学校代码：10246

学　　号：

# 硕 士 学 位 论 文

（专 业 学 位）

# 基于关联分析的智能在线培训系统设计

# Design for Intelligent Online Training System Base on Associati

1、标题包括核心技术和应用对象，注意第四章重点内容是核心技术的应用！
2、黑体且加粗，不超过 20 汉字

院　　系：　软件学院

专　　业：　软件工程

姓　　名：

指 导 教 师：

完 成 日 期：　2014 年 2 月 19 日

# 目 录

摘 要..... III
ABSTRACT.....
第一章 引 言.....
1.1 智能在线培训系统发展现状.....
1.2 海策智能在线培训系统存在的问题.....
1.3 本文的主要内容.....
1.4 本文的篇章结构.....
第二章 关联分析技术基础.....
2.1 数据挖掘的过程.....
2.2 关联规则.....
2.3 Apriori 算法简介 .....
第三章 海策智能在线培训系统需求分析.....
3.1 智能在线培训系统主要功能.....
3.1.1 教程管理模块 ..... 13
3.1.2 智能答疑模块 ..... 15
3.1.3 互动社区模块 ..... 17
3.2 智能在线培训系统核心流程..... 18
3.2.1 教程学习流程 ..... 19
3.2.2 在线答疑流程 ..... 20
3.2.3 社交关系管理流程 ..... 21
3.3 智能在线培训系统的安全需求..... 23
第四章 海策智能在线培训系统设计..... 24
4.1 智能在线培训系统总体架构设计..... 24
4.2 用户行为数据采集子系统设计..... 26
4.2.1 用户学习轨迹数据采集 ..... 27
4.2.2 问答匹配结果数据采集 ..... 29
4.2.3 社交行为数据采集 ..... 31
4.3 用户偏好分析子系统设计..... 33
4.3.1 用户行为数据预处理 ..... 33
4.3.2 用户偏好关联规则分析 ..... 37
4.3.3 用户偏好规则的存储和发布 ..... 41
4.4 智能助教子系统设计..... 42

1、 目录第一章、第二章只需要 2 级标题，其他各章节可以到 3 级标题（最后一章结论只需 2 个 2 级标题）。
2、 目录的字体用小 4 号，章名和下级标题留半倍行距,2 级和 3 级标题之间不要留间距！章名黑体加粗，其他宋体即可！
3、 目录页第一行“目录....I”删除
4、 目录格式要紧凑

4.4.1 相关教程主动推送 ..... 42
4.4.2 在线提问智能答复 ..... 44
4.4.3 社交关系智能推荐 ..... 46
4.5 与同类系统比较..... 47
4.6 海策智能在线培训系统应用效果..... 48
第五章 结 论..... 49
5.1 海策智能在线培训系统的特点..... 49
5.2 不足与展望..... 50
参考文献..... 51
致 谢..... 53

# 摘 要

随着互联网技术的迅速发展，在线培训的市
海策智能在线培训系统拥有在线学习、在线答疑
拥有一套专业、高效、科学的培训管理模式和学
增多，教程的查找越来越困难，另一方面由于学
师处理回复的工作量太大，导致问题无法及时回
系统的互动社区虽然提供了用户之间的交流和协
立，用户的社交网络没有形成规模，群体智能的
通过提高系统的智能度来改善用户的这些体验，
程、提高问题回复的及时性、提升用户社交网络
决的问题。

首先讨论了海策智能在线培训系统的现状以
联分析来提高系统智能度的必要性。在此基础上
模块、互动社区模块等主要功能，分析了教程学
系流程等核心流程，然后给出系统整体架构设计
用户偏好分析子系统、智能助教子系统进行详细
联分析实现相关教程主动推送、在线提问指定能
对系统的可维护性和可扩展性进行了分析总结，以及对系统的未来发展进行展望。

**关键词：**数据挖掘，关联规则，智能答疑，智能培训系统，在线培训

1、 摘要分 2 段，大约 300-500 字：第 1 段是本文的研究意义，简要介绍，注意不要泛泛谈，一定针对所做工作的价值；第 2 段概要介绍本文所用的研究方法、结果和结论等。注意语言流畅，一口气能读完。可以采用“首先分析…的现状，明确了…的问题（或不足）。在此基础上，应用…技术（或理论），先做…,然后…，最后…。其中重点做…。

**2、 开题报告如果写好了，可以把研究思路中的时态换一下，研究意义简化即可，不要重新写。**

3、 不要出现“本文“、”我们“等字样。尽量不要用”提出“。

4、 注意是“关键词“而不是”关键字“。关键词 5 个，可以选择本文所用的核心技术、理论、 对象等作为关键词。关键词之间用分号隔开。

5、 中文关键词不要用英文！

1、 注意全文的语言简明扼要，通顺易懂！

2、 在下面给出的注意事项中，仅仅在 1 处指出，全文类似处不再指出！

3、 一般情况下全文的篇幅请控制在 55-65 页，不要太少或太多！

# ABSTRACT

With the development of Internet technology, online training market is showing a trend of rapid growth. Haice Online Learning System is software that has online learning, online Q & A, online assessment, integration and management functions. It is a set of professional, efficient and scientific management mode and learning mode. With the increase in the number of tutorials, it becomes mor to search the tutorial. In the Q & A section, the numb every day that cost so much effert and time, which a not be answered promptly, and reduces the students' Furthermore, interactive community module provide collaboration features between users, but each independent, user's social circle is small-scale an dose not take effect. Therefore, there is an urgent intelligent features to improve the ser experience, response speed.

First of all, the status and issues of the online learing system are analyzed, and the necessity of developing the intelligent features with association analysis to improve the system is illustrated. On this basis, the main features of tutorials management module, intelligent answering module and interactive community module are discussed, and core process of tutorial learning, online Q & A and community management are analyzed, and the overall structure is designed. Then the key subsystems, such as user behavior data collection subsystem, user preference analysis subsystem and intelligent assistant subsystem are designed in detail. Meanwhile, the implements of tutorials intelligent recommendation feature, intelligent answering feature, and social circle recommendation feature with association analysis are discussed especially. Finally, the maintainability and scalability of the system are summarized, and the prospect of the system is discussed.

1、 Abstract 与中文摘要对应，只要意思符合即可，不要逐字对照翻译！
2、 注意用**被动态、第三人称**，注意语法不要有问题。
3、 关键词中英文对照，顺序要对应
4、 注意不要有语法错

**Key words:** Data Mining, Association Rule, Intelligent Q & A, Intelligent Training System, Online learning

# 第一章 引 言

页眉注意与标题一致！

互联网诞生以来，它的迅速发展每天都在改变着人们的[illegible]、商业、通讯、新闻、医疗等行业都有革命性的变化，同时对传统的教学方式也产生了一系列的影响，从过去的纸质教材到现在的电子图书、视频教程、互动教学，这些年的发展都为教学提供了更多的方式。通过互联网进行在线培训可以跨越因地域等方面造成的教育资源分配不平等，使[illegible]限制，提升学习效率。

从当前情况看，国内的在线培训市[illegible]在线培训系统如雨后春笋地涌现了出来，也有[illegible]该行业正在蓬勃的发展。海策公司在创立之初[illegible]发了自己的第一代在线培训系统，但由于系统[illegible]大量的教程中，很难找到自己感兴趣的内容，[illegible]常常因为数量巨大，导致无法得到及时的回复[illegible]的互动还处于简单的社区论坛的形式。因此[illegible]决以上的问题，使得产品在激烈竞争中更具竞[illegible]任务。

1、 引言分为四部分：与本文研究相关的现状、不足、本文的主要工作和章节安排。注意这里的现状和不足一定紧密结合本文的讨论，不要泛化，讨论大的概念！思路是由现状找出不足，由不足引出本文的研究意义和主要工作。可以在几部分之间加过度的句子，以便论文的逻辑清晰！

2、 现状既要讨论业务，又要讨论系统；既要讨论国内外同行，又要讨论本文涉及的企业！

3、 篇幅限制在 5 页左右！

4、 现状部分参考文献部少于 6-8 篇！

## 1.1 智能在线培训系统发展现状

在线培训是互联网培训的一种新兴技术和模式，随着互联网的普及和信息化程度的提高，以及学历教育在社会上的认可度的提高，在线培训的市场正在呈现出高速增长的事态[1]。

（1）我国智能在线培训系统的发展情况

随着互联网技术、数据库技术、多媒体技术的发展，我国智能在线培训系统的发展也经过了不同的阶段，大致分为三个阶段：点播教学模式，互动教学模式，社区化教学模式。这些模式在国内、国外都有很多代表性的公司和网站，国内在线培训有代表性的有多贝、YY 教育、万鹏学堂、网易公开课、传课、淘宝同学等[2]。

点播模式是最早出现的一种培训模式。一些高等院校、公司企业和培训机构为了更好的存储和管理培训资源，利用互联网的优势，开发网上培训系统来存储和管理培训资源，用户可以随时随地点播音频课件、视频课件、查阅电子教案等培训内容，节省了培训成本，增加了用户的学习方式[3]。但该阶段的主要特点是以课件为中心在网上进行培训资源的展示。这种模式的缺点是系统的智能性较

注意页脚用数字，编号从引言开始！

差，每个用户看到的内容都一样，功能上仅限于对培训内容的管理，而且培训内容也没有相关的标准，资源格式不统一，没有互动性，培训效果并不十分理想[4]。

互动教学模式的出现使得培训系统进入了智能的初级阶段。互动教学模式在点播式教学的基础上增加了很多互动教学方式，一般用户在线注册、登录后，系统会对用户的学习过程进行记录、统计，集成聊天工具、论坛系统、多人语音视频系统、在线测试等服务，提高了用户和授课老师之间的互动水平和用户的学习效果，其主要特点是以[illegible]交互功能，强调为用户提供及时有效的服务。[illegible]很多用户只有一位培训老师，多人语音和视频[illegible]线的互动对每次参与的人数有限制，非实时的[illegible]问题不一定有人答复，错过了直播时间就只能[illegible]还不高。

全文行距为固定值 20 磅，包括中英文摘要！

社区化教学模式的阶段又进一步提高了系统的智能性。社区化教学模式的特点是以用户的关系为中心、以用户和内容的互动为驱动，社区化的教学本质上也是人与人的互动、人与教学内容的互动。这种模式在互动教学模式的基础上，利用一些社交性功能和服务，建立起属于自己的学习网络，包括学习资源网络和社交伙伴网络。在社区化教学模式中，每位用户既可以是学生也可以是老师，每位用户都可以作为教学资源的提供者，也可以作为学习者，每位用户既是用户关系网的中心、也是用户关系网的节点，而用户关系体系的建立会反过来促进每个关系圈内的内容产生及用户互动，这样就形成用户到内容，内容到用户的良性循环，这不仅会增强每个用户个性化的教学体验、也增强了培训系统对用户的粘性，并使得系统处于智能的增进和优化过程中。这种模式的缺点是，教学内容因为都是由个人用户产生的，教学内容的质量很难控制，标准也很难统一[5]。

（2）国外智能在线培训系统的发展情况

国外目前点播教学模式、互动教学模式、社区化教学模式都有，但国外的发展水平要高于国内，已经有很多公司获得了风险投资，用户数量也在高速增长。国外一些优秀的培训系统在内容上没有照搬线下的课程，而是针对互联网用户进行了优化，而且大部分都有一定的智能性。

Coursera 是培训系统中互动教学阶段的代表，是由斯坦福大学教授创建的免费在线大学课程项目，包含了从商业到技术、再到社会学或文学的各种课程。目前已经和 83 多 个教育机构展开合作，提供 400 多 门免费课程，该公司 2103 年获得 4300 万美元 B 轮投资，由 GSV Capital、 International Finance Corporation、俄罗斯首富 Yuri Milner 等联合投资，此前该公司曾获得 2200 万美元的投资。还有 Udacity 是由斯坦福大学教授、Google X 实验室研究人员 Sebastion Thrun 创办，旨在为尽可能多的学生带来高质量的大学课程[6]。在形

式上除采用视频授课外，其它基本都与在真实的大学中接受教育一样，也有课前大纲、课后作业、课后测试等等。有报道称该公司已获得 500 万美元投资。出此之外还有麻省理工公开课、耶鲁大学公开课、哈佛医学院，互动教学阶段的培训系统多属于这些免费大学课程项目[7]。

以 Udemy 为代表的开放式在线培训系统，在群体智能方面做得最好。系统的所有用户都可以在该系统开设自己的课程，自 2010 年推出以来，该公司已经吸引了成千上万的人授课，课程可以是免费也可以收取一定费用。该公司曾表示，一些讲授者已在该系统上赚到数十万美元，截止 2013 年 Udemy 也已累计获得 1600 万美元投资。此外还有如可汗学院、Lynda、Course Hero、InstaEDU 等众多提供在线课程的创业公司。

走在智能培训系统前沿的是 Codecademy，它是一个免费有趣的在线互动编程学习网站，它像玩游戏一样，让你一步一步从易到难来完成学习，你可以每天利用碎片时间来从零基础入门到掌握一门编程语言，它展示出了在线培训的未来应该是基于标准算法、系统模型、数据挖掘、知识库等为用户提供个性化、定制化的学习服务，在这个过程中，老师授课的依赖会越来越小，并被技术部分取代。

总之，国外目前大部分都已经进入了培训系统的互动教学阶段，更有一部分在群体智[illegible]行业的前面[8]。

参考文献在论文中的标注用上标，句号前！

（3）[illegible]目前的情况

海策智能在线培训系统是从 2008 年开始发布的，是从互动点播阶段开始，在系统运行过程中，逐渐添加智能功能和社区功能。在几年的市场销售过程中，该系统为公司创造了利润，为公司进入智能在线培训行业积累了宝贵的经验。

海策智能在线培训系统从一开始就支持视频的播放，最初的智能功能体现在学习进度记忆功能，无论用户浏览到什么教程，系统都会记住用户上次在该教程停留的位置，可以从上次停留的位置继续播放视频；后来还增强了学习考核的一些智能模块，例如随机试卷、在线测试、在线评分、知识点掌握情况报告等；之后还增加了社区功能，在不断对社区功能优化的过程中，逐渐体现出系统的群体智能，通过社区功能很多教程内容错误、题库错误等得到了用户的反馈，甚至由用户自己解决，例如很多教程的错误一开始由用户反馈，逐步过渡到用户修改后提交管理员审核，该功能使得教程的正确性得到进一步的提升。

但由于原来对于系统智能性认识的欠缺，在系统设计中智能性的体现还不够充分，导致随着系统的发展，大量的教程搜索、大量的用户提问无法及时答复、用户社交网络没有发挥群体智能的优势等，也促使公司采用数据挖掘等方法对产品的智能性进行设计和改进。

## 1.2 海策智能在线培训系统存在的问题

虽然公司较早的进入了该行业，开发的智能在线培训系统也得到了很多用户的好评，为公司创造了一定的利润和积累了在该行业的经验，但是在互联网的强劲发展势头下，更多形态的在线培训模式出现并影响着市场，用户和市场也对该类产品有了新的要求，原来的系统显露出很多问题，如果想在该行业里有更好的持续性发展，公司必须分析当前的市场情况，找出和改进这些问题。

目前海策智能在线培训系统主要面临以下几个方面的问题：

（1）用户很难在日益增长的教程中找到感兴趣的教程

随着系统的运行，培训教程数量越来越多，用户要在众多的教程中找到自己感兴趣的概率越来越小。最初系统提供了通过目录进行检索的方式，但是由于目录分类也越来越多，每一个目录下面的教程数量也相当庞大，而且系统定义的目录分类和用户理解的可能不一致，导致用户有时候不知道某些教程会确切的在什么目录下。虽然之后提供了关键字的全文搜索，但是有时候用户不是很确定关键字或者确定的关键字查到了太多的内容，这样的方式也没有很好地解决用户的问题，而且全文检索往往是由用户主动发起，用户更愿意使用互联网搜索引擎，这就使得该功能使用率不高，常常出现的现象是用户通过搜索引擎搜索到一篇教程，学习之后就离开了，系统对用户的粘性较低。

