多标签知识点标注系统需求分析

1：项目概述

1.1 项目背景：

在智慧教学系统中需要通过利用程序设计习题的知识点的类型信息向学生推送相关知识点的程序设计习题。由此老师需要对原始的未被标注上知识点的程序设计习题所包含的知识点类型进行标注后，输入智慧教学系统以便于习题推送。

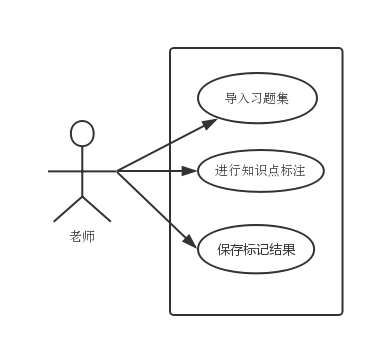
1.2 目标：

本系统的目标是老师向该系统输入未被标注知识点的程序设计题目的文本集合文件，系统对程序设计题目集合中的每条题目包含的知识点进行自动标注，然后将被标注的程序设计题目集合文本输出到指定文件。共有237种知识点类型，每条习题至少被标注上一种知识点，至多被标注上六种知识点。

1.3 用户特点：

该系统最终用户是老师以及其他需要对程序设计习题进行知识点标注的人员。该系统的使用人员需要具备一定的计算机科学知识，以便于对程序设计习题能有所甄别。

用例图如下：



1.4 假定和依赖：

本项目是否能成功实施，主要取决于一下几点：

（1）为了项目的开发下的条件和实施，在必要时对现有的业务流程进行合理的调整。

（2）学校为调研、开发和实施过程提供必要的工作环境和系统运行环境，这些环境有助于软件。

（3）学校为软件系统的运行提供必要的且能够满足系统运行条件的硬件环境和通讯环境，不合适的硬件环境和通讯环境将会影响系统的性能。

2 : 具体需求

2.1功能需求：

2.1.1 输入：输入数据的方式为向系统导入包含习题集的文本文件。输入分为三部分。

(1)首先系统需要通过输入训练集（已经被人工标注过的程序设计习题集文本文件），以便于系统中的算法对其算法模型进行训练。

（2）系统分类算法训练完成后。通过向系统输入测试集来对系统的标注效果进行评估。

（3）向系统输入未被标注的程序设计习题集文本文件，使系统对其知识点进行自动标注。

2.1.2 输出：

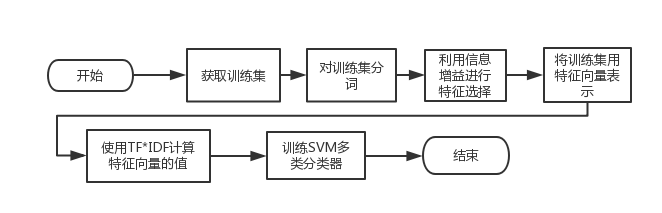
输出将按照训练集所规定的数据格式将标注后的程序设计习题数据集保存到用户指定的文本文件中。其中每行代表一条习题。习题之间以换行符号隔开。

2.1.3功能描述：

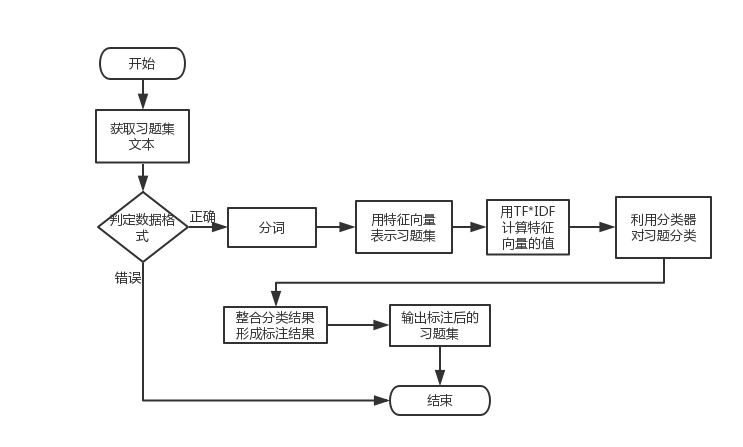
由于需要通过利用程序设计习题所包含的知识点的信息进行习题的推送。但是人工对习题的知识点进行标注费时费力，无法保证其标注结果的一致性。因此需要应用多标签学习的方法来实现对习题集知识点的自动标注。

