**北京航空航天大学计算机学院**

**硕士学位论文开题报告**

**论文题目**：适航领域软件需求跟踪方法研究与应用

**专 业**：计算机技术

**研究方向**：计算机软件与理论

**研 究 生**：赵腾

**学 号**：ZY1506321

**指导教师**：曹庆华

**北京航空航天大学计算机学院**

2016年12月15日

目 录

1 论文选题的背景与意义 1

2 国内外研究现状及发展动态 1

2.1 需求跟踪技术的国内外研究现状 1

2.1.1 静态跟踪技术 1

2.1.2 动态跟踪技术 2

2.2 需求跟踪技术的研究趋势 3

3 论文的研究内容及拟采取的技术方案 4

3.1 论文研究内容 4

3.2 论文拟使用的技术方案 4

4 关键技术与问题难点 6

4.1 标引词提取 6

4.2 聚类模型的选择 7

4.3 可视化内容的选择 7

5 论文研究计划 7

5.1 时间安排 7

5.2 预期成果 7

6 主要参考文献 8

**适航领域软件需求跟踪方法研究与应用**

# 论文选题的背景与意义

随着计算机技术在航空领域的应用，航空机载电子设备的发展迅速，同时软件起到的作用也越来越大，因此，软件的适航性成为飞机飞行安全的重要的因素。随着软件规模越来越庞大，软件需求跟踪作为保证软件开发工作与需求保持一致性的重要技术手段，成为影响项目成败的重要因素。2011年航空无线电技术委员会（RTCA）在DO-178B的基础上提出了DO-178C，即机载系统和设备合格审定中的软件考虑，在标准层面上约束了软件开发的流程以及应该达到的目标[1]。

需求跟踪关系包括不同抽象等级的需求之间的关系、需求和其他阶段（设计、编码等）输出元素之间的关系。在适航领域软件开发需求包括系统需求、高级需求、低级需求，需求本身的关系比较复杂，文档数量比较庞大，因此能够自动化地建立需求跟踪关系是必要的。

# 国内外研究现状及发展动态

## 需求跟踪技术的国内外研究现状

需求跟踪技术分为两类：静态需求跟踪技术、动态需求跟踪技术。静态需求跟踪技术主要使用跟踪矩阵、跟踪图和交叉引用等多种形式表示跟踪过程，静态需求跟踪链静态表示；动态需求跟踪技术是在软件开发过程和需求变更时自动建立或维护跟踪关系，主要使用基于信息检索、基于事件触发、基于规则的方法[4]。

### 2.1.1 静态跟踪技术

传统需求跟踪主要使用静态跟踪技术，但随着软件规模越来越大、软件声明周期越来越长，静态跟踪技术易于出错、不易维护的缺点暴露出来，如果继续使用则成为软件维护的一个负担。

需求能力跟踪矩阵用来表示需求和其他系统元素的联系链，每个功能性需求向后连接一个特定的使用实例，向前连接一个或多个设计、代码和测试元素。设计元素可以是模型中的对象，例如数据流图、关系数据模型中的表单、或对象类。代码参考可以是类中的方法，源代码文件名、过程或函数等。跟踪能力联系链可以定义各种系统元素类型间的一对一，一对多，多对多关系。

需求跟踪图可以直观的表示需求之间的联系和需求和各个阶段元素之间的关系，以及在需求变更时对各个部分的影响。不仅支持软件需求到系统设计文档、系统相关的规格说明、实现代码之间的跟踪，还支持软件开发过程中的中间制品之间的跟踪。在跟踪图中，用户可以自己定义图中的对象和关系，方便用户使用自己熟悉的语言对关系描述。

交叉引用主要应用在需求文档之间建立跟踪关系，可以在系统的需求说明文档、软件需求规格说明文档等中间建立需求跟踪关系。此方法建立的跟踪方法较直观，便于实际使用，但局限性是只适用于需求文档之间跟踪关系的建立[5]。

### 2.1.2 动态跟踪技术

动态需求跟踪针对手工建立需求跟踪代价过高且易于出错的问题,以自动化技术为手段辅助开发人员建立和维护跟踪关系[4]。

软件生命周期内出现的大部分软件制品都包含着大量格式化与非格式化的文本信息因此主流的动态需求跟踪方法大多将软件制品视为待检索的文档，并计算不同软件制品之间的相似度,对于相似度高于一定阈值的软件制品则之间存在跟踪关系[6]。