因此，如何提高相关教程查找的便捷性，使得系统用户粘度提高一个层次是当前系统的一个迫切需要解决的问题

（2）用户的提问不能及时

海策智能在线培训系统是采 题、模拟考试的方式来完成一个教程的学习。在学习过程中 统中进行提问，由培训老师在管理后台中进行回答。由于教 益增多，每天产生的学生问题也越来越多，培训老师每天都 回复，由于问题数量太多导致很多问题排队等待答复，常常 答案，而且学生对每个教程的提问一般都集中在几个难点， 是雷同的，培训老师大部分的答复都是在做重复性的工作， 保持学习的连贯性，在一定程度上降低了用户体验和学习的 师的重复工作量很大，也增加了教程的运营成本。

1、标题用黑体、加粗！用三级标题：
第一章 …
1.1 …
（1）
①
全文缩进统一
2、标题与正文前后之间留半倍行距！全文类似

因此，如何改善答疑的方式，缩短答复的响应时间，提高学习的连贯性和用户体验，减少培训老师的重复工作也是当前系统需要进行改进的方面。

（3）官方单向增加和更新教程内容的方式限制了系统的发展

原来系统的教学内容是来自公司教程研发团队自主编写或由某技术公司授权的，每个教程的内容都是由管理员通过后台录入、编排，包括设置课程、章节、

题库等，虽然用户也可以发现问题并且修改后提交审核，但系统的内容是否丰富还是局限于海策公司自己所拥有的资源，这样的由官方单向对教学内容进行增加管理的模式制约了教学内容的扩展，使得系统运行一段时间后，内容没有更新导致用户有流失的现象。

如何使系统的教程内容能更好更快的扩展，充分发挥用户的群体智能，获得更多渠道来丰富系统的内容进而提高行业竞争力成为了系统需要解决的问题。

## 1.3 本文的主要内容

本文首先对当前智能在线培训系统做了简要介绍，阐述了国内外智能培训系统的发展历程和现状，并对海策当前的发展状况也进行了分析。本文还对智能在线培训系统发展的各个阶段做了介绍，分析了每个阶段的智能功能特点，并给出了具体的案例，然后总结了目前海策智能培训系统存在的问题，说明了系统目前在如何让用户快速进行教程查找、问题答疑如何得到快速准确的答复、如何发挥群体智能来使培训内容得到良性的增长这些方面存在的问题，进而得出了要提高系统的智能性，需要采用数据挖掘关联算法应用到系统中的必要性。

在介绍了数据挖掘的过程、关联规则、Apriori 算法等关键技术的基础上，分析了海策智能在线培训系统的功能结构，讨论了系统的教程管理模块、智能答疑模块、互动社区模块三大主要功能，分析了海策智能在线培训系统的教程学习流程、在线答疑流程、社交关系管理流程三个核心流程，详细描述了系统在教程学习过程中、在线答疑过程中，用户在社交网络的协作中的系统需求，还对这些智能模块的重要性和应用效果以及系统的安全需求做了说明。

在此基础上，给出了海策智能在线培训系统的整体架构，对各个子系统的功能和工作原理做了介绍，并且对用户行为数据采集子系统、用户偏好分析子系统、智能助教子系统进行详细设计。首先，在数据采集方面，对用户学习轨迹、用户关注方向、问答匹配结果的数据采集设计做了介绍；再者，对用户偏好分析的数据预处理、关联规则分析、关联规则结果存储进行了设计分析；最后，还介绍了获得关联规则后应用到系统中的具体功能以及应用场景介绍。重点讨论如何基于关联分析的 Apriori 算法，运用开源的数据挖掘软件 WEKA 实现系统的这些智能功能。

本文还对海策智能在线培训系统在同类系统中的智能性的设计和实现进行了比较分析，并简单说明了系统最终的应用效果。最后，对本系统进行了总结，对今后的应用前景进行了展望。

## 1.4 本文的篇章结构

本文共分五章，首先简要介绍了论文的背景情况，引出了本文所作的主要工作内容。然后简单介绍了海策智能在线培训系统的现状和不足。在此基础上，详细分析了目前智能在线培训系统的需求和系统业务流程，并基于关联分析方法对智能在线培训系统的数据挖掘部分进行了设计和实现，最后分析了系统存在的不足及问题，并且在此基础上给出了改进

1、这里有一段概要的句子，在分别介绍各章节的内容安排
2、少用或不用“提出”字样
3、全文统一使用“本文”，不要用本论文、论文等字样
4、删除全文“我们”字样

第一章从分析海策智能在线培训系统 出了论文的主要研究内容，既解决上述问题 数据挖掘的方法改进学员的学习效率、提高 以此增强产品竞争力，突出了本文的现实意

第二章围绕数据挖掘原理和技术，介绍了关联分析的核心技术和方法，这是系统进行数据挖掘的技术基础，明确了本文研究的重点方向。

第三章在详细了解海策智能在线培训系统核心技术的基础上，对该系统进行了需求分析，讨论了智能在线培训系统的功能结构，对系统的核心业务流程进行了重点分析。

第四章阐述了该系统的的系统结构和软件架构，主要针对和数据挖掘相关的子系统给出了设计方案和实现方法。

第五章对基于关联分析的智能在线培训系统的特点做一个总结，对系统的不足之处提出了改进方案，最后指出了系统进一步发展的要点。

# 第二章 关联分析技术基础

1、 本章标题应改为具体的核心技术名称
2、 第二章是全文的理论基础，先分析使用某核心技术可以解决引言提到的问题，然后简要介绍后文分析用到的该核心技术具体内容部分即可。注意简要介绍 1 项核心技术即可！不要过多介绍概念！篇幅限制在 5 页左右！
3、 每一章开始从新的一页开始！

海策智能在线培训系统是以智 系统的市场竞争力，这就需要引入各 智能功能。系统目前已经已经存在的 大量的数据进行分析、挖掘和应用， 模块再次进行创新，会为产品增加更

数据挖掘就是从大量的、不完 含在其中的、但事先不知道的、且又 是数据挖掘的常用的方法之一，用于 相关关系，Apriori 算法是最有影响的挖掘布尔关联规则频繁项集的算法，系统将使用 WEKA 实现 Apriori 关联算法的计算。

## 2.1 数据挖掘的过程

数据挖掘是从数据中提取出隐含的、过去未知的、有价值的潜在信息，是一门从大量数据或者数据库中提取有用信息的科学，并且综合了统计分析、机器学习、人工智能、数据库等诸多方面的研究成果而成。整个数据挖掘过程主要包括：明确业务问题、数据集成、数据抽取、数据转换、数据清洗、数据分析、结果评估、数据应用。

首先需要明确业务问题，这不仅仅是指确定商务上的目标，还包括明确使用数据挖掘技术需要解决什么问题，只有定义了数据挖掘需要解决的问题，才能有的放矢的进行数据准备和预处理，最终给出解决方案，不至于迷失在大量的数据当中。

数据集成并非固定的步骤，一般当数据挖掘需要的数据来源于各种不同的数据库或者表，但他们有代表相同的数据含义，需要通过一定的关联关系把这些数据集成到一起，往往是代替使用多表关联查询来显示数据，以提高属于的利用率，这为提高数据分析的效率有着重要意义。

数据抽取通常在完成了数据的集成后对数据进行观察和分析，往往可以发现集成后的原始数据一般都是为了满足业务需要保留了大量的业务属性，但在建立数据挖掘分析模型时，并不是所有的属性都是关联分析需要的，只选取哪些有用的属性可以提高分析的效率和正确率，因此必须将原有数据与关联分析无关的属性进行删除。数据抽取将会减少数据分析的工作量、提高分析的效率，为生成一

个高效的分析模型提供了高质量的数据。

数据转换是数据预处理一个重要的环节，完成了数据的集成、数据抽取后的数据可能某些属性的意义是一样的，但是他们的字段类型不同，需要统一转换后的字段类型。还有一些属性虽然没有类型冲突，但在业务系统中记录的是用户的隐私信息，或者和商业有关的敏感信息，这些属性都把他们转换为指定的类型，不保存他们具体的含义，还有进行一些数据转换操作将大大缩小数据的扫描范围，提高效率，例如：如何改变数据之间的关系、属性的类型、是否需要汇总数据、过滤数据、去掉哪些明显没有相关性的数据等。

数据清洗是为了处理噪声数据、错误数据、缺失数据等问题，产生这些数据的原因可能是系统的漏洞、系统异常后导致的数据缺失，或者是人为的操作失误造成的，称之为“伪样本”。由于“伪样本”的存在会造成分析结果有偏差，所以在数据预处理过程中需要进行数据清洗。处理缺值主要的方法有删除元组、数据补齐和不处理三类。对于错误数据的问题，则如果可以推断最后可能的值则进行自动修正，也可以采取删除错误记录的方式。

数据分析需要选择一个合适的算法或是多种算法的组合，例如分类、聚类、预测、关联等，每种算法都有它使用的范围，常常同一个问题可以有多种算法，应该选择那种算法需要综合考虑各方面的因素。除了算法还有很多实现了该算法的工具，例如 WEKA 软件等，可以借助第三方成熟的计算软件实现数据分析的计算。

除此之外，数据分析得到的模型是否有效、可靠需要进行结果评估，可以使用模拟的数据、当前已有的真实数据进行演练，查看结果和预期的是否一致。数据分析得到的结果常常不能直接应用到系统中，还需要做进一步的数据转换、业务转换，甚至开发专门的业务系统对这些数据进行管理和调用。

## 2.2 关联规则

关联规则（Association Rules），是数据挖掘的常用的方法之一。关联规则是用于从大量数据中挖掘出有价值的数据项之间的相关关系，但不一定是因果关系[9]，关联规则多不考虑项目的次序，而仅考虑其组合。

关联规则最早被应用在分析零售企业客户的购物行为，又被称为购物篮分析（Market Basket Analysis），它从大量商务事务记录中发现有趣的关联关系，可以帮助许多商务决策的制定，如分类设计、交叉购物和促销分析[10]。经典的实例是沃尔玛公司对顾客的购买记录数据库进行关联规则挖掘，想知道顾客经常一起购买的商品有哪些，在沃尔玛数据仓库里各门店的详细原始交易数据的基础上，沃尔玛利用数据挖掘方法对这些数据进行分析和挖掘，发现跟尿布一起购买

最多的商品竟是啤酒，经过大量实际调查和分析，揭示了一个隐藏在"尿布与啤酒"背后的美国人的一种行为模式。在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有 30%～40%的人同时也为自己买一些啤酒。产生这一现象的原因是，美国的太太们常叮嘱她们的丈夫下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了他们喜欢的啤酒。

要了解关联规则需要知道以下几个基本的定义，结合当前海策智能在线培训系统对几个基本的定义做简单的介绍。

（1）事务

系统中用户可以对感兴趣的教程进行收藏，这些收藏教程的信息都存在数据库中，每个人所收藏的所有教程为数据库中的一条记录 ，称这样的一条记录为事务，所有事务的集合构成了事务数据库。在关联规则中，把该事务数据库记为 $D$，$D=\{t_1,t_2,...,t_n\}$，每个人所收藏的所有教程为 $D$ 中的一条记录 $t$。

（2）项集合、项目、项集

所有收藏的教程名称汇总在一起，去掉重复的项，可以得到一个集合 $I=\{I_1,I_2,...,I_m\}$，称之为项集合。每个用户收藏的教程是项集合的一个非空子集，称之为项目 $i_k,k=\{1,2,...m\}$，项集是由 $I$ 中项目构成的集合，若项集中包含的项目数为 $k$，则称此项集为 $k$ －项集。在项集中同时出现的次数超过人工定义的阈值的项集称为频繁项集。用来获取频繁项集的中间项集称之为候选项集，候选项集中满足支持度条件的项集保留，不满足条件的舍弃。

（3）支持度、置信度、强关联规则

关联规则是形如的 $X \Rightarrow Y$ 蕴涵关系，例如收藏教程 $X$ 的人有多大程度上会收藏教程 $Y$，$X$ 和 $Y$ 分别称为关联规则的先导和后继，关联规则 $X \Rightarrow Y$ 在 $D$ 中的支持度是 $D$ 中事务包含 $X \cup Y$ 的百分比，也就是说数据库中同时包含了 $X$ 和 $Y$ 的事务数占所有的事务总数的百分比，关联规则 $X \Rightarrow Y$ 在 $D$ 中的置信度是包含 $X$ 的事务中同时包含 $Y$ 的百分比[11]。

支持度和置信度是度量项目关联的重要指标，他们分别描述了一个关联规则的有用性和确定性。支持度是用户兴趣的重要度量，支持度很低的关联规则表示随机现象。置信度高的关联规则反映了 $X \Rightarrow Y$ 的正确程度，也就是发生 $X$ 的时候发生 $Y$ 的可能性有多大。如果同时满足最小支持度阈值和最小置信度阈值，则认为关联规则是有趣的，称之为强关联规则，否则称为弱关联规则[12]。如果不考虑关联规则的支持度和置信度，在事务数据库中可以发现无穷多的规则。事实上，满足一定的支持度和置信度的关联规则才是有意义的。一般这些阈值由用户或者专家设定，这些阈值太大或者太小都会影响关联分析的结果。

（4）关联规则的常见分类

根据关联规则所处理的值的类型，分为布尔关联规则和数量关联规则。布尔关联规则考虑关联规则中的数据项是否出现；数量关联规则中的数据项是数量型的，例如年龄(″20-25″)⇒购买(″牛奶″)，年龄是一个数量型的数据项。

根据关联规则所涉及的数据维数，分为单维关联规则和多维关联规则。单维关联规则只涉及一个维度，例如(″面包″)⇒购买(″牛奶″)只涉及“购买”；多维关联规则涉及两个或两个以上的维度，例如年龄(″20-25″)⇒购买(″牛奶″)涉及“年龄”和“购买”两个维度。

根据关联规则所涉及的抽象层次，分为单层关联规则和广义关联规则[13]。单层关联规则不涉及不同层次的数据项；广义关联规则是在不同抽象层次中挖掘出的关联规则。例如年龄(″20-25″)⇒购买(″蒙牛牌牛奶″)和年龄(″20-25″)⇒购买(″牛奶″)是广义关联规则，因为″蒙牛牌牛奶″和″牛奶″属于不同的抽象层次。

（5）关联规则的常见算法

关联规则有很多常见的算法，例如经典的 Apriori 算法是基于两阶段频集思想的递推算法，先找出所有的频繁项集，然后由平凡项集产生强关联规则，这些规则必须满足最小支持度和置信度。

还有 F-P 增长算法，又叫频繁模式增长算法，是 2000 年 Jiawei Han 等人提出的一种算法，该算法是一种不产生候选模式而采用频繁模式增长的方法挖掘频繁模式，该算法使用一种称为 FP 树的数据结构，由于 FP 树蕴涵了所有的频繁项集，其后的频繁项集的挖掘只需要在 FP 树上进行，该算法只进行 2 次数据库扫描。

再者还有 Eclat 算法，与 F-P 增长算法和 Apriori 算法不同，Eclat 算法加入了倒排的思想，具体就是将事务数据中的项作为 key，每个项对应的事务 ID 作为 value，这种数据处理方式很适合用关系型数据表示和实现[14]。

## 2.3 Apriori 算法简介

Apriori 算法是最有影响的挖掘布尔关联规则频繁项集的算法，核心是基于两阶段频集思想的递推算法，先找出所有的频繁项集，然后由频繁项集产生强关联规则，这些规则必须满足最小支持度和最小置信度，该算法在关联规则分类中国属于单维、单层、布尔关联规则[15]。

在关联规则，一般对于给定的项目集合，算法通常尝试在项目集合中找出至少有 C 个相同的子集。Apriori 算法采用自底向上的处理方法，即频繁子集每次只扩展一个对象，该步骤被称为候选集产生，并且候选集由数据进行检验，当不再产生符合条件的扩展对象时，算法终止。

在 Apriori 算法中，寻找最大项目集的基本思想是算法需要对数据集进行多步处理。首先简单统计所有含一个元素项目集出现的频率，并找出那些不小于最小支持度的项目集，然后从第二步开始循环处理直到再没有最大项目集生成。搜索所有的频繁项集需要多次扫描事务数据库，这也是影响关联算法性能的主要因素。

