基于 UML 的句子相似度计算小程序的设计

江阿古丽·哈依达尔　郭　玲
（昌吉学院，新疆 昌吉 831100）

摘　要　本文通过采用软件工程中提出的面向对象软件设计方法论，使用 UML 统一建模语言的图形符号和基本概念，在 Rose 建模软件中以工程化的形式设计出在研究基础领域使用的计算句子相似度的小程序，并创建系统模型。该设计方案可以在各种 APP、微信小程序和网站的搜索、查找模块中考虑使用，在黏着性语言类的自然语言处理研究领域具有较好的应用价值。

关键词　UML 模型　软件工程　句子相似度
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1007-0745(2022)07-0037-03

1 句子相似度计算小程序的研究意义

以前的研究工作中，从词性、词序、句长、相似单元角等几个方面提出了哈萨克语句子相似度的计算方法，另外初步研究了切分名词词根和词缀的有限状态自动机的构造[1]。随着 python 程序设计语言的广泛使用，目前也可以采用 numpy() 函数库、字符串和有序组合数据的内置函数等有关技术方法，从数据库中可以筛选出相似句子，并按相似比例的高低进行排序。粘着性语言类的词根和词缀具有较特殊的主从关系，因此，计算相似度前实现分解句子和切分词缀是必不可少的。本文主要探讨的小程序的详细设计方案对计算句子相似度技术，甚至对机器翻译、搜索引擎等领域的研究工作提供必要的技术条件。

2 句子相似度计算小程序的可行性研究

句子相似度计算小程序作为验证新的计算方法和其规则而推出的测试小系统，可以在人工智能、自然语言处理等领域，当作数据分析的辅助系统应用，该系统的设计和维护阶段的任务分解明确，操作简单易学，使用群体只限于研究人员和测试人员，用户不仅可查看数据分析结果以外，还可以组内交流意见。

3 需求分析

参与者分管理员、测试员和计算相似度后台系统。管理员可以进行维护测试员信息、维护数据库、审核新提交的数据、维护公告等操作。测试员首先登录系统后可以使用计算相似度，上传新数据，修改个人信息、查看公告和在分组讨论模块留言等功能。至于未注册系统的用户系统每一天只提供三次免费查询功能。计算相似度后台系统以辅助参与者的身份与外部环境进行交互。

4 句子相似度计算小程序的系统模块分析

系统由登录界面、测试相似度界面、查看公告界面和分组讨论界面组成。其中前台由测试员和管理员都可以登录，每个模块两类用户均可实现的操作有：登录界面中可以注册、登录、填写个人信息；测试相似度界面中可以搜索相似单词、搜索相似句子、上传新数据；查看公告（新闻）界面中可以完成查看新发布的公告、搜索公告、点赞公告（新闻）等操作；分组讨论模块中可以留言、点赞。

后台智能允许管理员访问，其中管理员可以完成的操作有：（1）用户管理模块：审核注册、注销的测试员账号，维护测试员基本信息，并更新测试员实体类数据库表；（2）相似单词（句子）管理模块：添加新的计算规则、维护有误数据，并更新单词（句子）实体类数据库表；（3）公告（新闻）管理模块：上传新公告(新闻),维护已经上传的新数据,并更新公告(新闻）实体类数据库表；（4）分组讨论模块：维护留言信息，并更新留言实体类数据库表；（5）系统历史记录管理模块：维护系统使用记录信息，并更新历史信息实体类数据库表。

5 句子相似度计算小程序的功能分析

系统中除了注册和登录的前提条件为参与者打开系统界面外，维护、上传、留言、搜索等功能的前置条件是参与者登录系统成功，其基本的操作流程如下：

1. 注册功能：测试员登录前先需要注册，新测试员通过填写姓名、联系方式、单位和验证码进行注册系统，并提交信息。管理员对其参与者信息进行审核，如果审核通过，将分配 ID 账号编码，并发送登录密码，设置其用户权限。测试员收到管理员信息后对账号密