[4]中介绍了一种基于句法分析的跟踪关系恢复方法，通过句法分析我们可以识别出最有可能刻画软件制品特征的部分动词与名词，并减少制品中存在的噪音对需求跟踪关系恢复过程所造成的不利影响。方法分为如下步骤：1）将制品中的句子切分为句子块（代码特殊处理）；2）词性标注；3）对句子进行块分析，利用句子的上下文来修正词性标注过程中可能引入的错误；4）减少标引词中存在的噪声；5）通过聚类簇映射为分类建立映射关系。该方法关键是使用语义聚类对得到的不同分类进行映射，然后通过得到的结果来建立跟踪关系。

[7]和[4]是同一个作者，[7]的是从命名实体识别的角度来恢复需求追踪关系，主要解决需求和源代码的跟踪关系。该文章提出基于文本的需求追踪方法严重依赖于文本质量[7]，因此，文中提出一种方法，通过代码实体上下文构建命名实体识别模型，解决了抽象语法树和正则表达式无法解析非源代码形式的软件制品问题。文中使用基于最大熵模型的需求追踪方法，将源代码中的命名实体识别看做分类问题，同时代码的类别包括：class、method、invocate、comment、param、normalText，建立分类模型后对代码进行分类；然后对上述已经找到命名实体的制品文本进行预处理，包括分词、过滤停用词和词干化处理，将软件制品转为文档集合，然后使用[4]中相同的聚类方法对标引词进行分类和映射，从而建立跟踪关系。

[6]提出使用代码注释辅助动态需求跟踪的方法，用来建立需求文档和代码之间的跟踪关系。文中指出动态需求跟踪运用信息检索等技术，自动化建立需求文档和工作产品的跟踪关系，在跟踪精度等方面仍然存在问题，文中认为问题一部分是出在没有利用注释信息上面。使用的方法和[7]相似，主要有两点不同：1）预处理数据阶段，使用自动翻译工具将中文的软件需求文档翻译成英文，因为文中认为代码都是英文，也省去了中文分词的步骤；2）使用向量空间模型对文本进行检索，得到文本和代码的相似度，达到一定阈值的即可建立跟踪关系。

上面介绍的是国内的一些工作，下面将介绍几个国外的工作。

[8]提出基于LSI（Latent Semantic Indexing），即潜在语义索引，重建需求和设计制品、测试用例之间的跟踪关系。LSI是信息检索技术中的一种，可以将信息降维，文中提出一旦所有的文档都表示为LSI的子空间，那么就能通过余弦相似度方法计算文档间的关系。同时指出LSI不依赖于提前定义的单词表或语法，也就是说，可以省去花费巨大的数据预处理过程。

[9]是比较早提出通过信息检索的方式来增强需求跟踪的文章。文中主要工作包括：1）将问题构造成信息检索问题；2）选择IR算法；3）准备算法输入的需求文本；4）分析算法输出；5）选择合适的策略整理算法输出；6）比较算法性能和人工性能。比较重要的是，文中提出可能在这个过程中存在的问题：1）需求文档不全或语义模糊；2）有些缩略语没有定义；3）领域或工程只是缺少亲密性；4）高层需求和底层需求中使用不同的术语表示。这篇文章使用VSM算法，同时增添了两个扩展：一个是使用关键词列表，另一个使用简单的词典。

## 需求跟踪技术的研究趋势

目前，动态需求跟踪关系的自动化构建方向尽管很早就有人在做，但是在模型构建方面还是不够成熟，同时在具有相当复杂软件需求文档的适航领域的工作基本没有，未来还有很多工作可以进行探索。

在适航软件需求跟踪方面，由于涉及较多的航空航天术语，并且很多资料都涉密，所以该方面的需求跟踪应该与普通软件有所区别。首先，语料库的建立应该遵循军方标准，同时符合保密性和安全性要求；其次，与正常软件开发不同的软件生命周期会产生与正常软件不同的数据，在建立跟踪关系时的链接表会有较大区别。

当前，动态需求跟踪技术的自动构建主要集中在需求和源代码之间的联系建立，而复杂需求之间的关系建立案例较少，以后随着软件复杂性的提升，或者不同应用领域的发展，复杂需求间跟踪关系方面的工作会增长。

# 论文的研究内容及拟采取的技术方案

## 论文研究内容

本论文总目标为基于适航领域软件开发各个阶段文档，利用文本挖掘的技术提取需求中关键概念和属性，建立需求跟踪关系模型，研究动态需求跟踪关系自动构建方法并实现，同时编写可视化工具，能够动态显示需求跟踪关系。