该算法的优点是简单、易理解、数据要求低[16]；缺点是在每一步产生侯选项目集时循环产生的组合过多，没有排除不应该参与组合的元素，每次计算项集的支持度时，都对数据库全部记录进行了一遍扫描比较，对于比较大型的数据库，会大大增加计算机系统的 I/O 开销[17]。

经典的关联规则数据挖掘算法 Apriori 算法广泛应用于各种领域，通过对数据的关联性进行分析和挖掘，挖掘出的这些信息在决策制定过程中具有重要的参考价值。Apriori 算法广泛应用于商业中。在消费市场价格分析中，它能够很快的求出各种产品之间的价格关系和它们之间的影响，为商家指定产品的价格提供决策参考[18]。还可以通过数据挖掘，分析用户的消费历史信息，猜测这些年来顾客的消费习惯，从而指定有针对性的促销策略和优惠活动。

本系统将使用 WEKA 进行 Apriori 的关联规则计算。WEKA 是一款数据挖掘开源软件，全名是“怀卡托智能分析环境”（Waikato Environment for Knowledge Analysis）[19]。该软件是用 Java 语言实现的，是一款免费的、非商业化的机器学习以及数据挖掘软件，包含了一个图形用户界面来与数据文件交互并生成可视结果，例如曲线、图表等。WEKA 集合了很多机器学习算法用来承担数据挖掘任务，包括对数据进行预处理、分类、回归、聚类、关联规则[20]，它还可以通过简单的方式快速实现自己的数据挖掘算法，来改进和扩展个性化的算法。重要的是它可以在 Java 项目中引入 WEKA 的类库，以完成诸如服务器端自动数据挖掘这样的任务 [21]，这正是本系统中采用的方式。

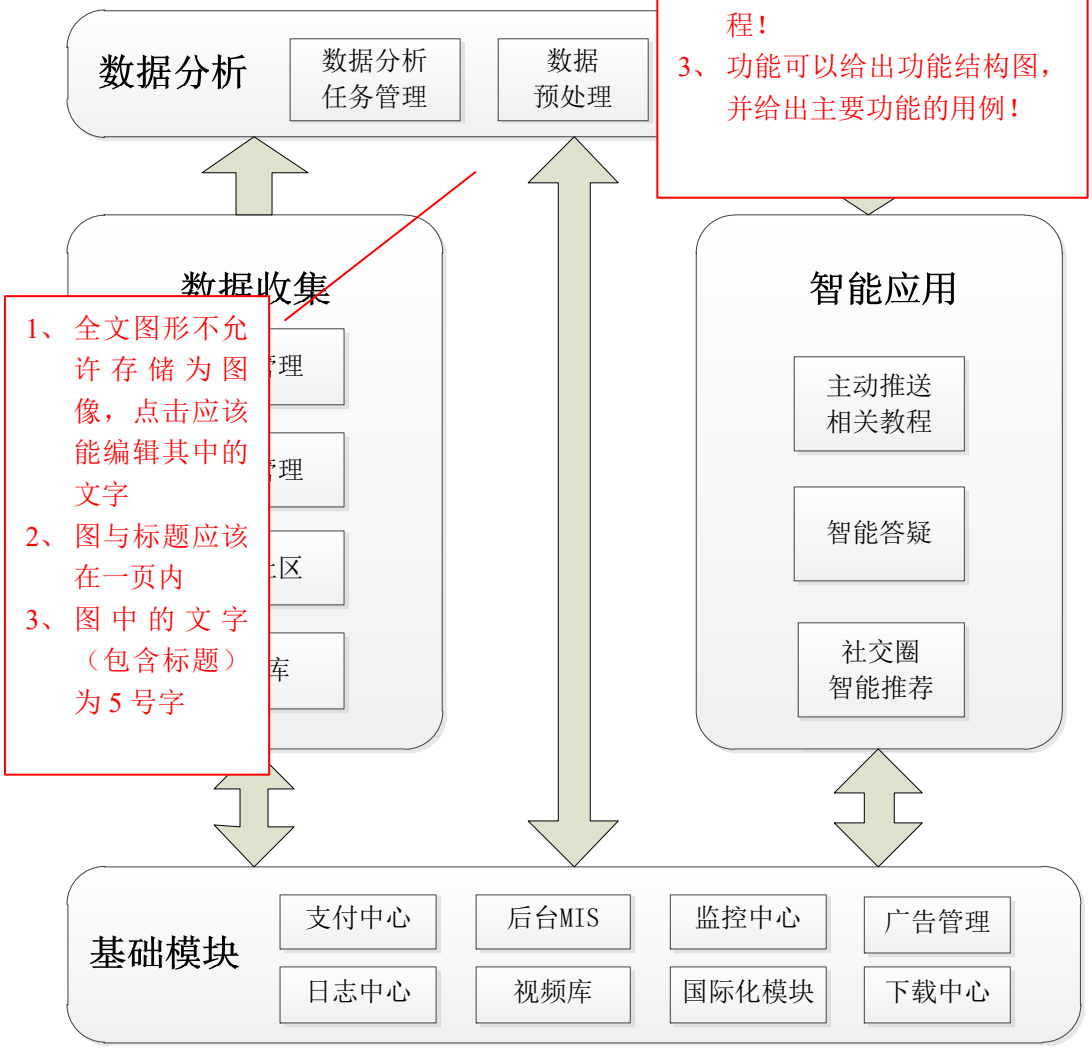
# 第三章 海策智能在线培训系统需求分析

海策智能在线培训系统拥有在线学习、在线答疑、在线测评、积分管理等功能，是集一整套专业、高效、科学的培训管理模式和学习模式的系统。本章对系统需求进行了概述，着重介绍该系统的教程管理、社区模块三大功能模块的需求，并对教程学习流程、流程进行了分析。

1、开发类的论文，第三章是系统的需求分析，主要讨论系统的**核心**功能和**核心**流程（尤其是与后文重点相关的内容，无关的功能和流程删除），**不要象项目报告，面面俱到**！流程尽量用 UML 用例等图画！
2、注意本章先分析系统的主要功能，再分析系统的核心流程！
3、功能可以给出功能结构图，并给出主要功能的用例！

## 3.1 智能在线培训系统主要功能

该系统用户分为学习型用户、分享型用户、作人员，他们的需求构成了该系统的主要功能，




1、全文图形不允许存储为图像，点击应该能编辑其中的文字
2、图与标题应该在一页内
3、图中的文字（包含标题）为 5 号字

图 3-1 海策智能在线培训系统功能结构图

### 3.1.1 教程管理模块

教程管理模块是智能在线培训系统的基础功能模块之一，它是系统产生内容的基础模块，为用户进行教程学习、问答互动等功能提供了前提。教程管理模块功能用例图如图 3-2。

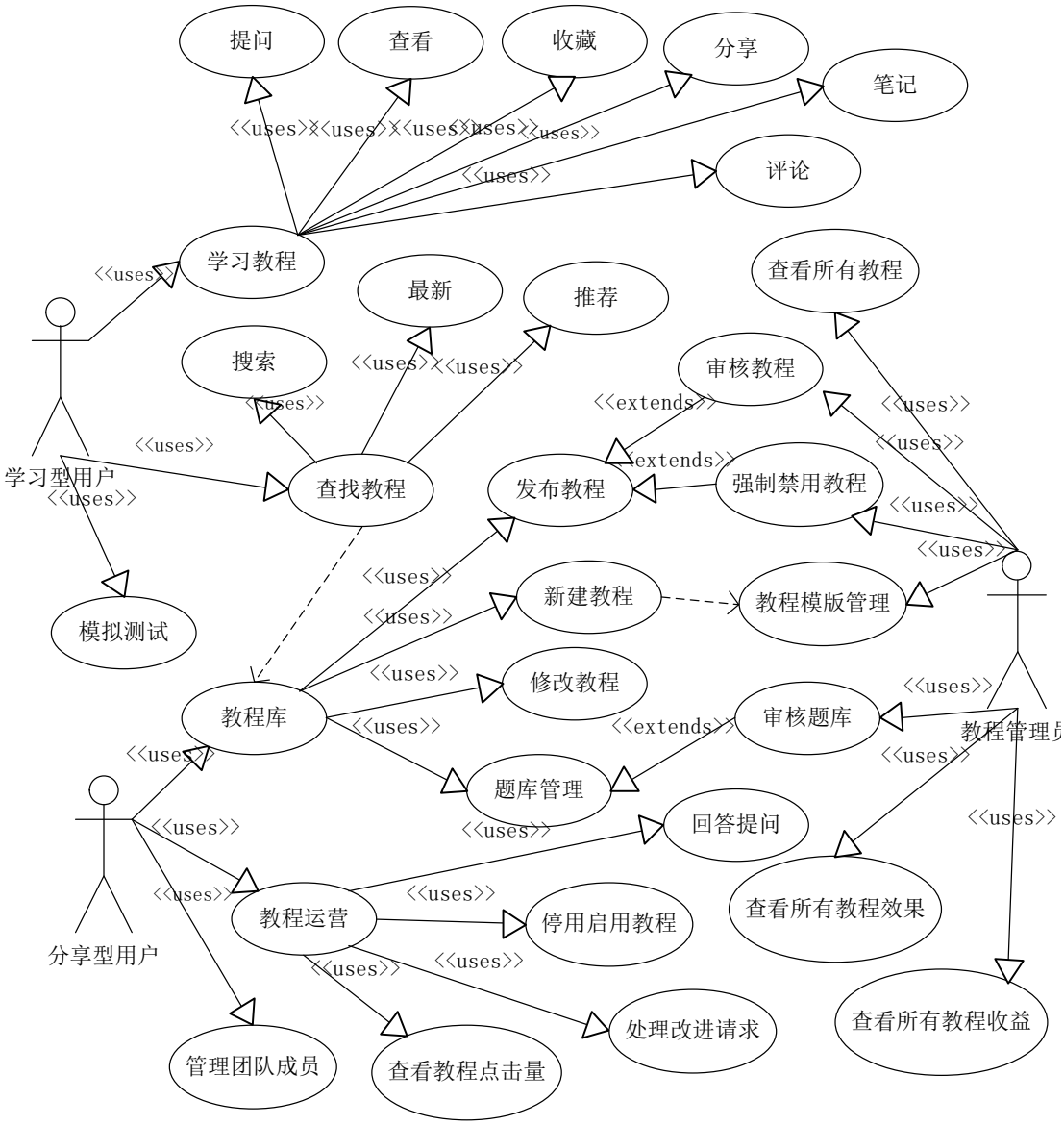



图 3-2 教程管理模块功能用例图

教程管理模块主要分为教程制作、教程审核、教程展示三个部分，涉及的用户包括分享型用户、教程管理人员和学习型用户。

（1）教程制作

教程制作主要是由分享型用户参与，分享型用户包括了海策内部的教程制作

团队、主动分享的互联网用户，任何通过注册的用户都可以成为系统的分享型用户。教程制作的功能需求包括教程排版、修改、发布、题库管理、教程运营、团队管理。

新建一个教程，用户可以选择一个系统提供的模板，模板可以帮助用户快速建立教程。教程分为标题、图文区、视频区、视频字幕、题库。其中标题为必填，图文区和视频区必须有其中之一，其他为选填。新建的教程并没有发布，学习型用户暂时也无法看到，需要确认所有内容都编辑完成，点击发布后，经管理员审核，才会出现到网站供其他用户查看。审核期间无法编辑，如果还需修改可以先取消发布。

教程的发布人可以为教程设计题库，并不是所有的教程都必须有题库，如果需要提供题库，则可以在某个教程下建立题库，输入问题、可选答案、正确答案。题目支持单选题、多选题、不定项选择题、是非题，不支持问答题。还可以为教程设置知识点，把题库的题目分别归类到各个知识点，支持一个题目属于多个知识点。

教程运营是为了检测教程是否运行正常、受欢迎程度如何。教程发布后，分享人可以通过系统查看用户的提问并回答，根据修改意见进行优化，查看该教程的点击量，查看该教程得到的广告分润收益。

教程支持多人协作修改和发布，成员管理也就成了必不可少的功能之一。教程默认的所有者就是用户自己，如果用户升级账号为组织，还可以邀请其他用户加入该教程的编辑，成员同意后将加入组织，成员也可以主动离开某个组织。教程所有人可以把不同的成员按组分类，不同的教程可以映射到不同的组或者成员。

（2）教程审核

教程管理员的功能需求包括一系列的审核、管理功能。凡是非官方团队发布的教程都需要审核，管理员确认内容与标题符合、没有违反法律法规的内容，则可以审核通过，如果有问题，则填写驳回理由，并审核不通过。题库的审核与审核教程相同，对其内容的匹配性合法性进行审核。由于审核过程中没有发现，之后遭到用户投诉的，管理员可以强制禁用某个教程，也可以恢复启用。管理员可以添加、修改教程模板，所有的分享型用户在新建教程时就可以使用它来快速创建教程。管理员可以查看所有教程的点击量汇总、广告分润收益汇总，来观察和改进教程的运营。

（3）教程展示

教程通过审核后即发布到系统供所有学习型用户进行查阅。查找教程是学习的第一步，用户在这里可以通过分类目录、最新教程、最热教程等方式找到自己

想要学习的内容。用户还可以通过关键字搜索教程，支持多关键字的组合搜索。用户还可以根据热门排行选择教程，热门排行是根据该教程的实际点击量进行统计的，并且包括全部分类排行和指定分类排行。系统还会根据教程最新的发布时间排序，给出最新教程排行。用户可以在系统中获得个性化的推荐排行，每位用户的推荐排行都是根据用户自身的学习轨迹分析后得出的，这也是系统的智能功能之一。

用户找到相应的教程后可以通过看视频、看文字、做测试来学习，教程的文字和视频的字幕还支持有不同的语言版本，系统会记录当前的学习进度，例如每一个教程最近一次浏览的章节，每个视频最近一次播放的位置。用户在学习过程中，如果发现有重要的部分，还可以使用笔记功能，通过笔记功能可以实现截图、备注、涂鸦等。当用户来不及学完教程或者希望以后再次进行学习，还可以使用收藏夹功能，被收藏的教程当有更新时还会以邮件、系统消息的方式提醒用户，用户收藏教程的偏好被记录到系统，作为分析用户学习行为的依据，从而发现规律，在用户学习、收藏时以此规律找出用户可能感兴趣的教程提供给用户。用户还可以分享系统的某一篇教程到微博、博客和社区网站，也可以直接在教程的互动区进行评论、提问，如果是提问，系统通过对问题的分析，给出可能正确的参考答案。

模拟测试是学习教程之后检验知识点是否掌握的重要环节。用户学习完毕教程之后，根据教程的参数设置，教程可能会提供可选的测试环节或强制的测试环节，一般来说属于系列教程的要求顺序学习和需要达到一定的测试成绩才能继续下一课程。整个题库分为若干个知识点，教程制作人指定了每个知识点占整个测试的比重，每次测试都是从教程附带的题库中随机生成试卷，系统会按照指定的比重随机抽取知识点的试题。通过测试题进行测试后，用户除了获得数字成绩以外，还会获得一份测试报告，告知用户对知识的掌握情况，即对个知识点测试的正确率。

### 3.1.2 智能答疑模块

智能答疑模块是为增强用户学习的连续性、减少人工答复而设计的智能模块之一，体现了系统的智能优势。用户在学习过程中，常常遇到很多学习问题，系统过去的方式是通过论坛或者特定的栏目进提问，由人工进行答复。本次设计增了该智能答疑模块。系统的功能用例如图 3-3。

对于普通用户，在学习的过程中可以对遇到的问题进行提问，可以对系统给出的答案进行评分，用户对无论是系统自动答复的答案还是人工回答的答案，可以选择有限个答案进行评分，且只有一个最佳答案，评分记入系统后将会影响以

后的智能答疑的结果。用户可以同时是问题的提出者，也可以是其他问题的回答者。

分享型用户和系统工作人员也需要对一些问题进行回答。当学习型用户觉得当前给出的答复不满意时，问题都处于待答复状态，除了系统实时的智能答疑，更了解该教程的原理和知识点，回答选的功能之一，官方发布的教程都有提问，系统工作人员在这里进行解答，例如访问时遇到问题，教程无法显示等系统问题。智能答疑模块功能用例如图 3-3。