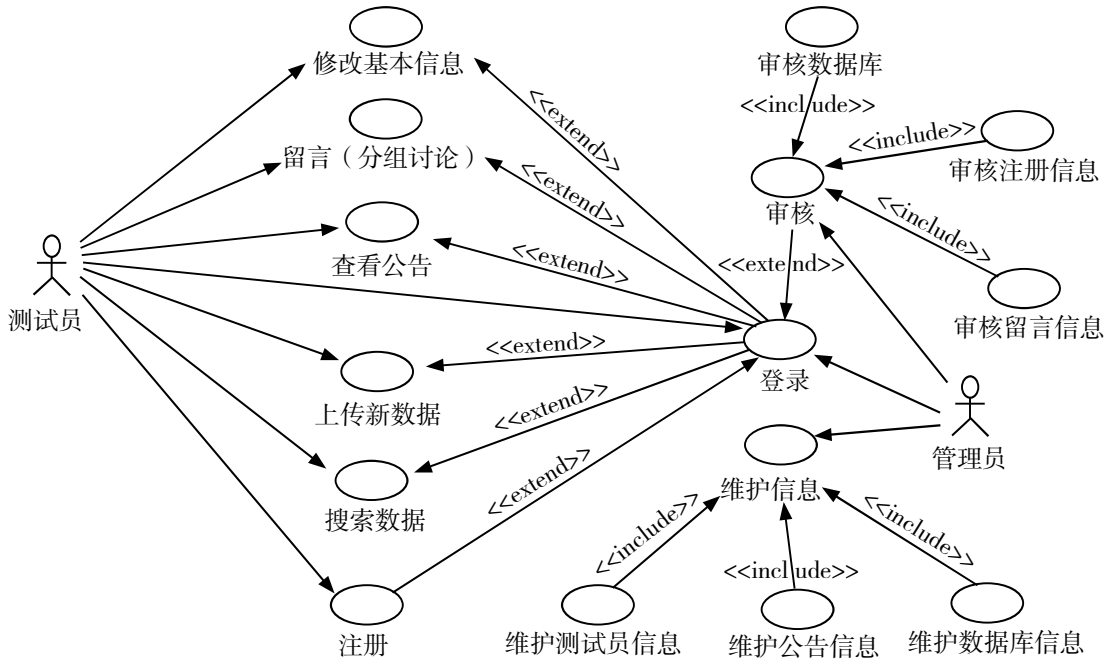

图 1 系统用例分析图

码进行修改。

2. 登录功能：管理员和测试员首先登录成功后可以正常使用小系统。为了保护测试员的基本信息，忘记账号密码时只允许输入三次，如果输入有误，将系统提示错误信息。测试员忘记密码时，可以通过手机验证等措施找回账号密码。

3. 维护测试员信息：新测试员注册系统时，管理员查看其参与者基本信息的完整性，并审核其身份；如果测试员注销其账号，管理员将数据库中删除其基本信息，并取消权限；部分信息安全级别较高的，隐藏的用户个人信息由管理员亲自审核、修改、补充和删除。

4. 维护数据库：句子相似度系统的数据库规模可无限地扩展，随着新数据的上传，器容量可以不断增大。但参与者提交新单词或句子之后，管理员先审核该数据，如果数据中有拼写错误或者语法错误等问题，将数据审核结果发送给上传数据的测试员；该测试员第二次重新编辑数据，确保无误后，再次提交到系统上。管理员再次审核该数据，如果新上传的数据中没有任何问题，下一步检索在数据库中是否存在该数据，如果没有则添加至数据库中。

除此之外，管理员还可以删除数据库中重复的信息或者错误的信息，还可以进行修改、更新等维护操作。用例维护信息的前置条件是数据库在系统中存在，管理员登录管理模块成功。管理员与系统参加该用例，基本流程为：首先，管理员在数据库中检索相关单词或句子；其次，如果数据库中存在，则维护其内容，并提交；最后，数据库覆盖原数据后，则提示维护成功。替代流为：如果数据库中找不到关键字，则系统提示管理员该数据数据库中不存在；如果元数据覆盖失败，则系统提示管理员数据维护失败，请重新设置。

5. 上传新数据：测试员登录系统后可查看某一个关键字在数据库中的相似句子或者单词，如果搜索数据时，数据库提示其数据不存在，则测试员可以将数据作为新关键词上传至数据库中，同时还可以提交使用该单词的句子和它的近义词，从而可以不断更新和充实系统数据库。

6. 搜索相似句：测试员首先在搜索框中输入关键字，并点击确认后，系统将其关键字的相似单词或句子从数据库中进行检索，并把最终结果输出。其中相似度的计算方法在查找相似单词和相似句子中应用。

搜索相似单词：测试员登陆成功后再计算相似度模块选择单词相似度计算功能，输入关键词，点击确。则系统会自动检测数据库中出现该单词的句子和近义词，通过对比单词的长短、对比字符串的序号和字符类型等方式计算其相似度比例，并把相似单词、近义词和相似度显示在界面上；如果数据库中找不到该关键字数据库则提醒测试员数据库中不存在该数据， 并提示是否确认添加到数据中。

搜索相似句子：该功能的基本用例模板与搜索相

似单词的基本流程是一样的，只不过计算数据模型是按句子所包含的有序组合中的字符序号、其来长度和词序进行匹配，搜索过程中找不到的句子时，测试员可以通过上传新数据的方式提交至数据库中，管理员审核通过后更新系统数据库。

