**项目需求分析**

**需求分析来源：**

1. 用户访谈

用户访谈是一种最基本的需求获取手段，通过与用户面对面的交流我们可以直接得知用户的具体需求，有利于我们更加直接、准确地获取需求。这次我们采访的对象为周围的青年大学生，因为当代大学生是非常庞大的一个群体，即使大家感兴趣的行文获取会不同但是大家在获取信息的渠道方面或多或少会有一些共性。

1. 用户调查

在进行用户防谈时，由于很多关键人员的时间有限，不易安排过多的时间或者项日涉及的客户面较广。不可能——一访谈。因此，就需要借助用户调杏的方法，通过精心设计要问的问题，然后下发到相关的人员手中，让他们填写，再从所填写的内容中获取系统的需求倍息，这样就可以克服上述的问题。

用户调查最大的不足就是缺乏灵活性，而且可能存在受调查人员不能很好表述自己想法的限制。

1. 文档考古

文档考古是指对历史存在的—些文档进行研究，从带有数据的文件、表单、报表等文档中获取所需信息的过程。对于一些数据流程比较复杂的、工作表单较多的项目来说，就可以应用这种方法。

1. 建立联合分析小组

在系统开发时，系统分析员和用户之间由于知识结构的差异，难免存在难逾越的交流鸿沟。用广提供的需求信息，在系统分析员看来可能是零散和片面甚至无法理解的。因此，为了能够减少交流上的问题，就需要一个领域专家来帮助进行沟通，即可以建立一个由用户、系统分析员和领域专家参加的联合分析小组来共同完成需求的获地。

1. 原型法

原型是在[软件开发](http://www.sytm.net/ruanjiankaifa/" \o "软件开发" \t "https://www.sytm.net/ruanjiankaifa/_blank)中被广泛使用的一种工具，在软件系统的很多开发阶段都起着非常重要的作用。原型法就是尽可能快地建造一个祖糙的系统，这系统实现了目标系统的某些或全部功能，但是这个系统可能在可靠性、界面的友好性或其他方向上存在缺陷。建造这样一个系统的目的是为了看，考察某一方面的可行性。如算法的可行性，技术的可行性，或考察是否满足用户的需求等。原型是在最终系统产生之前的一个局部真实表现，可以让人们能够对一些具体问题进行基于文物的有效沟通，从而帮助人们尽早解决软件开发个存在的各种不确定性。

1. 模型驱动

前面的面谈、原型、观察以及文档审查等方法可以通过执行一些具体的获取行为来对系统需求进行认知和理解。但是大多数软件系统，尤其是对于复杂的系统而言，它们的需求获取任务绝不是可以通道一两次这样简单的获取行为就能够完成的。为了能够使得获取行为相互配合、减少不必要的精力耗费和防止出现获取信息的遗漏，可以采用模型驱动的方法。

1. 基于上下文的方法

软件系统是作为一个整体存在的，它通过和环境的交互来解决用户的问题，满足用户的需求。软件系统中的每项功能都是依存于一定的背景和上下文环境，因此，要正确地理解系统的功能就必须要正确地理解它的背景和上下文知识。基于上下文的方法就是注重于系统的环境、开发组织的业务背景、涉众的特征以及目标等。与前面的方法相比，它更加注重用户在—定环境下表现出来的行为，通过分析用户的行为得到信息。

总之。进行需求分析时，应注意一切信息与需求都应站在用户的角度上，尽量避免分析员的主观想象，并尽量将分析进度提交给用户，让用户进行检查与评价，从而达到需求分析的准确性当然，在需求人员进行需求获取的过程中，往往可能是多种方法的结合，取长补短，从而达到更好地获得系统需求的目的。