在正文引用出所用图的编号，注意不要用“上图”、“下图”字样！句子后用句号，不要用引号！

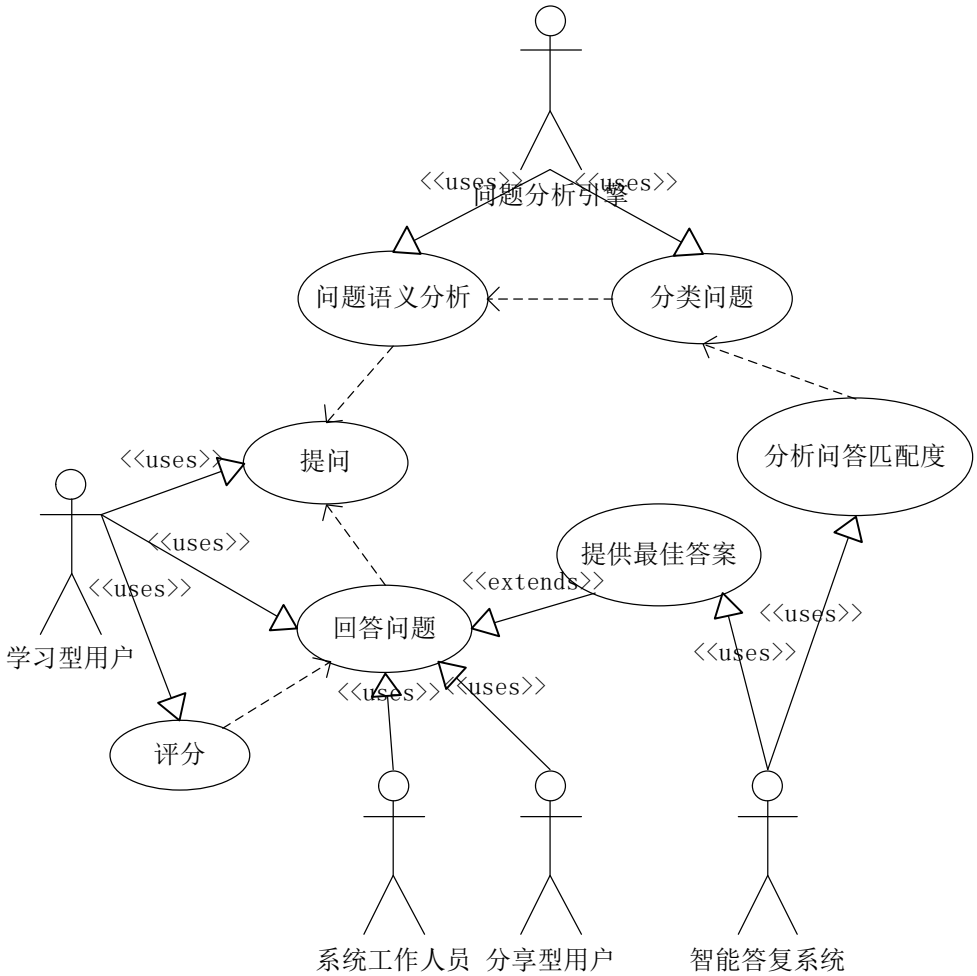



图 3-3 智能答疑模块功能用例图

问题分析引擎在智能答疑模块中是进行语义分析和问题分类的模块。用户的提问可以通过问题分析引擎进行单词分解、分析后得到一定的语义，根据一定规则，这些语义对应着一些问题分类。问题分析引擎负责把用户抽象的问题具体到某个问题的分类，为之后的智能答疑提供了数据基础。

智能答疑模块是实时的、自动的问题回复模块。智能答疑模块根据问题分析

引擎提供的问题分类，和该分类历史的匹配情况，按照匹配度的高低进行排序，把结果返回给用户，有用户人工确定问题的正确性。用户也可以不采纳提供的备选答案，等待人工答复。一般来说，有过人工答复或[illegible]答疑问题，智能答疑模块都能在系统中进行比较高的[illegible]

1、**不同级标题**之间不要直接相连，其间用一段总括的句子隔开。
2、全文所有的标题用黑体、加粗，不同级标题之间字号逐级大 1 号！

### 3.1.3 互动社区模块

互动社区模块是为用户和用户之间，用户和系统[illegible]之间进行交流互动的子系统，它形成了以用户为中心的社交网络，是教程多人协作开发的前提，是系统内容自更新的基础。互动社区模块功能用例如图 3-4。

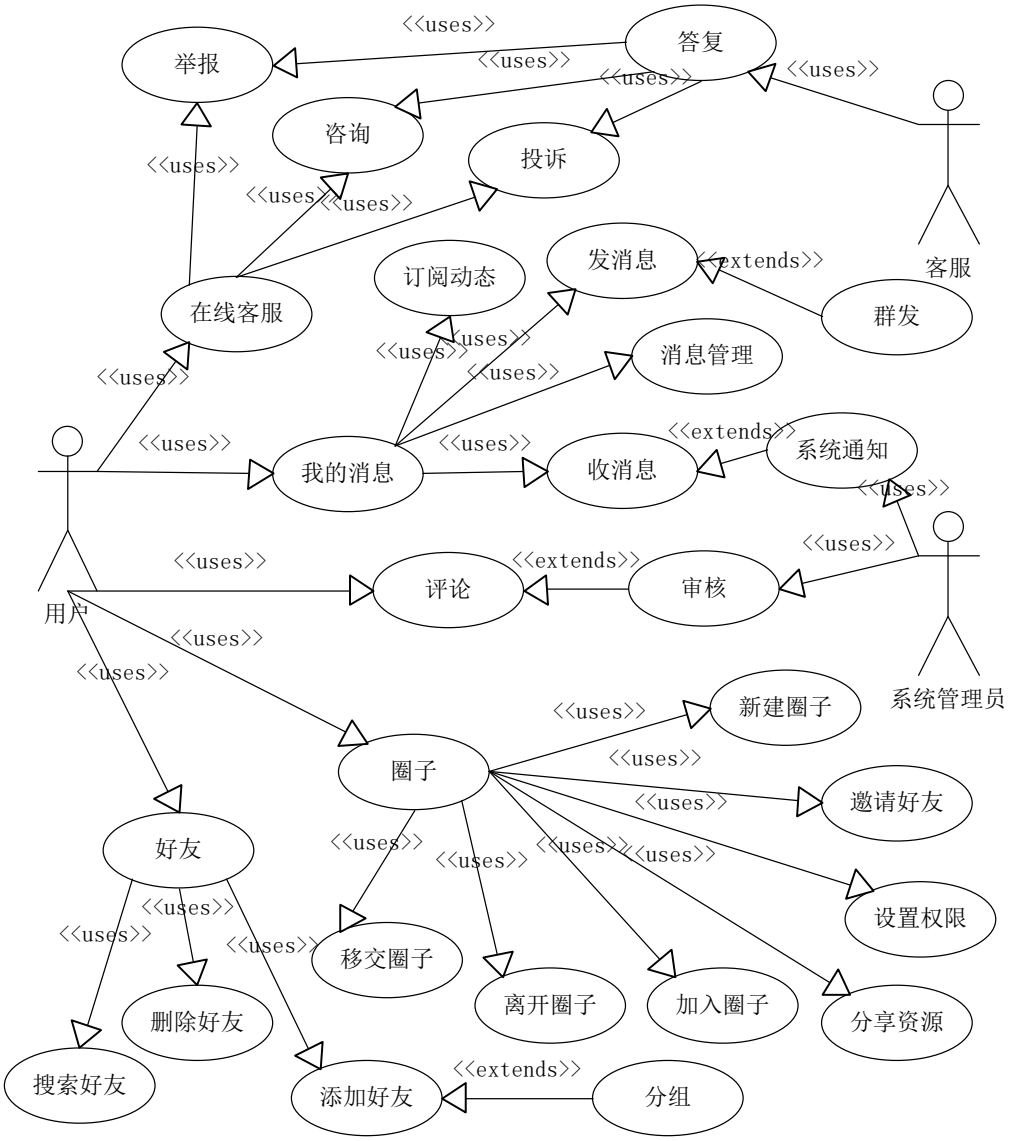



图 3-4 互动社区模块功能用例图

在互动社区模块中，涉及到普通的学习用户、客服、系统管理员，包含了在线客服、我的消息、评论、圈子、好友等主要模块。

在线客服是系统运营的最后一环，可以通过该模块获得直接的用户反馈。作为普通的用户，可以通过在线客服反映使用中遇到的问题，可以对非法的言论进行举报，可以咨询使用问题、商业合作，可以投诉系统故障、服务态度等。系统客服专员需要在规定的时间内对用户提出的问题进行答复，遇到无法处理的问题可以转移给系统管理员。采用在线答复的方式代替电话，为公司节省了大量成本，而且在线客服还可以在线截屏、传送文件，并对客服事件比电话客服有更好的记录，从而可以大大提高服务效率和质量。

我的消息模块是给用户和用户之间及时沟通设计的，同时也支持收取系统信息。通过消息模块，普通用户可以给其他用户发送消息，可以订阅动态消息，诸如某个教程有更新、自己团队的教程有更改、自己的问题有回复等消息。系统管理员可以使用消息模块来发布系统通知。

评论是保证教程质量的一个重要手段，用户可以使用该模块，评论该教程的效果、指出错误、进行讨论等，每个用户都有一次机会对该教程进行评分，可以多次以文字的形式进行讨论和评价。系统管理员审核评论的文字是否合法，如有人生攻击、非法言论等则进行屏蔽。

好友功能是为用户和用户之间提供点对点的交流服务。对学习型用户来说，可以实时查看好友的动态，了解好友的在线情况，还可对同一门课程进行测试分数比拼，提高学习的趣味性。用户可以为好友分组，一个好友可以重复出现在多个组。系统会根据关联分析，找出和用户兴趣相同的好友推荐给用户。

圈子是系统群体智能的基础，为兴趣相同的用户建立一个群，提供了用户之间的学习交流、教程制作等社交和协作功能。对学习型用户来说，圈子可以为用户参与一些感兴趣的主题提供了可能，且能在圈子中认识更多的好友；对于分享型用户来说，可以在圈子里建立组织，加入成员，映射成员到相应的教程中，达到多人开发的目的，组织里支持群体聊天、文件共享、修改日志等协作功能，任何一个成员都可以自由的加入和离开某一个组织。系统会根据关联分析，找出和用户可能感兴趣的圈子推荐给用户。

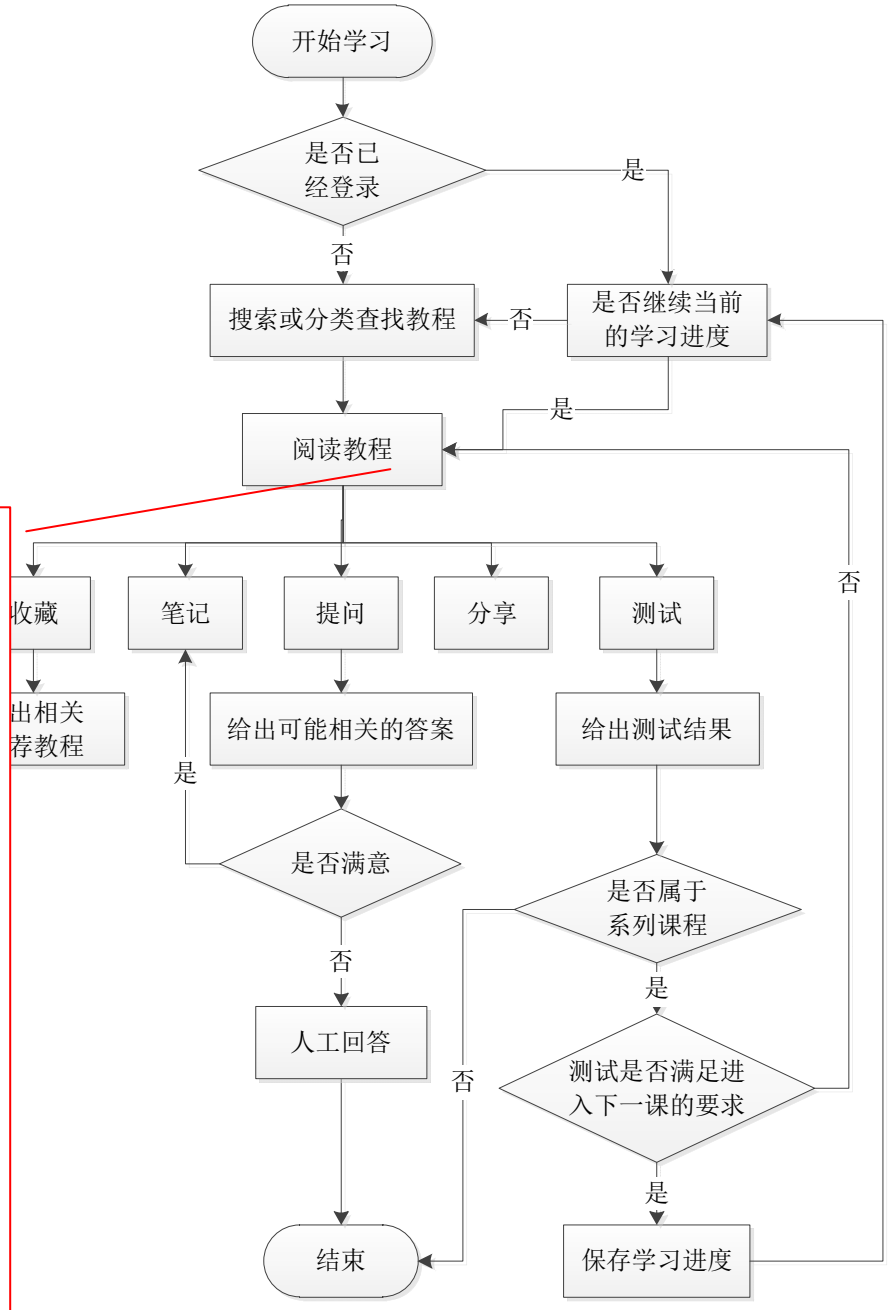
## 3.2 智能在线培训系统核心流程

系统的流程中，业务核心流程是教程学习流程、在线答疑流程、社交关系管理流程。教程学习流程是指用户学习单课程或者一系列培训课程时的主要流程；在线答疑流程是指用户对有疑问的问题进行提问，系统自动给出答案，或者由教程发布人、其他用户进行回答的主要流程；交关系管理流程主要是指用户对好友、

圈子、协作等社交功能的管理的主要流程。

### 3.2.1 教程学习流程

系统的教程学习流程主要包含：教程搜索、教程展示、教程测试以及一些学习辅助操作，如图 3-5 所示。




1、图居中，边距不要超过正文边距。
2、图及其名称的字号比正文小 1 号，即 5 号字！
3、图的编号各章统一给定！
4、图表的名字应该与图表位于一页，图的名称放在图下面，而表的名称放在表上面

图 3-5 教程学习流程图

进入教程学习模块，系统首先会判断用户是否已经登录，如果没有登录会提示用户注册或登录，也可以跳过该步骤直接进入下一步搜索教程并学习；如果已

经登录，则会判断是否有学习进度被保存，如果有，会提示用户是否继续上一次的学习进度，如果继续学习进度则直接进入上次的教程，如果用户选择否，则会进入下一步搜索教程并学习。

学习教程可以通过观看视频和阅读文字来完成，系统允许只有视频或文字，也支持两者都有。打开教程如果包含视频，系统会自动开始播放，并会根据当前所选的语言显示字幕，如果没有视频，则会显示相应语言的文字版教程，如果只有一种语言版本则显示原始版本，否则显示符合用户的语言支持的版本，如果没有找到符合用户的语言支持则显示英文教程或作者第一语言的教程。

学习过程中可以做笔记、可以评论、可以提问。提问后系统会尝试进行智能答复，显示可能匹配的答案，并按照曾经用户评价的匹配度排序，如果用户满意该答案可以进行答案匹配度评价，同时可以记录到笔记中，如果没有满意的答案则等待其他用户或者教程发布人回复。