2.1.4 系统的处理流程：

(1)多标签分类器训练流程：



（2）对未标注程序设计习题集进行知识点标注流程：



2.2 性能需求：

为了保证系统能够长期，安全，稳定，可靠高效的运行，多标签知识点标注系统需要满足以下性能需求：

（1）系统处理的准确性：系统处理的准确性是系统的核心性能。必须保证系统对程序设计习题集进行知识点标注的准确率。由于标注的准确率是系统的核心，所以在训练分类器的时候必须保证训练数据的正确性和代表性，同时也要保证训练数据数量足够分类算法进行分类器模型训练的需要。

（2）系统的时效性：系统处理的时效性是系统的必要性能。必须保证系统在面对大量输入数据的时候能够保证其在合理的时间内完成对数据的标注并进行输出。因此在对底层分类算法的选取上必须选取其时间复杂度相对较小的算法以保证系统的时效性。

（3）系统的易用性：多标签知识点标注系统是直接面向老师的。而老师平时的工作量较大，任务繁忙。所以就要求系统能够提供良好的用户接口，易用的人机交互界面。要实现这一点，就需要系统尽量使用用户熟悉的术语。针对用户可能出现的使用，要提供足够的帮助文档，缩短用户对系统熟悉的过程。

（4）系统的易维护性：多标签知识点标注系统中所涉及到的数据是非常重要的信息。系统需要将在对未标注程序设计习题集进行知识点标注流程中的每个环节中的原始数据和产生的中间数据进行即时的备份。以便于系统崩溃时进行数据回复以及当标注结果出现较大偏差时对系统进行分析等工作。

2.3 数据需求 : 数据字典

(1)训练集和测试集

描述：人工标注好知识点的程序设计习题集文本文件，每条数据占据一行。训练集共包含1000条数据，测试集共包含593条数据。

数据格式：如例子所示：每条习题被标注为6种知识点，如其中 93,79,32,13,0,0为该条习题所包含的六个知识点,后边的数据是习题的具体内容。非0的数值代表237个知识点中的序号,0表示该位置无知识点，即该条习题包含93,79,32,13共4个知识点。

例：“93,79,32,13,0,0 以下选项中合法的用户标识符是(B)A.long B.\_2Test C.3Dmax D.U.dat”

(2)未标注程序设计习题集

描述：需要进行知识点标注的程序设计习题集文本文件。每条数据占据一行，数据之间以换行符分隔。

数据格式：每条习题数据为一段文本。如：

“（ ）是C语言提供的合法的数据类型关键字。 A)Float B)signed C)integer D)Char”

(3)知识点

描述：所有的知识点类型数据的索引和知识点名称的对照文本文件。共273种知识点。

格式：索引 ： 知识点名称。如： “1 ： c语言”

(4)程序设计习题集文本数据的特征向量表示

描述：在系统对习题集进行知识点标注的过程中。需要将习题的文本表示方式抽象为以特征向量表示的方式，并需要计算特征向量的值。

格式：如：

原始文本：“安徽合（肥）巢（湖）芜（湖）高速公路巢芜段日前竣工通车并投入营运。合巢芜高速公路是国家规划的京福综合运输网的重要干线路段，是交通部确定１９９５年建成的全国１０条重点公路之一。该条高速公路正线长８８公里。（彭建中）”

向量表示：“20:-3.152736022363656 101:-3.6686767467964168 150:-2.5952547069568657 195:-4.379523504455763 215:-7.16703787691222 325:-3.258096538021482 341:-2.7212954278522306 369:-1.5686159179138452 415:-4.433194921248281 416:-2.70805020110221 437:-4.767289035464526 463:-3.0633909220278057 535:-1.6094379124341003”

3：多标签知识点标注系统的运行要求

(1)操作系统：windows7及以上

(2)运行环境：jdk1.7及以上

(3)数据库管理系统：My SQL

(4)硬件要求：Pentium 133以上，516M RAM