**项目背景：**

新闻发展越来越快，每天各种各样的新闻令人目不暇接，对新闻进行科学的分类既能够方便不同的阅读群体根据需求快速选取自身感兴趣的新闻，也能够有效满足对海量的新闻素材提供科学的检索需求。新闻发展越来越快，每天各种各样的新闻令人目不暇接，对新闻进行科学的分类既能够方便不同的阅读群体根据需求快速选取自身感兴趣的新闻，也能够有效满足对海量的新闻素材提供科学的检索需求。文本作为信息的重要载体，对文本数据的提取与表示是解决文本信息管理问题的关键手段。其中，新闻数据作为日常最为常见的文本数据之一,对新闻的精准分类,有助于缩短查阅时间,提升阅读体验。文本表示是文本分类的基础，传统的文本表示主要基于统计计数。这种表示方法默认为单词与单词之间是相互独立的，这会导致文本语义信息的丢失;另外，提取的特征具有高纬度和高稀疏的特点，不利于文本信息的表示。伴随着深度学习提出与推广，通过层次神经网络模型，可以有效地完成从低层特征到高层特征的逐层提取融合，对于重要特征给予高比重权值。深度学习提出为文本的特征提取与文本分类提供了有力的理论与技术支持。本文通过分析新闻文本数据特点，构建了基于层次神经网络的新闻分类模型，主要工作归纳如下:  
 首先把文本根据需求设计相应的新闻采集爬虫，并对采集的新闻数据进行清洗、中文分词以及文本表示：在中文分词时，采用了基于python的分词工具，为了获得更好的分词效果，对分词词典进行未登录词扩充。考虑到传统文本表示方法通常会面临高维稀疏问题，为避免该现象对文本表示带来的影响，采用分布式词向量表示方法，使用语言模型将文本数据转化为低维实值向量，并对使用的语言进行模型分析对比，确定合适的语言模型以及词向量维度。在各层级特征提取提取过程中引入注意力机制以解决权值配比问题。最后选用标题与正文数据组合的另一目的在于减少模型计算量，使用标题信息为主干、正文数据为特征扩充，合理缩减文本数据处理量，进而提高新闻文本分类效率。

**研究意义：**

随着互联网技术的高速发展以及电子终端的广泛普及，信息数据呈现指数型增长，在自媒体的时代下，每个人都是数据的制造者。用户在社交网络、门户网站的参与交互行为，加速了数据的规模效应，实现了从数据贫乏到信息过载的转变。据有关数据统计互联网每天产生的数据量大约有800EB,在享受丰富数据便利的同时，也面临着数据过剩所带来的烦恼:如何从繁琐的数据中快速、精准的获取所需要信息。信息数据的有效的组织和管理成为解决问题的关键[”。高速增长、展示形式多样化是大数据背景下网络数据的新性。数据表现形式上主要有文本、图像、音频等几种表达方式。其中，文本类型的数据与表现形式相比，数据格式小，占有资源少，更易于数据的分享、传播;并且，其他形式的数据信息也可以转化为文本数据形式进行表达，文本的这一-特点，使文本数据在信息存储中占有重要地位。新闻数据作为文本领域的一个重要组成，也是日常最常接触的文本数据之一，对其数据合理类别划分有着重要意义:可以快速定位读者的兴趣点、缩短查找时间，提高阅读质量。

**现状跟发展趋势：**

根据2007-2008中国商业智能巿场发展报告统计:从全球范围来看,商业智能已经成为最具有前景的信息化领域;从国内来看,商业智能是企业信息化中最重要的组成部分。文本挖掘是数据挖掘技术的分支，它能够使人们从大量冗余的信息中迅速发现对自己有用的信息，并在一定程度上揭示信息与信息之间的关联，是一种重要的商业智能技术。典型的文本挖掘方法包括文本分类，文本聚类﹐概念/实体挖掘，生产精确分类﹐观点分析，文档摘要和实体关系模型等。

已经有多个通用的文本挖掘模型被提出，最早的是Feldman等提出的 KDT( Knowledge Discovery in Texts )框架。该系统有两个输入，一个是已经用关键字标注过的文档集合，另一个是以关键字作为节点的层次结构。一般来说﹐关键字层次结构是面向特定应用领域的﹐给出了某一应用领域涉及的主要概念之间的层次关系，是领域背景知识的一部分。知识发现模块在背景知识的支持下,对标注过的文本文档进行挖掘操作﹐得到模式通过知识表达模块表达出来。另外一个具有代表性的文本挖掘模型由Tan提出﹐它将文本挖掘分为两个阶段，即先将无结构文本表示为中间表示的文本求精阶段﹐以及从中间表示提取模式或知识的提取阶段。文本数据时以自然语言表示的，这些数据对计算来说是不可识别和分析的﹐不管哪种模型﹐从文本数据中抽取特征信息形成文本的中间表示是文本挖掘的基础，文本的中间表示应包含足够的信息来支持后续的文本挖掘操作。通常的特征提取步骤包括︰词语切分，词频统计﹐特征选择三部分。

