法获得所需要的证据资料向审查方表明产品对于适航条款的符合性。在标准领域，为了实现DO-178C标准的认证，美国联邦航空管理局（FAA）提供了软件审批指南（Software Approval Guidelines），帮助适航审定部门审定适航软件是否符合DO-178C标准。该指南对DO-178C中规定的活动以及输出进行了更为详实的解释说明，指出了其中需要关注的重点，为建立目标满足性论证模式提供了重要的指导。由于软件生命周期数据完整地、真实地记录了软件的整个研制过程，它就成了软件适航审定的举证依据。审查方通过评审研制单位提供的软件生命周期数据，来判断机载软件的研制过程是否满足DO-178C相应软件等级软件的目标，从而判断该机载软件能否能够获得适航批准。

## 配置管理过程关键技术

为什么变更管理能够代表了配置管理的关键技术？

由于变更管理中，必然涉及到了基线管理、标识配置、所以建立变更管理的模型。基线的建立是基于变更的，而进行标识配置是为了基线管理和变更管理服务。

包含了配置标识活动以及基线管理活动的，以变更管理为中心的配置管理模型。能够反映配置管理过程的关键技术环节

## 信息检索技术

信息检索技术是一个比较老的概念，指的是根据给定的查询项在一个数据集合上查找出相似的、相关的内容[23]，最后将结果按照相关度排列。对于信息检索，大部分的工作是基于文本检索，即查询项和待查询集合的内容都是以文本形式存储，主要方法可以分为两类：基于统计检索和基于语义检索。

### 基于统计检索

基于统计的方法是根据查询语句和待查询集合的内容的统计结果来确定查询结果。将查询和待查询集合放在同一个统计空间中，对文本进行分词，将所有的内容基于统计信息进行表示，然后计算得到与查询项相关的内容。当前常用的基于统计检索的方法包括向量空间模型和概率模型。

向量空间模型把文本信息用向量空间中的点表示，然后计算向量之间的距离来量化两个文本之间的相似度，最后通过对计算结果排序得到查询结果。文档向量构造最简单的方法是使用词袋模型（Bag of Words，BOW），即文档向量的长度是数据集中所有的词的数量，文档向量每一维的值为当前词在该文档中的出现频率或其他的一些特征，从而将一个文档表示为一个向量。在计算文本间距离时，可以使用余弦公式等方法，得到两个文本之间的相似度值。还有一些方法通过对关键词加权重的方式表示文本向量[24]。

潜在语义索引也是基于向量空间的检索方法。LSI使用one-hot方法表示文档，并将所有文档放在同一个矩阵中，通过奇异值分解（Singular Value Decomposition，SVD）的方法对矩阵降维，并且得到文本的主题和主题之间的关系。但是由于使用向量表示之后矩阵规模巨大、同时矩阵会过于稀疏，因此需要进行降维操作。LSI方法虽然表示和计算过程都比较简单，但是也有如下几个缺点：（1）通过SVD生成的新矩阵解释性较差；（2）无法解决“一词多义”的问题；（3）和词袋模型一样，忽略了文章中单词的先后顺序。

概率模型基于多个随机变量之间概率关系计算查询项和候选文档之间的关系，将查询项与文档集合中的文档同时出现的概率定义为二者的相似度，形式化表示如公式(2.1)所示。在计算完成后，对相似度进行排序，然后设置阈值过滤得到某个查询对应的相关内容，从而完成信息检索任务。

（2.1）

其中表示文档集合中第*i*个文档，表示查询语句，公式(2.1)使用贝叶斯公式对二者同时出现的概率进行求解。

### 基于语义检索

基于语义的方法是在一定程度上通过对查询句子进行句法和语义的分析，即通过其他的表现形式对文本的语义进行表示。由于当前大部分信息检索都是基于文本检索，因此基于语义的信息检索也和自然语言处理技术相关。使用句法分析的方法是借助文本的句式和一些词性信息对文本语义进行强化处理，如在文献[25]中提到的使用子树加速对文本相似度的计算，并将子树与文本特征向量进行比较匹配，从而得到更准确的文本相似度；使用语义分析的方法有基于本体[26]和基于外部语义词典（如Hownet[27]和Wordnet[28]）等方法，这些方法基本是将查询文本中的单词或词组抽离出来，然后用本体或者外部语义词典中的含义来表示，从而将整个文档集合中的内容统一表示，然后使用给定的算法计算相似度，从而确定检索列表。