教程可以被用户收藏以便日后查找和复习。收藏之后系统会在收藏成功提示页面给出用户可能感兴趣的教程，用户可以通过该方式可以不断的进行收藏新教程，自动推送感兴趣的教程，避免了用户在茫茫的教程库中不断的搜索。

学习完成之后系统会判断该教程是否包含题库，如果包含会则提示进入测试，完成测试后会给出成绩和知识掌握情况报告，如果不是系列课程那么就完成了教程的学习，如果是系列课程并且分数达到进入下一课程的标准，那么保存学习进度并询问是否继续学习，否则需要重新测试直到获得合格的成绩才能进入下一课。用户可以对同一课程进行反复测试，系统每次给出的题目都是随机的，但是题目覆盖哪些知识点和覆盖的比率是教程发布者指定的。

### 3.2.2 在线答疑流程

系统在线答疑流程是体现系统智能性的一个流程，和传统的用户提问老师回答不同，该流程优先进行系统自动回答，再者是教程发布人回答，还支持其他用户回答。用户对答案进行评价后会自动增加到问题库，使得该功能会以自动、智能的方式进行扩展和优化。

当用户在学习一门课程或进入用户中心时，可以触发在线答疑流程。学习课程时可以对当前的课程进行提问，或对其他用户提出的问题进行回答；在进入用户中心时，作为学习型用户可以看到所有曾经参与过的问题，无论是提问还是回答，作为分享型用户可以看到有多少问题等待人工答复。

用户每次新增加一个提问，系统都会在进行问题语义分析，确定问题的分类以后，通过关联分析和比较该分类问题所有的匹配答案，给出系统认为正确或者密切相关的几个答案，如果用户认可答案，需要对该答案进行评分，该问题状态

将改为已答复；如果用户觉得答案不正确可以选择系统重新给出答案或者等待人工答复。凡用户进行评分的答案都会记录到问题库，为下一个用户的提问作为参考。

当学习型用户进入用户中心时，可以看到没有答复的问题，此时可以再次尝试让系统给出答案，如果仍然没有需要等待人工答复，或者取消该提问。当分享型用户进入登录中心时，可以看到有多少问题等待答复，此时可以进行人工答复，该答复如果得到认可，将会被评分并记录到问题库中。具体流程如图 3-6。

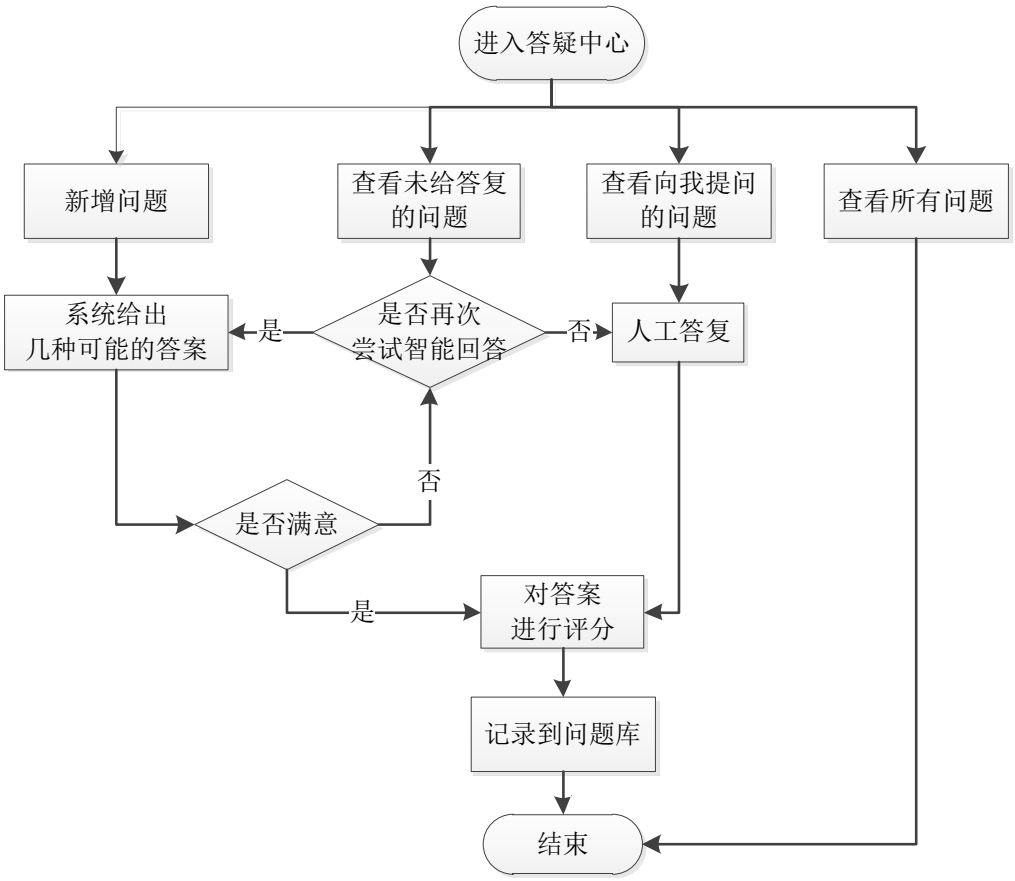



图 3-6 在线答疑流程图

### 3.2.3 社交关系管理流程

社交关系管理只针对登录的用户进行授权操作，用户登录之后可以管理自己的个人资料、好友、圈子、协作等信息，主要流程见图 3-7。

资料管理是对用户公开到社区中的信息进行的管理，包含用户的用户名、昵称，还包括用户愿意公开的性别、年龄、学校、专业、行业、职务、邮箱、及时

聊天工具、微博等信息，用户可以对部分信息设置为隐藏。每次输入个人资料系统都会进行校验，例如邮箱是否是用户自有邮箱，学校和专业是否存在，微博、及时聊天工具名称是否合法等，确保用户信息尽量真实有效。

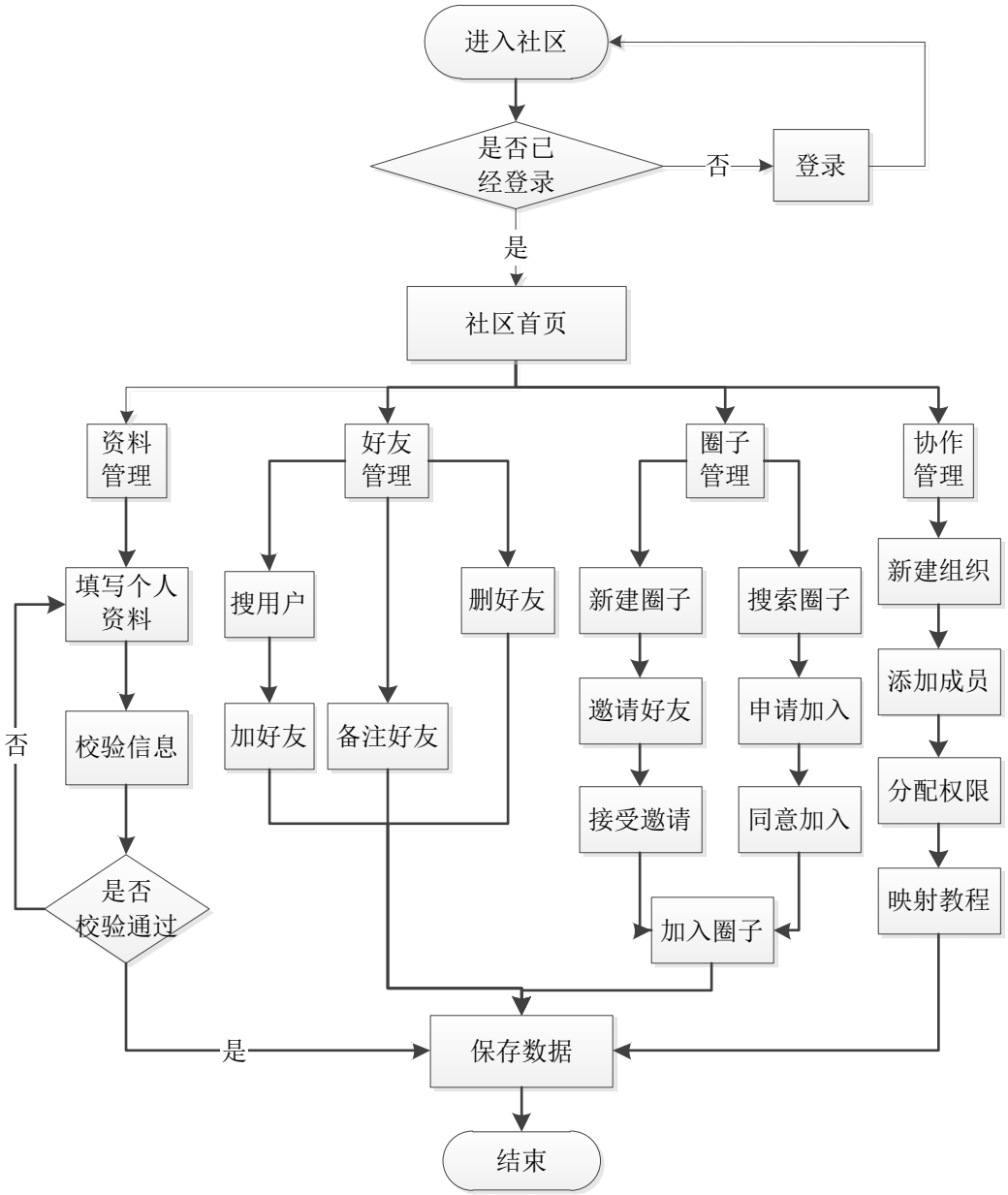



图 3-7 社交关系管理流程图

好友管理提供了简单的类似及时通讯软件的功能，为用户提供好友搜索、添加好友、删除好友、查看好友实时动态、查看好友在线情况、给好友添加备注等功能。圈子管理和好友功能不同，好友是用户与用户之间点对点的社交方式，圈子是用户与用户群点对面的社交方式。圈子为用户参与各种学习主题提供了便捷的方式。每个用户都可以新建一个圈子，为圈子命名、分类，然后邀请好友加入，或等待其他用户主动申请，管理员同意后加入圈子。圈子管理员对圈子有最高的

权限，可以删除用户、更改名称、移交管理权等。协作管理是为多人开发教程提供的协同工作模块，用户可以新建组织来管理协作成员，通过添加成员、分配权限、映射成员和教程的关系完成协作管理。

## 3.3 智能在线培训系统的安全需求

整个系统拥有大量的用户个人信息、广告商的资金信息、分享型用户的收益信息，一旦出现安全问题将会导致大量的用户隐私泄露，而且金融信息的丢失将会带来不可估量的损失。系统是对整个互联网公开的，这也决定了它存在较大的安全隐患。为了确保系统的安全性，需要从终端设备、数据传输、业务系统和系统管理等方面进行考虑，建立完善的安全体系。

（1）终端设备的安全

个人电脑、智能手机等终端设备面临的安全威胁主要在于设备的丢失、非授权使用、来自互联网的攻击和入侵等。这些威胁可能由公网用户入侵或使用者人为造成。因此，终端设备需要有一套安全有效的身份认证和访问控制，用户在进行相关操作前需要登录，保证只有合法的用户才能使用终端访问系统，而且在进行资金等重要操作时需要双重认证，即除了登录密码之外在验证动态密码。

（2）数据传输的安全

由于系统是基于互联网的，所以敏感信息和一些重要数据在传输过程中需要采用相对安全的协议进行传输，对数据传输前需要采用一定的加密方法，并且数据在传输到后台服务后进行校验防止篡改，保证数据的完整性和安全性，这包括终端设备和后台服务的通信、服务和服务之间的通讯都采用默认不信任的安全原则。

（3）业务系统的安全

对于密码等敏感信息，采用合适的加密算法加密后再存入数据库，即使信息被盗，也能够进行保护。所有的业务模块都合理的设计权限访问机制，只对有相应权限的用户开放访问权限，并且对异常 IP、异常流量、大额充值提现等操作进行监控、主动报警，避免由于设计不完善带来的安全问题。

# 第四章 海策智能在线培训系统设计

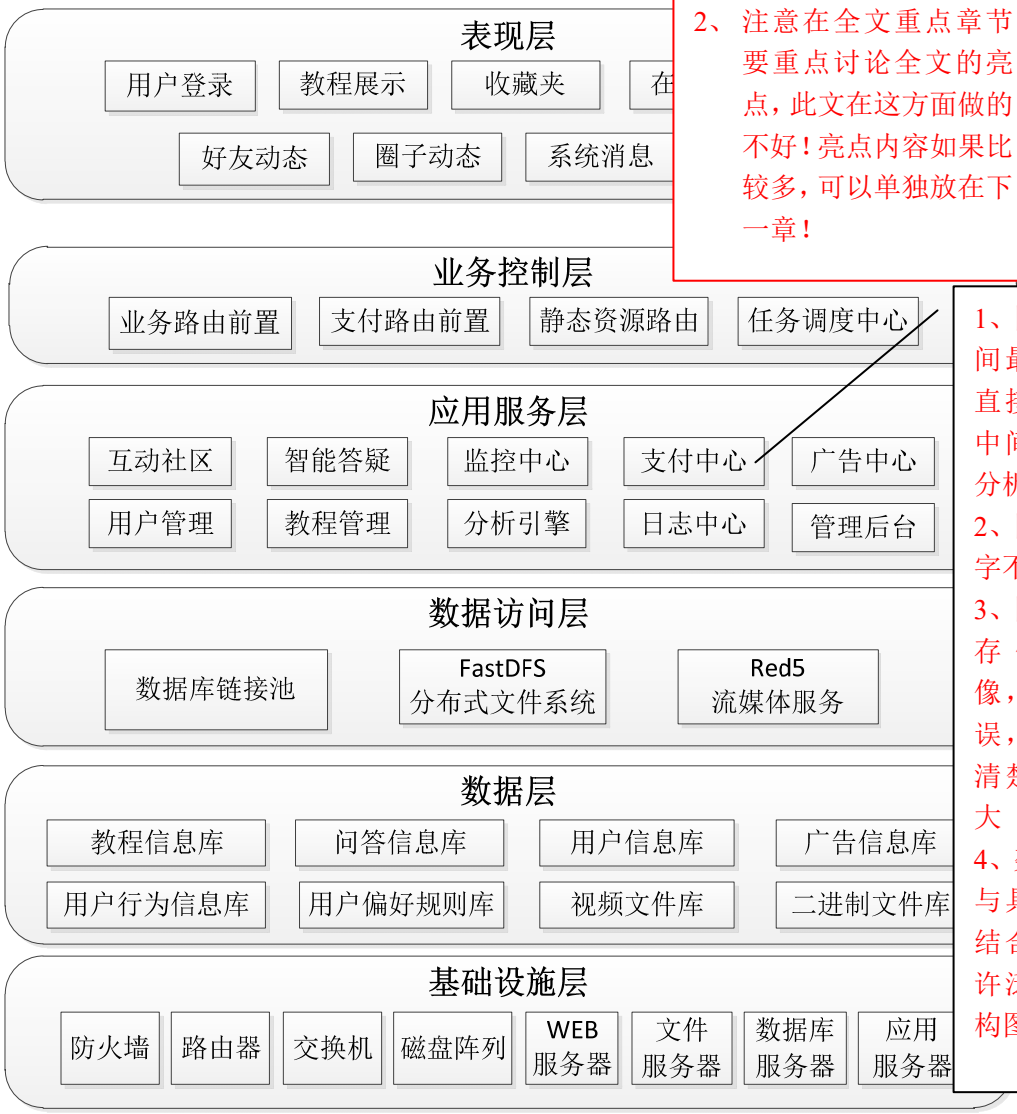
通过前一章对系统整体设计及需求的分析，确定 将详细介绍海策智能在线培训系统的系统架构、核心子 ，并对关联分析在子系统中的数据采集、数据挖掘、成 详细介绍，最后会比较同类系统优缺点并分析系统的应

1、开发类论文系统设计与系统需求都是全文的主要组成部分，注意比较分析，而不是简单介绍！系统设计主要包括系统的架构、核心子系统（尤其与重点内容相关的内容）等。如果数据库的设计在某方面有特色，可以简要介绍一下！注意功能和子系统的区别。