而文本分类是文本挖掘课题中讨论最多的技术之一，它是在给定的分类体系下，根据已经掌握的每类若干样本的数据信息﹐总结出分类的规律性而建立的判别公式和判别规则﹐从而达到自动对新文档的归类识别。一般来讲，文本分类需要四个步骤:(1)获取训练文本集;(2)选择分类方法并训练分类模型;(3 )用导出的分类模型对其它待分类文本进行分类;(4）根据分类结果评估模型。目前有多种基于向量空间模型的训练算法和分类算法，比较著名的有KNN ，ID3 ,C4.5，贝叶斯分类法，SVM分类法等。

新闻的处理，是一种对事实的选择、安排、解释等意义化过程﹐作为文字传播的一种特殊形态﹐新闻在语言表达上有着较为明显的个性特征，如篇章短小精干、表达客观公正、语言准确简洁等。新闻数据作为一类

重要的文本数据，其重要性不言而喻。由于互联网的出现﹐新闻数据的发布、采集、存储变得相对容易。然而，如何从浩如烟海的

新闻中检索、分析感兴趣的新闻类型却是很头疼的事。新闻分类作为新闻文本挖掘的基础和关键技术有着及其广阔的应用前景。

**特色与创新点：**

特色：提供简单的可视化界面。能够输入单条新闻，输出新闻的分类，或者支持本地上传csv/xlsx文件，批量输入新闻，并输出新闻分类。

创新点

1. 可以优化检索速度、提升搜索体验。
2. 一词多义问题

**可行性分析：**

通常情况下，新闻分类是新闻发布前的一个重要步骤，按照新闻内容信息可以实现多层级粒度的划分，其中一级类别主要包含财经、体育、娱乐、科技、军事等十几个类别，二级类别数量可达上百种，主要是对一级分类的一个细化,如体育可以细化为足球、篮球等等，三级类别是对上一级类别的进一步细化。面对海量新闻数据，依赖于传统的人工手段对数据进行标注、分类已经无法适应爆炸式增长的数字信息的管理需求,而且，由于不同的人对新闻认知方面也偏差，分类效果也不近理想效果。新闻自动分类通过告知机器训练规则，进而实现新闻文本上数据的高效、精准分类，同时还可以极大地节约人工成本。自动文本分类作为一种高效识别的文本处理技术在众多领域得到了广泛的应用，在信息时代下，极大地提高了信息的处理速度，使信息的实效性有了保证。新闻分类主要包含两个步骤，文本表示与文本分类。文本分类(Text Categorization)作为自然语言处理领域中的基石，在解决文本信息格式以及很大程度上处理信息杂乱无章的问题有着很强的适用性。文本表示则是文本分类的技术支撑，文本非结构化特征以及文字之间组合的特殊语义含义使得计算机理解文本产生了很大的困难，因此需要对文本数据进行己有算法处理形式的转化。文本表示效果的好坏会直接影响自然语言处理任务的分类结果表现。目前在文本表示时主要基于词袋模型（Bag-of Words，BOW)或者向量空间模型(Vector Space Model VSM)，但这两种表示方法缺陷在于只进行词频信息的统计，而忽略单词与单词之间语义联系。基于这种浅层语义的文本表示会造成数据信息的流失，偏离实际语境含义，进而增加文本分类任务的难度，导致文本数据的分类没有足够的信息支撑，在进行主题的细化方面表现尤为突出。传统的文本表示方法，主要是停留在词汇级别的浅层语义信息的表示，缺乏深层的文档级别的语义的获取，具有片面性，所以在进行文本任务中，当准确率到达一定程度后，就很难有能一步的提升。在进行复杂文本分类方面，单纯的使用词串或者词袋模型，分类效果就更差强人意。而基于深度学习的模型分类则有很大不同，其通过仿生大脑对信息处理过程，构建出多层级神经网络，实现信息数据的逐级提取与表示，进而完成高层抽象特征抽取，底层信号到高层语义的抽象关系可以使用映射函数描述。为克服浅层网络中的缺陷问题，深度神经网络引入了多层非线性映射结构，通过对神经网络结构分析与设计使得模型具有良好的特征学习能力，并且通过训练少量的参数完成复杂的函数逼近。在新闻文本分类任务中，引入深度学习，可以改善文本特征的提取方式，使分类效果更为准确，从而让读者在进行查阅时拥有更好的阅读体验。

参考文献链接：<https://www.sytm.net/ruanjiankaifa/20131105133049.html>