## word embedding介绍

Word2vec是Google的Tomas Mikolov[29]等人提出的文本表示方法，将单词表示成word embedding的形式，是一种分布式表示模型。Word embedding的向量表示反应了词的共现关系。在给定的语料库下，通过算法模型将词表示成向量的形式，向量的维数可以由用户指定，通常为几十到几千不等，因此可以远远小于词典的大小，这样就避免了向量空间模型中的矩阵稀疏问题。同时，Mikolov等人还开发出了训练词向量的工具，训练效率非常高，非常适合在现在大数据集的应用中。

Word embedding基于分布式假说提出，即出现在相同上下文中的词的含义相近[30]。训练word embedding的流程类似前馈神经网络，但是仅包含输入层、隐藏层和输出层三层结构，并且由于Mikolov等人对隐藏层和输出层做了优化，使得模型的训练效率更高。训练模型主要包括skip-gram和CBOW两种，skip-gram模型的输入是特定词的词向量，输出是特定词上下文相关的词的词向量，而CBOW和skip-gram的输入输出相反。skip-gram和CBOW模型分别如图8和图9所示。



图8 skip-gram模型示意图



图9 CBOW模型示意图

## 软件工程领域学习排序算法

学习排序（Learning to Rank， LtR）算法作为一种监督或半监督的机器学习算法[31]，在信息检索、数据挖掘和自然语言处理等领域起着重要的作用。在软件工程领域，学习排序算法经常用在故障定位和数据重复检测等方面。对于每个查询与候选文档的组合，我们能够抽取出若干特征，比如词汇相似度、语义相似度等，作为机器学习模型的输入。学习排序的流程如图10所示，更加形式化的表示如公式(2.2)和公式(2.3)所示。

(2.2)

(2.3)

其中表示查询语句，表示相应的候选文档集合，函数将查询-文档对映射为特征向量，是表示每个特征向量的权重矩阵，是排序方法。排序方法有三种类型：单文档方法（Pointwise）、文档对方法（Pairwise）和文档列表方法（Listwise）。

单文档方法的处理对象是单独的一篇文档，将其转化为特征向量后，该问题转化为分类或回归问题，可以使用等级回归或分类算法来解决问题。该方法只考虑单个文档的绝对相关度，忽略了文档间顺序关系。文档对方法则考虑了文档间的关系，把任意两个文档组成的文档对（文档A，文档B）作为机器学习的输入，最后得到文档A是否应该排在文档B的前边，此时学习排序算法可以转化为二分类问题。该方法通过考虑两两文档对之间的关系进行排序，因此比单文档方法的效果有所提升。文档列表方法的处理对象是每个查询所对应的所有搜索结果的列表，该方法直接训练优化算法模型输出的整体序列，因此其结果能够更接近真实的文档序列。已有的且常用排序算法如表1所示。



图10 学习排序的流程

除了排序算法，选择合适的特征作为模型的输入对学习排序算法的性能也有很大的影响。在软件工程领域，大多数使用学习排序算法的应用都选择词或者文本的相似度作为其中一个特征，比如文献[32]中作者使用了词表面信息相似度、通过API增强的词相似度和类名的相似度，文献[33]中作者选择文本相似度和上下文相似度作为学习排序算法的特征。值得一提的是，在计算以上相似度时，作者们都是用了余弦相似度。除此之外，特征可以分为两类：依赖查询的特征和不依赖查询的特征[33]，其中依赖查询的特征即与查询文本相关的特征，如相似度特征；不依赖查询的特征则与查询文本无关，仅仅反应候选结果的特点，如候选结果的词频[33]、故障修复频率[32]等。同时，在选择特征时，特征数量不应过大或过小，过大会导致过拟合问题，过小则会降低结果精确度。

表1 排序方法的常用算法

|  |  |
| --- | --- |
| **排序方法分类** | **常用算法** |
| 单文档方法 | 回归：Subset Ranking[34] 分类：SVM、McRank[35] |
| 文档对方法 | IR SVM[36]、Ranking SVM[37]、RankBoost[38] |
| 文档列表方法 | ListNet[39]、AdaRank[40]、SVM Map[41] |

## 本章小结

本章首先介绍了适航领域中的DO-178B/C标准中对于软件生命周期的定义、标准中各个基本要素的关系以及标准中对于软件需求跟踪的一些严格的约束；然后介绍了信息检索领域中常用的技术，包括基于统计检索方法和基于语义检索方法；接下来介绍了word embedding的概念和两种训练模型的区别；最后对软件工程领域的学习排序算法的应用以及常用排序算法进行介绍。本章通过对适航领域软件特点的分析和需求跟踪技术研究中相关的技术的介绍，为下一章适航领域软件需求跟踪算法模型的构建提供了理论基础和支撑。