因此，针对这一总目标，细化为以下几个研究内容：

1. 结合适航审定标准和软件生命周期文档，研究适航领域文档与普通文档的区别。
2. 研究基于信息检索的需求跟踪关系恢复方法，基于软件结构提出改善精度的策略。
3. 研究适航领域标引词聚类算法，将其作为跟踪结果的增强策略。
4. 结合适航领域特点，对软件需求和其他制品的跟踪关系进行不一致性分析。
5. 建立原型系统，对需求跟踪关系进行可视化，并对适航软件开发生命周期的数据进行评估，提出建议。

## 论文拟使用的技术方案

本文拟按照如下方案开展研究工作：

1. 从614所收集数据和预处理

在中航工业航空动力控制系统研究所（614所）收集某一型号发动机控制软件开发需求文档、设计制品和源代码等数据。对文档数据，在不破坏原始语义的情况下对数据进行脱密处理，同时对非结构化的数据进行人工或半自动化的结构化处理；对代码数据，进行自动化标注，标注代码类型，类似[7]中代码处理过程。

1. 采用关键词提取技术准确抽取标引词，并标注标引词重要程度

由于软件架构的分层设计以及各个需求优先级的差别，我们要根据文档结构，并分析句子的语义，获取概念的重要程度。为了完成这项任务，可以采取关键词抽取技术[10]。关键词抽取的难点在于如何评估单词在文档中的影响力，基本的方法是使用权值来识别词语的重要程度。最传统的关键词抽取方法很依赖词频，例如经典的TF-IDF算法, 文献[11]中的KeyGraph以及文献[15]中基于卡方检验的方法都依赖于词语共现性和词语频率。当然还有一些其他的算法，但基本是基于出现频率估计重要度，区别在于计算哪些频率以及关联频率。

针对适航领域，考虑建立领域知识库（项目词典），知识库中预先标记一些制品的重要程度，同时知识库还包括中英文符号表、术语定义、缩略词表、同义词表等，结合通用词典，使用这些先验知识和上述提到的频率计算方法来得到关键词的重要程度。

1. 使用LSI建立软件需求和其他制品之间粗略的关系，通过软件文档层次结构和重要程度提升LSI准确度

LSI（Latent Semantic Indexing，潜在语义索引）是信息检索的一种模型，其假设文本中的词与词之间存在某种联系 ,即存在某种潜在的语义结构 ,所谓语义结构是指自然语句中存在的语义范畴和语义关系所形成的抽象的语义格式。通过奇异值分解对检索项-文本矩阵进行降维，通过较少的、关键的向量来表示原文档或其他制品，因此其可以过滤噪声、节省空间，同时提高效率。通过该方式，可以建立起软件需求和其他制品之间的相似度由高到低的列表。

LSI忽略了不同层次软件文档、不同代码结构中关键词的重要程度的不同，因此在计算时将所有同义词赋予了相同的权重。同时，同一个文档中的文档层次结构也决定了文档内容重要程度的不同。因此，在本文的工作中，还将分析软件文档和代码的层次结构，来提升LSI结果的准确度。对软件文档而言，得到不同文本段中标引词的重要程度系数；对代码数据而言，根据类的多种关系或者函数之间的调用关系得到不同类和函数的签名的重要程度系数。

1. 对标引词进行聚类，作为增强策略

假设某一软件需求和某一其他制品之间存在跟踪关系，那么如果对他们中的标引词进行聚类，二者一定会出现在同一个聚类中。因此，可以通过聚类的方式，在语义的角度对软件需求和制品进行重新分析，将得到的新的跟踪关系加入到最终结果中。将聚类方法作为一个增强策略，进一步提升结果准确率。

聚类方法较多，基本包括：K均值、凝聚层次聚类（HAC）、DBSCAN、簇评估等。聚类之后的结果大致如图 1所示，图中同一颜色的点表示同一个文档中的标引词，同一个虚线框中的点表示同一个分类的标引词，离散的点表示没有被聚类的噪声点。



图 1 聚类结果

1. 使用多种可视化方式显示软件文档之间的关系

可视化的形式有多种，如：标签云、文本地图、TreeMap、ThemeRiver等[12]。其中文本地图是根据文本和聚类的相似度决定点与点、点集合与点集合的距离，相似度越高，它们在平面图上距离越近；TreeMap在绘制出最外层的矩形即树结构的根节点后，递归地细分内部空间成矩形块，即每一层节点的子节点都递归地划分各自的父节点，各个矩形块的面积由各自的权重比例决定。