7. 查看公告（新闻）：前置条件为管理员发表公告，只能高级管理员对公告进行上传、删除、编辑和置顶等维护操作。管理员发布公告成功后，测试员在登录系统的状态下，可以查看其公告，并公告左下方可以进行点赞和举报操作。

8. 分组讨论（留言）：只要注册系统的参与者都可以参与到分组讨论中。测试员同样先登录后发表自己的意见；管理员审核通过后以“组内留言”方式组内可以开展讨论。留言审核通过后，如果发现有错别字或者别的问题，留言的参与者可以自行修改、删除其内容。其他参加互动的人员通过引用该评论发表自己的意见，点赞自己赞同的留言，有意见的内容可以举报给系统管理员。

以上功能的后置条件为操作成功，更新数据库中对应的实体类数据库表，并生成每个阶段的操作记录文件。

6 数据库分析

系统数据库是通过连接多种子表的方式创建。其中，数据库子表可分为实体类和边界类数据库表。实体类表示参与系统交互的人员和系统关键信息存储的表格，如：用户实体类表、单词实体类表、句子实体类表、公告（新闻）实体类表、留言实体类表等。边界类表示系统界面数据存储的表格，如：主界面边界类表、公告页面边界类表、搜索页面边界类表、登录页面边界类表、留言页面边界类表等。

测试员和管理员完成一项操作后数据库表中以下几个实体类表格数据会被修改：一是账号实体类的属性包含用户名、ID、注册日期、联系方式、单位、账号密码和账号级别（管理员或测试员）组成。二是单词实体类的属性包含单词编号、单词内容、近义词和相似度信息组成。三是句子实体类的属性包含句子编号、句子内容、句子相似度信息组成。四是留言实体类的属性包含留言者名称、留言编号、留言时间和其内容组成。五是公告（新闻）实体类的属性包含编号、发布时间、标题、内容、点赞次数等内容组成。六是历史记录实体类的属性包含记录编号、详细内容、记录时间和操作者名称组成。

计算句子相似度的过程中，如果对单词进行切分时，需要在数据库中添加粘着性语言类的词缀实体类表，通过匹配单词和词缀表，完成词根和词缀的分解操作 [2]。

系统主界面包含搜索界面（计算相似度）、留言界面、登录界面和公告界面，这几个界面的基本信息通过边界类的类型保存至数据库，测试员与系统之间通过以上边界类来进行交互。

7 状态机分析

1. 数据状态：数据包含数据库中的实体单词、句子和测试员输入的关键词；数据状态根据词性判断，分别有词根状态、词缀未切分状态和连词状态等。

2. 操作状态：测试员输入关键词在系统进行搜索时处于正在搜索状态、匹配方式查找相似句子或单词需要时间，时间长度由数据库规模决定。

3. 参与者状态：测试员和管理员的状态第一阶段可以分为未注册、注册状态和注销账号等，完成注册小系统后由登录状态、未登录状态和退出系统状态等组成。

8 总结

句子相似度计算方法是目前人工智能领域普遍使用的技术，机器制造、自然语言处理等领域普遍使用的搜索，匹配功能中必须研究的项目之一。通过不同的数据模型计算相似度的方法识别模式具有庞大的数据库系统可以提高匹配工作效率，其操作流程和数据库之间的关系可以采用 UML 模型表示 [3]。本文中推出的小程序严格遵守系统需求分析和详细设计阶段的基本原理和任务要求，完成了对粘着性语言类计算句子相似度小系统的设计和建模工作，该模型在各种级别的管理系统的搜索模块的开发工作中均可嵌套使用。设计方案中提出的关键字、属性等数据字典部分包含的字段没有重复出现，测试员的需求在允许范围内可扩展，系统的数据库和功能在维护过程中也需要不断地改造和升级。

参考文献：

[1] 江阿古丽·哈依达尔，卡哈尔江·阿比的热西提，阿里木江·亚森，等 . 一种哈萨克语句子相似度计算方法的研究 [J]. 新疆大学学报（自然科学版），2012,29(04):471-474,479.

[2] 江阿古丽·哈依达尔，吐尔根·依布拉音，艾山·吾买尔，等 . 哈萨克语名词构形词缀有限状态自动机的构造 [C]// 第三届全国少数民族青年自然语言信息处理、第二届全国多语言知识库建设联合学术研讨会，2010.

[3] 魏金津，任女尔，蔡建军 . 基于相似度计算的 UML 图匹配算法设计模式检测技术研究 [J]. 电脑知识与技术，2018,14(28):165-167,171.