2、注意在全文重点章节要重点讨论全文的亮点，此文在这方面做的不好！亮点内容如果比较多，可以单独放在下一章！

## 4.1 智能在线培训系统总体架构设计

海策智能在线培训系统架构分为基础设施层、数 服务层、业务控制层、表现层。系统的体系架构如图 4




1、图或表之间最好不要直接相连，中间用文字分析隔开

2、图表中文字不能太小

3、图表不要存储为图像，此文错误，打印不清楚且文件大

4、架构图应与具体项目结合，不允许泛泛的架构图

图 4-1 海策智能在线培训系统体系架构图

基础设置层为整个系统提供了硬件和操作系统的支持，包含了网络设备和各种服务器，系统服务器包含了 Linux 和 Windows Server 2008 两种操作系统的机器。数据层依赖基础设施层的硬件和操作系统，数据层包含视频文件、数据文件、和数据库管理软件，其中数据管理软件包含 MYSQL 和 MS-SQLSERVER 两种，是为满足不用操作系统下的应用服务的需要。应用服务层通过 ODBC 和 JDBC 访问数据层的数据库，通过 FashDFS 分布式文件系统访问文件数据，通过 Red5 流媒体服务器支持对 Flash 的视频文件的访问。

海策智能在线培训系统的客户端采用 HTML5、CSS、FLASH 等技术展示教程内容。多种客户端的访问方式，但对系统兼容各设备提出了较高的要求。应用服务层，需要通过业务控制层，其中业务控制层判断路由、转发数据给应用层的业务系统，由应用层后返回结果给业务控制层再反映到表现层。

注意全文分析的力度，增加比较分析的内容，以便使论文有学术性。多问几个为什么？例如为什么用这样的架构，给定集成方案，通过比较分析确定一种比较适合的

系统的应用层采用面向服务的体系结构，面向服务的架构很好的封装了各个业务子系统，各业务对外以提供标准接口的方式提供服务，服务之间的松耦合给系统的稳定性、灵活性、扩展性都带来了很多好处。应用层中的所有服务通过在服务管理中心进行管理，由服务管理中心对各个服务进行注册登记、检查服务状态、为其他服务提供接口检索，解决了大量的服务之间调用关系复杂不便管理的问题。

海策智能在线培训系统设计为 B/S 架构，出于成本考虑数据库大部分都采用 MYSQL，因为历史原因也保留了部分 MS-SQLSERVER 的数据库。编程语言采用 JAVA，每个子系统都采用 MVC 三层架构，并使用了 SpringMVC、Spring、IBATIS 成熟的第三方框架，前端采用 HTML5、Flash、AJAX 展示培训内容和进行用户交互的实现，服务之间采用 WebService 进行通信，为了增强系统的安全性，除了嵌入的视频、图片等资源外在通信协议上基本都采用了 HTTPS 代替 HTTP。

系统对用户行为数据进行采集和整理，根据系统需要解决的智能功能的业务目标对数据进行分析，再使用这些分析结果应用到实际的功能呢和场景。通过以上三个步骤实现系统的智能功能，并明确各个子系统的职责。

以上的步骤具体表现为下面要介绍的三个子系统：用户行为数据采集子系统、用户偏好分析子系统、智能助教子系统。其中智能助教子系统是最终关联分析的成果应用子系统，提现了部分人工智能的特点，具有较基本的推理、学习、交流能力。

下面围绕智能功能实现的三个步骤对相关子系统进行设计，三个子系统的关

系如图 4-2。

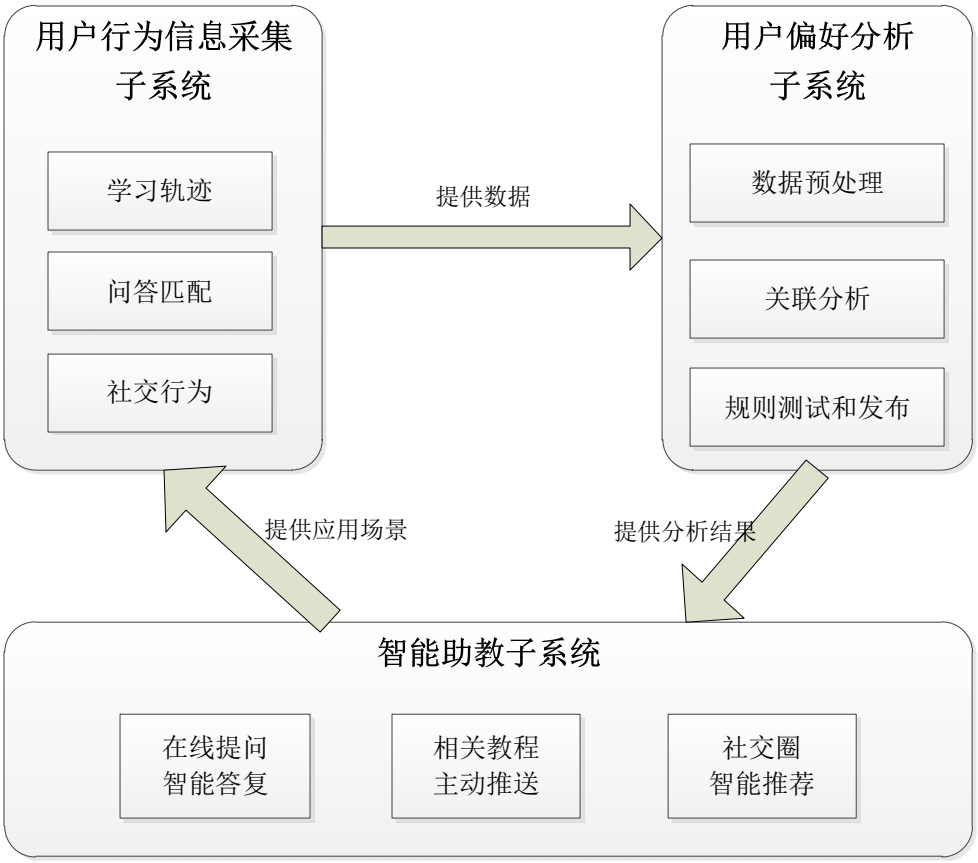



图 4-2 智能相关子系统结构关系图

用户行为数据采集子系统在关键的事件中设计一些数据采集点，对用户的学习轨迹、在线提问、社交行为的数据进行收集，为用户偏好分析子系统提供数据样本；用户偏好分析子系统对数据进行预处理、数据分析、获得关联规则、测试和发布规则，为智能助教子系统提供规则依据；智能助教子系统为系统其它子系统提供智能应用场景，同时也影响下一次的数据，形成一个良性的循环。

## 4.2 用户行为数据采集子系统设计

通过对海策智能在线培训系统的需求分析，可以发现要实现该系统的智能功能，需要基于关联规则对用户学习偏好、问题和答案的关联度进行分析获得他们的关联关系，这些关联关系需要通过数据挖掘生成，而想得到正确的关联规则需要大量的样本数据进行分析，数据采集的设计直接影响到关联规则生成的质量。

数据挖掘需要的样本数据很多，并且分布在各个业务系统中。智能推荐需要的数据主要来源于用户档案资料、学习历史信息、教程收藏信息；智能答疑需要

的数据主要来源于问题库、答案库。如此大量的数据要如何使用，首先需要对系统中所获得的数据信息进行研究、分析数据的属性以及包含的意义，下面对所需数据的采集进行设计。

### 4.2.1 用户学习轨迹数据采集

要实现为用户智能推荐相关教程需要了解所有用户的学习行为，而通过收集用户的学习轨迹数据可以用于分析用户学习行为。用户学习的轨迹是教程搜索、浏览教程、模拟测试、在线提问，同时还提供了很多学习工具，例如笔记、收藏、评论、相关教程推荐等，经过分析在系统进度管理模块、收藏夹模块记录用户的行为，为数据挖掘进行数据收集，用户学习轨迹交互时序图如图 4-3。

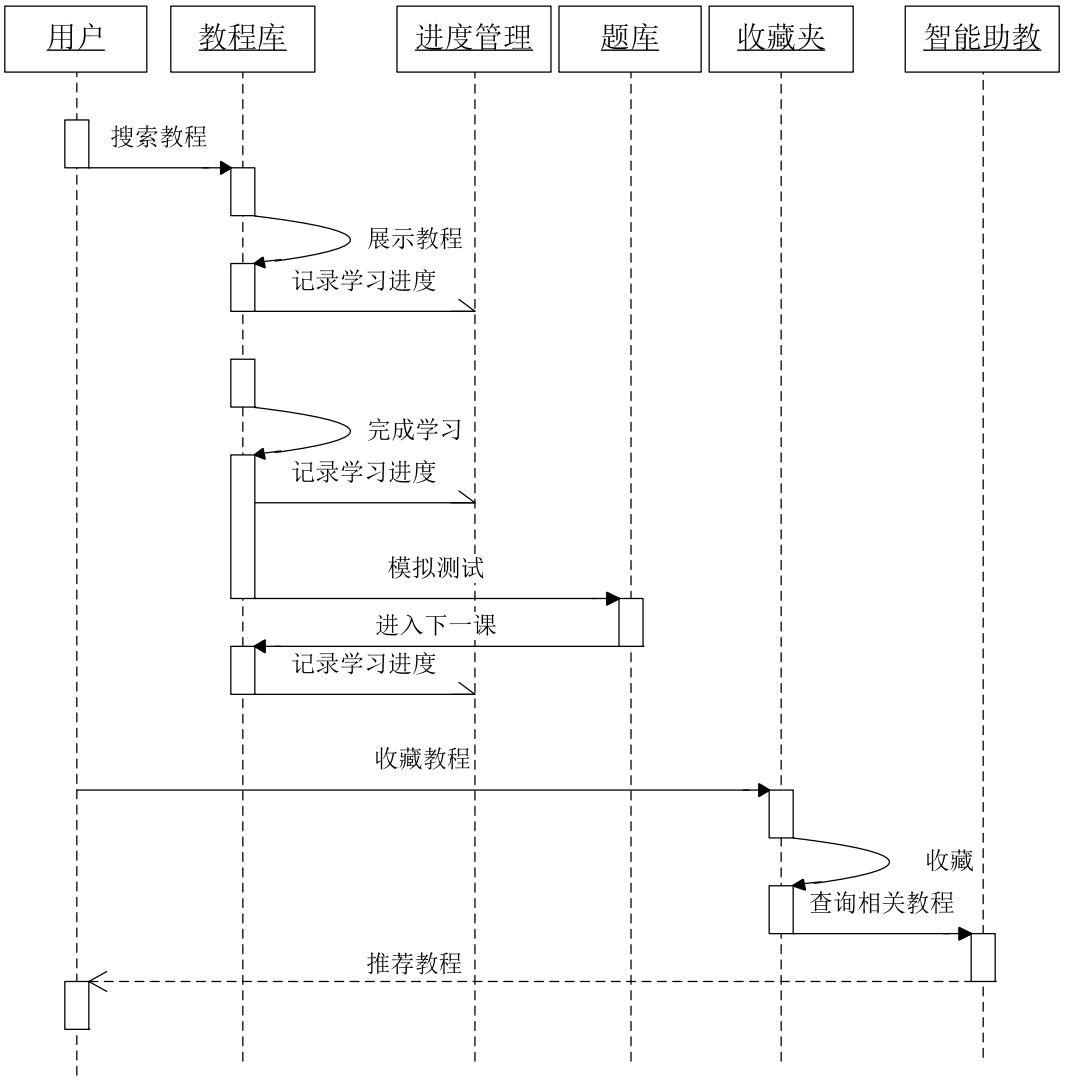



图 4-3 用户学习轨迹交互时序图

（1）分析用户学习轨迹数据流

教程库对外提供的接口包括关键字全文搜索、获得热门教程排行、获得最新发布列表、获得指定编号的教程详细内容、记录教程点击量等主要接口。用户需要学习一门课程，首先需要找到课程，根据用户输入的关键字，系统会调用全文搜索接口获得查询结果，在首页和查询结果等页面，系统都会调用接口获得热门排行列表、最新发布列表，这些接口完成用户的检索操作。然后进入该教程的学习页面，系统根据教程的唯一编号调用接口获得教程的详细内容进行展示，之后用户可以通过看视频、看文字、选择语言、选择字幕来更改教程的展示效果。

进度管理模块为记录一些用户学习进度提供了以下接口：记录用户教程的唯一编号、记录用户视频播放位置、记录最近一次学习的课程、记录系列课程学习当前到达的进度、记录考试成绩、记录考试次数等。用户在学习过程中，每次浏览任何一篇教程系统都会调用接口进行记录，并且记录了用户浏览该教程停留的时间。每次用户新登录系统，系统都会获取最近一次学习的课程，提醒用户是否直接继续上一次的学习。学习完毕后，系统会调用进度管理模块的接口记录学习进度，如果该教程有测试题库会提醒用户进行测试，否则自动进入下一课，如果没有下一课则完成本次学习。

模拟测试模块需要提供的接口有判断某个教程是否有测试题库、获得测试配置信息、生成随机试卷、计算和保存测试成绩、生成测试报告、获得历史测试结果等。每次用户完成一个教程的学习，如果该教程包含了测试题库，则会提醒用户进行测试。测试是否必须需要教程发布人预先设定,是在教程制作时配置，一般系列的课程都要求完成测试并达到一定的成绩才能算完成本课的学习任务。系统先调用是否需要测试接口，该接口会判断是否有题库而且是否已经完成测试，如果需要测试，则显示开始测试按钮。进入开始测试，系统将会调用题库模块生成一张随机的试卷，试卷的知识点覆盖比例是教程发布人之前指定的，测试完毕之后系统将记录测试成绩和生成测试报告告知知识点掌握的情况，并根据成绩判断是否可以进入下一课。每个学生都可以进行多次测试。

收藏夹模块需要提供的接口有收藏教程、删除收藏教程、为收藏教程添加标签和删除标签、获得已收藏教程的更新动态。用户无论在哪个学习的阶段，都可以对教程进行收藏，收藏夹模块会记录这些信息，在收藏成功的同时，系统还会分析被收藏教程和其他教程的相关性，然后推荐和收藏的教程相关的教程。已经收藏的教程会定时查询并提示用户是否有更新。用户可以删除已经收藏的教程，可以对已经收藏的教程添加和删除标签，这些标签为用户提供了分类和快速查找的功能。

（2）定义学习轨迹数据采集方式

经过以上的分析，可以得知记录用户的学习轨迹数据主要分布在进度管理、收藏夹中。为了获得更准确的学习轨迹数据，需要对这两个模块所记录的数据确定数据的采集方式、定义更丰富的属性。

在进度管理模块中，对用户任何一次浏览教程、每次学习教程、每次完成教程的学习都需要做记录，并且包含每次在教程的停留时间，采用异步提交进度数据保存在学习进度信息表中，减少对原来流程的性能影响。这些信息是用户学习行为最基础的数据，这些学习进度数据和教程的分类情况能反映出用户的学习偏好，即喜欢或者需要学习什么类型的教程，大部分时间都在学习什么方面的内容。学习进度信息包含了学习开始时间、学习结束时间、教程唯一编号、教程章节、视频播放位置、教程所属分类、测试次数、测试成绩、最近一次学习位置。学习开始时间和学习结束时间可以计算出停留时间，反映了用户对该教程的兴趣度，教材的分类反映了用户对各分类的兴趣度，测试次数和测试成绩可以反映出教程对该用户的难度。

在收藏夹模块中，用户可以收藏感兴趣的教程。教程收藏信息需要去掉重复的收藏信息。教程收藏信息包含教程唯一编号、教程名称、教材所属分类、收藏时间。教程分类反映了用户兴趣的分布，结合用户的年龄、职业、专业等可以得出很多关联关系，这些数据和学习进度一样可以作为发掘用户学习行为和学习方向的数据依据。

### 4.2.2 问答匹配结果数据采集

要实现对用户的提问智能的给出匹配的答案需要了解所有的问题和答案的匹配关系，问题和答案的匹配关系包含答案对应的问题，它们之间是多对多的关系，匹配度是根据用户对该问题下的答案评分来衡量的。经过分析需要在系统问题库、答案库、答案评分模块记录问答匹配数据，为数据挖掘进行数据收集，在线答疑交互时序图如图 4-4。