工作中将使用文本地图显示动态需求跟踪关系，向开发和维护人员提供一种直观显示各个软件需求点和软件文档之间的跟踪关系；各级需求之间使用TreeMap的形式，将分层需求直观显示，并根据跟踪关系得到不同抽象层级需求之间的联系。

# 关键技术与问题难点

## 标引词提取

在适航领域，由于保密性、安全性的限制，当前很多关键词提取工具中提供的语料库都不会包含我们的软件文档中涉及的词汇，同时，我们也不能直接使用保密信息直接在网络上操作，因此这给标引词的提取带来了困难。由于缺少先验知识，在标引词的重要程度标注上也会遇到问题。

因此，标引词提取之前，首先手动或半自动化地建立该领域涉及到某型号发动机控制软件的知识库，同时，由于软件文档和代码使用中英文混合的方式，因此应该建立领域的中英文对照词典和缩略词表。在上述基础上，结合通用词典和现有技术，提取标引词。

## 聚类模型的选择

由于聚类方法种类繁多，在该领域不同聚类算法得到的效果肯定也不同。通过查找文献，之前还没有对于适航领域软件文档中的内容进行聚类分析的工作。因此，具体哪一种算法对该工作行之有效尚不可知，必须通过大量的实验进行测试。

工作中将对比多个聚类算法，并对适合该领域文本的算法进行改进，寻找能够达到最好的效果的算法。

## 可视化内容的选择

适航软件开发的过程要求较为严格，因此文档内容数量庞大，全部显示一定比较乱，反而失去了可视化的意义，因此有必要设置显示的内容以及显示的优先级。

在软件适航方面，有DO-178C、DO-333和Order 8110.49chg1等多种标准对软件开发过程和整个生命周期进行约束[1]，因此应该在仔细研读这些标准的基础上，然后结合软件开发人员在编写文档、编写代码时的习惯，对应该可视化的内容进行选择。

# 论文研究计划

## 时间安排

2016年12月 至 2017年2月 进行数据整理，阅读相关文献，深入学习并分析所需技术方法。

2017年3月 至2017年4月 研究并实现具体的适航领域软件需求跟踪技术，并与已有算法进行对比实验。

2017年5月 至2017年6月 完成适航领域需求跟踪关系恢复和重建，撰写阶段成果小论文。

2017年7月 至2017年9月 编写可视化工具，方便维护人员使用。

2017年10月 至2017年12月 整理工作，撰写毕业论文，准备毕业答辩。

## 5.2 预期成果

1. 完成所有研究目标，实现软件需求跟踪关系可视化工具
2. 发表学术论文一篇
3. 完成硕士学位论文

# 主要参考文献

[1] 沈小明, 王云明, 陆荣国. 机载软件研制流程最佳实践适航标准 DO-178B/C 研究[J]. 2013.

[2] Gotel O C Z, Finkelstein C W. An analysis of the requirements traceability problem[C]//Requirements Engineering, 1994., Proceedings of the First International Conference on. IEEE, 1994: 94-101.

[3] 裴志远. 基于本体的软件需求跟踪的研究[D]. 合肥工业大学, 2011.

[4] 王金水, 翁伟, 彭鑫. 一种基于句法分析的跟踪关系恢复方法[J]. 计算机研究与发展, 2015, 52(3): 729-737.

[5] 王慧玲. 支持动态需求跟踪的构件关联机制研究[D]. 湖南大学,2014.

[6] 董刘, 李引, 李娟. 基于注释改进动态需求跟踪的方法[J]. 计算机工程与设计, 2009 (1): 113-115.

[7] 王金水, 薛醒思, 唐郑熠. 一种基于命名实体识别的需求跟踪方法[J]. 计算机应用研究, 2016, 33(1): 132-135.

[8] Lormans M, Van Deursen A. Can LSI help reconstructing requirements traceability in design and test?[C]//CSMR. 2006, 6: 47-56.

[9] Hayes J H, Dekhtyar A, Osborne J. Improving requirements tracing via information retrieval[C]//Requirements Engineering Conference, 2003. Proceedings. 11th IEEE International. IEEE, 2003: 138-147.

[10] Timonen M, Toivanen T, Teng Y, et al. Informativeness-based Keyword Extraction from Short Documents[A].KDIR[C]. 2012: 411-421.

[11] Ohsawa Y, Benson N E, Yachida M. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor[A].Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on[C]. IEEE, 1998: 12-18.

[12] 唐家渝, 刘知远, 孙茂松. 文本可视化研究综述[J]. 计算机辅助设计与图形学学报, 2013, 25(3): 273-285.