## 1. 研究目标

本课题将综合运用SSM（Spring+SpringMVC+Mybatis）框架技术、jQuery的AJAX技术和**相似度检测方法**等，实现理论与技术相结合的科技项目管理系统，充分利用大数据时代数据带来的有用信息。其研究目标分为两大部分：1）基于SSM框架的科技项目管理系统的开发，

为了解决现有的科技项目重复立项问题，帮助审查人员做出科学的评审决策，确保科研经费的有效利用，推动我国科技的快速进步。课题拟将改进的基于向量空间和图模型的相似度计算方法应用于科研项目的相似度计算中去，通过更深层次对科技项目申请书包括项目名称、项目内容、核心团队、可行性报告等多方面整合，进行相似度计算的研究，改善目前的项目审查模式，对科技项目进行科学的分类管理，在立项之初找到与其相同或几乎相同的项目。以**“一种领域适应性的跨领域未登录词识别”“改进的基于领域自适应中文分词研究”“基于领域本体和向量空间的科技项目相似度计算方法”**为主要的理论研究目标。通过“理论研究、算法设计、实验验证、系统设计”的技术路线，验证理论成果，优化算法性能，形成理论框架，并以某市科技项目管理平台中的项目申请书为例，将所提出的模型与算法应用于其中，设计并实现科技项目管理系统。

## 研究内容

国内外对科技项目管理系统的开发已有大量的研究，在某些性能已取得很好效果，但还有待进一步提高。随着，近年电子商务、社交网络等大型互联网应用的高速发展导致互联网数据极速增长，大数据的概念被逐渐引入到大众的视线。而数字文档（如项目申报文档，论文文档，网页等）呈几何级数增长，同时，由于其本身的易复制性，导致项目重复申请等不良现象频频出现；大量相似文档的存在和数据孤岛数量不断的增加，也降低了信息检索的效率和精度。以此我们引入**基于科技项目的领域自适应的中文分词，试图在前人研究的基础上，将科技项目中跨科技计划的未登录识别引入到领域自适应的中文分词算法中，提出改进的基于领域自适应中文分词研究**，并将该算法结合SolrCloud检索，采用高可靠性分布式方案，应用于科技项目管理系统，改善海量数据运行时间长、处理效率慢和文本相似度检测等问题。

**主要研究内容如下：**

1. **一种领域适应性的跨领域未登录词识别**。首先，研究现有的未登录词识别算法，其识别方法可分为基于规则的方法、基于统计的方法、规则与统计相结合的方法以及其他的改进方法。尝试在现有的未登录词识别算法理论中提出一种领域适应性的跨领域未登录词识别算法使其适用于科技项目申请书中。最后利用广州市科技业务平台提供的科技项目申报书作为实验数据进行实验验证，同时实验选择准确率（Precision）、召回率（Recall）和F值等指标作为未登录词识别结果的评价标准。
2. **改进的基于领域自适应中文分词研究**。整理现有的基于领域自适应中文分词模型研究现状，提出一种改进性的基于领域自适应中文分词算法，并与传统算法进行比较研究。同时根据实验结果选取准确率、召回率和F值作为评估指标，以验证所提算法的有效性。
3. **基于领域本体和向量空间的科技项目相似度计算方法**。由于传统的基于向量空间模型方法没有考虑词语之间的关系和句子的结构等信息，只是简单的把文档分为单个的词语，缺乏对句子语义的理解。本文尝试将基于向量空间模型的方法与基于领域本体的方法相混合成新算法进行相似度计算，最后我们将该方法结合solr，采用高可靠性分布式方案，形成基于SolrCloud分布式科技项目查重系统。
4. **研究将所提出的相似度计算方法应用于科技项目管理系统**。主要研究现有的中文分词与中文文本相似度计算的相关技术，并将上述研究提出的基于领域本体和向量空间的科技项目相似度计算方法模型应用于科技项目查重系统。

## 3. 拟解决的关键问题

基于前述研究内容，本课题研究过程拟解决的重点和关键问题在于以下3个方面：

1. 如何根据现有的未登录词识别算法提出一种领域适应性的跨领域未登录词识别？本课题通过研究现有的未登录词识别方法的优缺点、适用范围等作比较分析，研究提出一种领域适应性的跨领域未登录词识别算法，同时使其适用于现有的大数据时代下的文本领域。
2. 如何对现有的中文分词中跨领域分词带来的问题？这是本课题首要解决的问题之一。近些年丰富的文本领域环境下，中文分词的领域适应问题给中文分词任务带来了巨大挑战。所以本课题提出一种改进的基于领域自适应中文分词研究方法。
3. 如何将基于领域本体的计算方法和向量空间的计算方法相结合？基于领域本体的计算方法在当前的研究工作中也占有一席之地，主要是通过语义来计算相似度，而向量空间的相似度方法只是简单的把文档分为单个的词语，缺乏对句子语义的理解。所以本课题将两种算法相结合，应用于科技项目申报这一领域的相似度计算。
4. 通过使用solr平台进行科技项目索引库的建立算是否能有效提高运算效率？我们拟通过利用solr平台的SolrCloud实现分布式索引结构，通过构建索引库加快搜索到相似文本。