（1）分析在线问答数据流

在线答疑对外提供的接口包括问题列表查询、问题关键字查询、问题关联答案查询、问题状态修改、设置问题匹配答案、答案评分等主要接口。在用户提交问题后，系统会调用问题分析引擎对问题进行语意分析，确定问题的分类，根据问题分类系统会在已有的关联模型中找出最佳答案并返回给用户。

用户获得智能回复的答案后，如果用户认为找到了匹配答案则需要对答案进行评分，否则该问题总是显示为未解决；如果用户认为没有找到匹配答案，可以等待人工答复，教程发布人和其他网友都可以进行答复，但智能答复是实时的，而人工答复周期较长，甚至有可能没有答复。用户评分完毕将会保存问题与答案

的关联关系、答案的评分、修改问题的状态为已答复。

问题库中记录了所有用户提过的问题，包含了问题唯一编号、问题名称、问题时间、问题是否已经答复。答案库中记录了所有回答过的问题，包含了所属的问题唯一编号、回答时间、回答人、回答内容。答案需要和问题结合才能够反映出有意义的信息，答案和问题是多对多的关系，一个问题可以有多个答案，一个答案也可能同时对应多个问题。答案评分是独立的模块，记录了答案的每次所得评分。

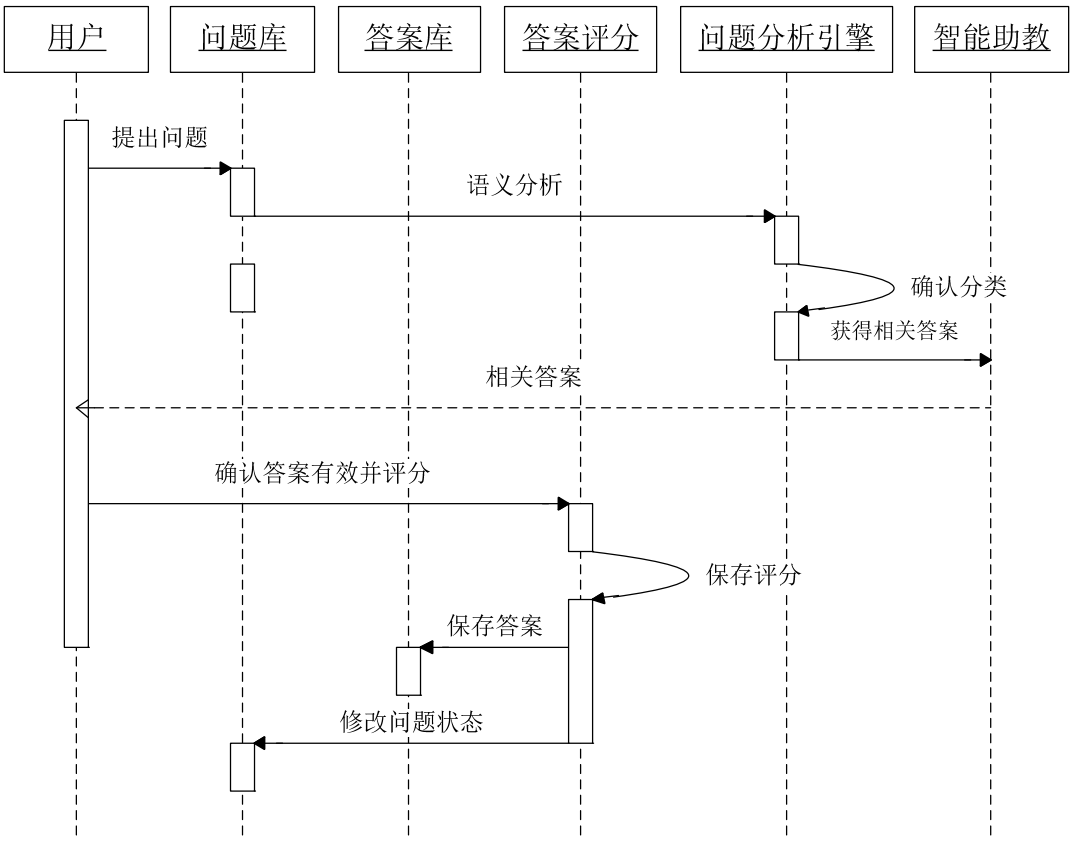



图 4-4 在线答疑交互时序图

（2）定义问答匹配的数据采集方式

经过以上的分析，可以得知记录问答匹配结果数据主要分布在问题库、答案库、答案评分中。为了便于之后的数据挖据，需要对这三个模块所记录的数据确定数据的采集方式、设计数据的属性。

在问题库中，仅仅记录问题的名称远远不够，还需要增加问题关键字、问题分类。其中问题关键字和问题分类不是用户输入的，而是系统调用问题分析引擎经过某种规则和算法自动生成的，问题的名字最终被分解为若干关键字，问题关键字的某种组合形成一个问题分类，问题分类是查找关联问题的关键，相同的问

题分类表示可以使用相同的答案。

在答案库中已经记录了答案与问题的关联关系，有关联关系的答案都是智能回复的备选答案，是经过用户验证的有一定匹配度的答案，在数据采集过程中，为了更好给数据挖掘计算时提供更好的反应答案匹配度的数据，每次记录答案的评分是，还需要记录该答案与该问题之间的恶平均评分、最高评分、最低评分。其中某次的评分反映了作为某个问题该答案的正确率；最高评分和最低评分可以反映出匹配的正确度，评分越高表示越接近正确，反之则表示不太正确；平均评分反映了答案和问题关联的正确性，平均评分低说明答案匹配给了很多不相关的问题。

### 4.2.3 社交行为数据采集

要实现为用户推荐相同兴趣的好友、可能感兴趣的社交关系需要了解该用户的背景资料、已经拥有的好友、已经参与的圈子。经过分析包含这些数据的模块有用户档案、好友管理、社交关系管理，用户社交管理对象交互如图 4-5。

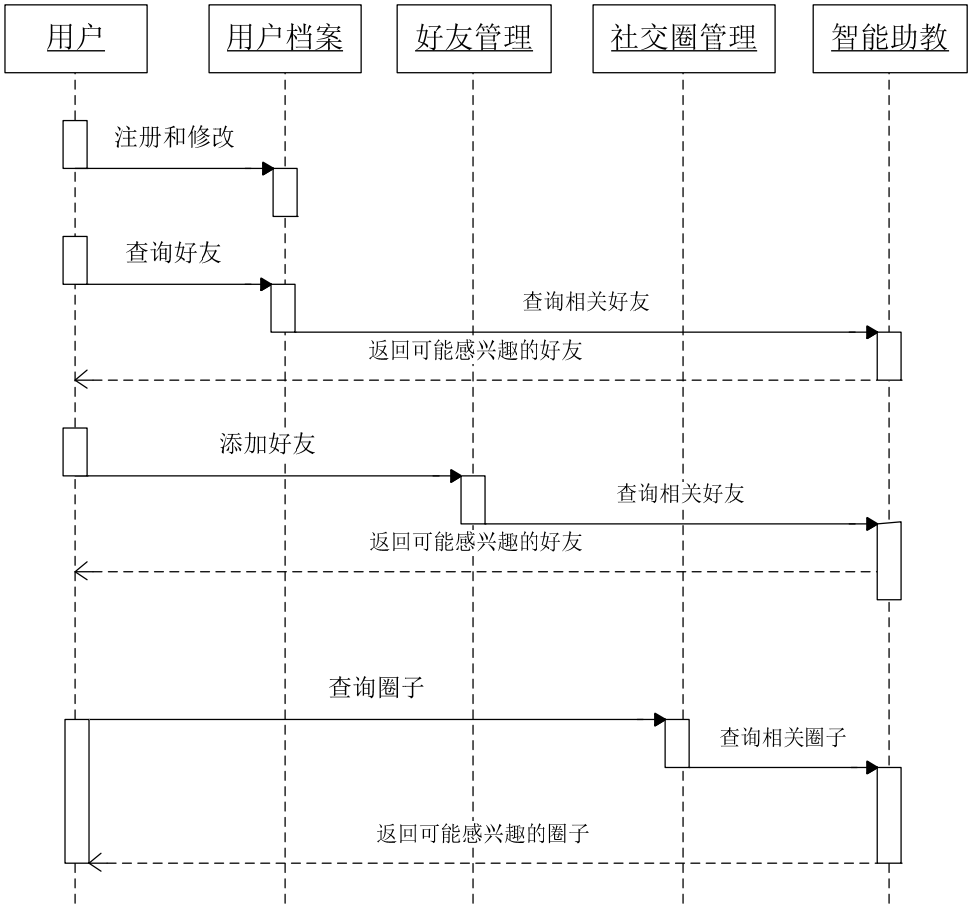



图 4-5 用户社交管理对象交互时序图

（1）分析用户社交行为数据流

用户档案在用户注册时产生，可以登录系统后进行修改。用户档案信息是组成用户管理模块最基础的数据，系统通过返还积分和各种物品奖励，鼓励用户填写真实的资料，所以系统中的用户档案相对还是比较完整和真实的，这些信息反映出了用户的行业分类、年龄段等特征。

好友管理模块提供的接口有查询好友、添加好友、删除好友、设置权限、查看好友动态、查看好友在线状态、所学教程的好友得分排行等。好友管理模块中记录了好友账号、昵称、备注、好友分组等。用户在查询好友和添加好友时，系统都会根据目标好友的一些特征进行分析，推荐可能认识或者感兴趣的好友，用于便捷和快速的扩大用户好友数量。

圈子是为兴趣相投的用户提供独立互动空间的一种方式，通过加入不同的圈子用户可以参与不同主题的讨论，可以是不同的兴趣小组，也可能是某个培训课程的同学。一个用户可以参与多个圈子，也可以自由退出任何一个已经参与的圈子。用户还可以创建圈子，圈子的所有人是自己，所有人拥有圈子的最高权限，可以邀请其他好友、删除好友、甚至转移圈子的所有权。圈子包含了分类、名称、所属好友唯一编号、共享资料、聊天记录等。

（2）定义用户社交行为的数据采集方式

经过以上的分析，可以社交行为相关数据主要分布在用户档案、好友管理、社交关系管理中。为了便于之后的数据挖据，需要对这三个模块所记录的数据确定数据的采集方式、丰富数据的属性。

用户档案资料里需要采集的信息主要包含了用户名、性别、年龄、学历、学校、专业、职业、工作所在行业。性别和年龄在学习偏好上会有所不同，学历、职业、专业和工作所在行业很大程度上决定者学习的方向。

好友管理资料里主要是用户之间的关系，需要增加的是好友活跃度的信息，所以还需要预先统计和记录用户最近一次登录时间、拥有的好友数、参与的圈子数、提问数量、回答数量都是反应用户活跃度的数据，在给用户推荐感兴趣的好友时，可以去掉那些活跃度低的用户，提高好友推荐的质量。每次好友管理中的数据变化都会触发统计数据的更新，为数据挖据提供及时可靠的数据。

圈子管理资料里需要增加人气指数的统计，和用户活跃度类似，人气指数反应了圈子的质量，人气指数越高说明圈子的质量越好、为用户带来的价值越高。由于圈子关联的深度层级较多，实时的统计会耗费较多的资源，系统采用定时任务进行圈子人气指数统计的方式进行数据更新。

## 4.3 用户偏好分析子系统设计

用户偏好分析子系统是利用用户行为数据采集子系统采集的数据样本，基于Apriori 算法进行数据挖掘，得出用户对教程的学习偏好关联规则、问题与答案的关联规则、用户与用户之间兴趣关联规则，为之后的关联规则应用提供基础模型。下面分别对生成关联规则所需的数据预处理模块、关联规则分析模块、关联规则成果发布模块进行设计。

### 4.3.1 用户行为数据预

1、注意这里的子系统应该与重点内容有关的子系统，分析子系统是为了说明重点部分的内容！与重点内容无关的子系统讨论删除！项目报告与学术论文的区别要搞清楚

2、第三章提到的功能和流程应该是与第四章核心子系统对应，注意主要功能、核心流程与核心子系统的名称不要雷同

根据以上对相关业务的数据分布在各个业务系统，存储在不同据库表都只能从自身的业务角度去记录和关分析，并非所有的数据都有用，所以需所有用的，需要提取哪些属性，如何处理那含许多有噪声、不完整、甚至不一致的数须进行预处理，以改善数据质量、缩小数据范围，最终达到关联分析建模的要求。

以当前海策智能在线培训系统拥有用户档案信息 5 万多条，教程信息 3000 多条，学习历史信息 18 万多条，问题库记录 12 万多条，答案将近 20 万条。如此多的数据如果都要扫描分析，将会耗费大量的资源，并且有些数据对分析过程没有什么意义，有些数据需要先进行处理后进行关联分析会有更高的效率。

首先，数据分散于多个数据表中，不能有效的对数据进行统一处理，需要进行数据集成。再者，不是所有的属性都对本次的分析有作用，所以需要进行数据抽取，选择那些有用的属性。其次，不同的属性可能代表相同的意义但有可能类型不同，需要把数据转换为统一的类型。还有，数据中存在噪音数据，例如缺少值、错误数据、遗漏的数据，这些噪音在数据挖掘中可能引起模型的效果不佳。

根据以上的分析，我们需要对数据进行预处理，具体的数据预处理工作流程是数据集成、数据抽取、数据转换、数据清洗、数据更新。

（1）用户学习轨迹的数据集成

智能推荐需要的数据来源于用户账号信息表（UserAccount）、用户档案信息表（UserBaseInfo）、用户背景表（UserBG）、课程表（Course）、课程分类表（CourseClassify）、课程分类关系表（CourseClassifyRelationship）、用户背景关系表（UserBGRelationship）、学习历史信息（LearningHistory）、教程收藏信息表（CourseFavorites），需要通过用户唯一编号关联把这些表里的数据集成到一起，

代替使用多表关联查询来显示数据，以提高属于的利用率。用户学习轨迹的数据预处理数据如图 4-6。

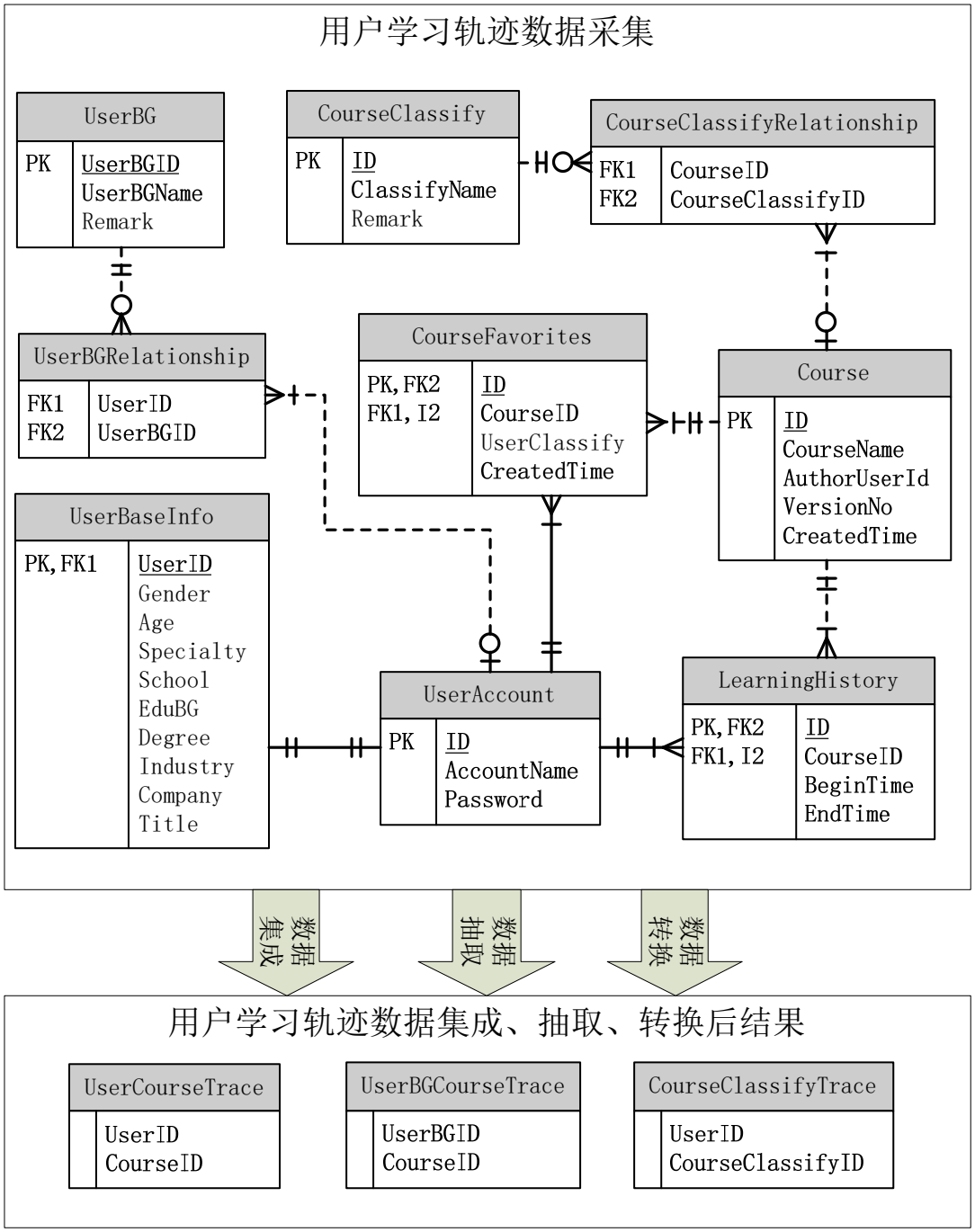



图 4-6 数据预处理数据模型结构图

学习历史信息和教程收藏信息从关联分析的角度都是从某个教程找到关联教程，而不区分他们是因为学习被记录的历史信息还是主动收藏的，学习历史信息的数据是用户的学习行为，一条数据记录代表一次学习经历，而教程收藏信息仅代表了用户对该教程感兴趣，因此这些并不能立刻使用这些数据进行数据建

模，需要做的工作是将这些数据进行集成，合并这两个表的信息得到一个用户教程关系表，用来反映某个用户经历过的教程，即包含了学习过的教程和收藏的教程。

（2）问答匹配的数据集成

智能答疑需要的数据主要来源于问题库、答案库，目前主要有问题表、答案表、问题答案关系表、答案评分表、问题关键字关系表组成。问题表中仅包括了问题的标题，并不能反映出该问题是否有答案，如果有答案是否有评分等，而系统智能答疑是通过新问题的关键字找到拥有同类关键字的问题，且需要找到的历史问题有答案，答案的评分较高，以给问题的答案作为新问题答案的参考，因此集成这些数据让它能反映出问题关键字、问题题目、问题的答案、问题答案的评分，将会减少数据分析的工作量、提高分析的效率，为生成一个高效的分析模型提供了高质量的数据。

（3）智能推荐和智能答复所需属性的抽取

完成了数据的集成后，对数据的观察可以发现集成后的原始数据为了满足业务需要，系统的数据库设计保留了大量的业务属性，但在建立关联分析模型时，并不是所有的属性都是我们需要的，只选取哪些有用的信息可以提高分析的效率和正确率，因此必须将原有数据与关联分析无关的属性进行删除。

智能推荐中主要需要知道是哪位用户收藏和学习了哪些教程，关于用户的身份在用户档案信息中的唯一编号、姓名、登录账号等都可以表示，但我们只需要用户唯一编号进行标示就可以，还有诸如用户密码、密保问题、注册时间等都是本次关联分析不关心的属性，因此将这些属性删除。知道用户之后还需要知道这些用户所经历过的教程，我们只需要知道教程的唯一编号，关于教程的名称、分类、所有人等都没有太大的价值，所以这些属性也要删除。

智能答复中主要需要知道哪些问题有答案且答案的评分高于某个最小值，关于原始数据中问题的标题、答案的内容、提问人、答复人都不是本次建模关心的属性，因此也要删除。数据抽取的过程我们仅仅保留那些可以被关联分析模型使用的最少的属性就可以了。

（4）冲突数据和敏感信息的数据转换

完成了数据的集成、数据抽取后的数据将会被保存在数据库中的新表中，这些表是仅提供给关联分析使用的。但在集成的过程中发现，我们选取的属性中，合并为同一张表时某几个属性的意义是一样的，但是他们的字段类型不同，需要我们统一转换后的字段类型。还有一些属性虽然没有类型冲突，但在业务系统中记录的是用户的隐私信息，或者和商业有关的敏感信息，这些属性本系统都把他们转换为指定的数字类型，不保存他们具体的含义。

以用户学习轨迹采集为例，经过数据集成、数据抽取、数据转换后主要表现为三张表：用户教程跟踪表（UserCourseTrace）、用户背景教程跟踪表（UserBgCourseTrace）、教程分类跟踪表（CourseClassifyTrace），如上图 4-6。用户教程跟踪表是为根据教程 ID 获得相关教程规则提供分析数据，用户背景教程跟踪表是为根据用户背景获得相关教程规则提供分析数据，教程分类跟踪表是为根据教程分类 ID 获得相关教程分类规则提供分析数据。

（5）数据清洗处理方案

完成了以上几步数据的预处理后，数据还存在大量噪声数据、错误数据、缺失数据等问题。产生这些数据的原因可能是系统的漏洞、系统异常后导致的数据缺失，或者是人为的操作失误造成的，称之为“伪样本”。由于“伪样本”的存在会造成分析结果有偏差，所以在数据预处理过程中需要进行数据清洗。

处理缺值记录。由于各种原因导致系统中的部分属性值缺失，属性值缺失的情况经常发生甚至是不可避免的，但缺值的存在对分析结果的正确性有很大影响，无法想象在虚假、劣质数据泛滥的数据集上，如何能找到有用的、隐藏的关联规则。首先，由于缺值可能导致系统丢失了有用信息；再者，因为脏数据能够使挖掘过程陷入混乱，导致不可靠的输出，因此，在进行数据挖掘前，对数据的缺值进行相关的处理是非常有必要的。处理缺值有很多方法，主要有删除元组、数据补齐和不处理三类。对于出现缺值的记录处理采用删除记录的方式，可能会丢失隐藏在数据中的知识点，因此不采用删除空值记录的方法。数据补齐又有手工填补、利用缺省值填补、利用均值填补、利用同类别均值填补、利用最有可能的值填补等方法。本系统主要是用户档案资料有大量的缺值，主要原因是有些信息并非必填项目，用户可以选择空值，因此，大部分都采用缺省值填补。

删除重复记录。由于数据从不同的业务表中集成到一起，很多记录可能会有重复，例如用户学习过一个教程同时也收藏了该教程，在最终集成的信息表里其实用户只经历了一个教程,只需要保留一条数据。还有由于系统缺陷或者异常带来的重复数据。系统经过统计某些字段是否联合唯一来判断是否有重复数据，如果发现只保留其中一条。

处理错误记录。很难避免在系统的数据库中有一些错误的数据,这些样本的数据可能是由录入错误造成的，也可能是系统缺陷。对于错误数据的问题，则如果可以推断最后可能的值则进行自动修正，例如一条已经评分的答案记录有关联的问题唯一编号，但问题的状态错误的表示为等待回答，则可以修正该问题状态为已答复。但如果遇到无法推断可能值的情况，采取删除错误记录的方式，删除后能提高数据挖掘算法的效率和准确度。

（6）预处理任务的执行策略

海策智能在线培训系统的关联分析采用定时任务的方式，在资源不紧张的时段进行计算，每次计算任务开始时都需要对最近一次已经预处理的数据再补充增量数据，数据更新的及时性、有效性将决定分析结果是否正确。

系统把以上的数据预处理过程[illegible]行关联分析模块，就会先调用该[illegible]一步的关联分析。

- 全文务必围绕重点内容（此部分是全文重点）组织，与之无关内容删除，否则项目报告嫌疑
- 重点内容占第四章的绝大部分篇幅

### 4.3.2 用户偏好关联规则分析

根据上述的数据预处理，完成了海策智能在线培训系统的数据准备，可以开始对数据进行分析。本节主要对如何使用 WEKA 实现关联规则的计算进行介绍，获得关联规则并验证。

（1）关联规则分析的环境配置

本文采用 Apriori 算法进行[illegible]关联规则的分析，使用数据挖掘开源软件 WEKA（Waikato Environment for [illegible]软件是用 Java 语言实现的，与本系[illegible]中调用相应的接口进行计算，一[illegible]关联规则，满足最小支持度和最[illegible]

列举分析问题时，列举的方面不要超过 5 点，不允许简单的几个标题（每个标题都应有详细的分析）！

WEKA 软件可以从官方网[illegible]安装后会得到一个图形操作界面、weka.jar 类库文件及 weka-src.jar 源码文件。前期使用 WEKA 的客户端图形界面可以很方便的进行小规模数据的调试和验证，项目投入运行时采用在 JAVA 项目中引入 WEKA 的 JAR 包、调用相应的接口完成关联规则的计算。

首先需要在项目工程中引入 weka.jar，然后配置相应的数据库连接参数，包括数据库连接地址、用户名、密码和数据查询语句，编码时引用相应的类，创建具体的实例进行计算。此外还需要设置最小值支持度、最小置信度等参数，参数的配置可以实现 weka.core.OptionHandler 接口，这个接口为各种数据挖掘方法都提供了设置、获取参数的功能，至此就完成了关联分析计算任务前的基本的配置和程序编码。

（2）关联规则分析数据准备

在上一节中已经完成了数据预处理，数据的属性都是关联分析需要的，数据也采用了增量更新的方式扩展数据样本，但这些数据仍然有优化的空间，还可以进一步缩小数据扫描的范围，提高分析效率。

关联规则分析数据准备首先需要除去那些明显不存在相关性或者相关性很低的数据。例如用户教程跟踪表（UserCourseTrace）中，某个用户仅收藏了一

个教程，也就是一个 USERID 只对应了一条记录，那么他和其他教程就没有明显的相关性，这样的记录就可以删除。那为什么不在数据预处理中删除这些记录呢？数据预处理之后的数据虽然是历史数据，但它仍然会随着时间的推移增加，数据之间的关联关系也会改变。例如上面的例子，随着时间的推移，该 USERID 收藏的教程越来越多，就会从原来的无相关性变为有相关性，因此有些数据的处理是在每次计算之前进行的，删除的数据只是表示在本次分析任务中无相关性，可能以后会产生相关性，所以某些数据优化的步骤是在分析任务执行过程中进行。用户偏好分析子系统关联分析任务数据准备中删除的数据如图 4-7。

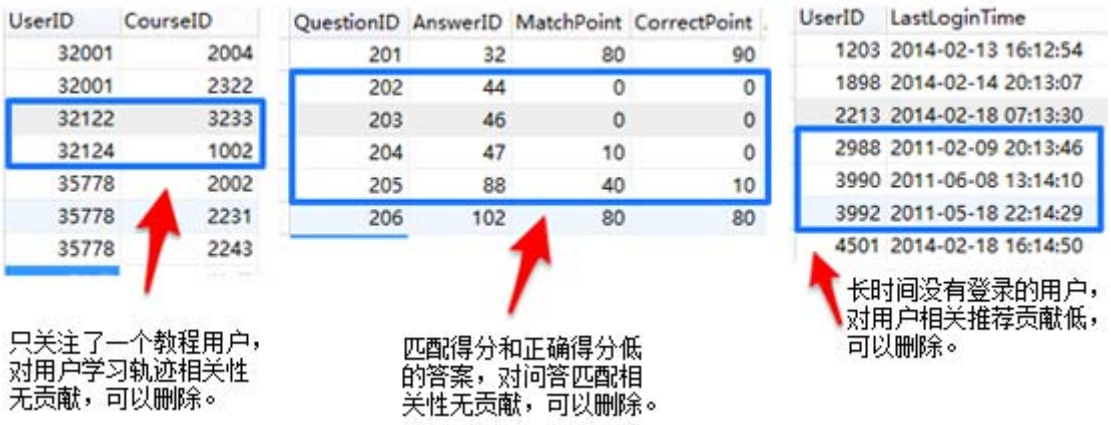



图 4-7 数据准备中删除的数据示意图

删除多余的数据之后，还需要把剩下的数据样本转换为 WEKA 支持的 ARFF 格式的文本数据文件，该文本数据文件由系统的专门模块负责生成，生成完毕后才出发分析模块进行关联规则分析。用户学习轨迹的数据文件如图 4-8。

```
@relation UserCourseTrace

@attribute 'c1' { t}
@attribute 'c2' { t}
@attribute 'c3' { t}
@attribute 'c4' { t}
@attribute 'c5' { t}
@attribute 'c6' { t}
@attribute 'c7' { t}
... ...(省略部分)

@data
?,?,?,?,?,?,?,?,?,?,?,t,t,t,?,t,?,t,?,?,t,?,?,?,t,t,t,t,?,t,?,t,t,?.....(省略部分)
t,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,t,t,t,?,?,?,?,?,t,?,?,?,t,t,?,?.....(省略部分)
?,?,?,?,?,?,?,?,?,?,?,?,?,t,t,?,t,?,t,?,t,?,?,?,?,?,?,?,t,?,t,?,?,?,?,?.....(省略部分)
t,?,?,?,?,?,?,?,?,?,?,?,?,t,t,?,t,?,t,?,?,t,t,?,?,t,?,?,?,?,?,?,?,t,?,?.....(省略部分)
?,?,?,?,?,?,?,?,?,?,?,?,?,t,t,?,t,t,t,?,?,?,?,?,?,?,t,t,t,?,?,?,?,t,?,?.....(省略部分)
?,?,t,?,?,?,t,?,?,?,?,?,?,t,t,?,t,t,t,?,?,t,t,?,?,?,?,?,t,?,?,?,?,t,?,?.....(省略部分)
t,?,t,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?.....(省略部分)
... ...(省略部分)
```

图 4-8 用户学习轨迹的 ARFF 数据文件截图

WEKA 首先导入 ARFF 格式文本数据，然后创建一个数据实例,创建 Apriori 算法实例，设置相关的属性和参数，最后进行计算并返回计算结果，主要实现代码如下。

```
import weka.associations.Apriori;
import weka.core.FastVector;
import weka.core.Instances;
... ...
public class UserCourseTrace {
    ... ...
    public FastVector[] execute() throws Exception {
        //从数据文件中导入数据
        BufferedReader reader = null;
        reader = new BufferedReader(new FileReader(dataFile));
        //创建实例
        Instances data = new Instances(reader);
        reader.close();
        //设置数据属性
        data.setClassIndex(dat
        //创建 Apriori 算法 sche
        Apriori scheme = new A
        //设置 Apriori 算法参数
        scheme.setOptions(weka.core.Utils.splitOptions(
                "weka.associations.Apriori -N 10000" +
                " -T 0 -C 0.9 -D 0.05 -U 1.0 " +
                " -M 0.4 -S -1.0 -c -1"));
        scheme.buildAssociations(data);
        FastVector[] rules = scheme.getAllTheRules();
        //返回规则结果
        return rules;
    }
    ... ...
}
```

1、除主要数据结构、类、本文提出的（改进）算法、消息及其处理程序外，其余代码删除
2、代码不能太简单

除非本文设计或改进的（而非直接应用现有的算法）核心算法，尽量不要在正文放入代码！注意这里给出的样文只是相对好一些，并非没有问题的！

计算任务运行一定时间后将获得 Apriori 的计算结果,计算后获得的关联规则不能直接使用，需要转化为数据库表的记录，在此基础上封装一些接口来查询

和访问这些关联规则。通过 WEKA 计算最后获得的关联规则部分结果如下。

```
Best rules found:
1. c4=t 74 ==> c1=t 74    conf:(1)
2. c3=t c4=t 54 ==> c1=t 54    conf:(1)
3. c4=t c9=t 42 ==> c1=t 42    conf:(1)
4. c1=t c3=t 55 ==> c4=t 54    conf:(0.98)
5. c1=t 77 ==> c4=t 74    conf:(0.96)
6. c1=t c9=t 44 ==> c4=t 42    conf:(0.95)
... ...
```

（3）关联规则分析任务的拆分和部署

系统的数据日益增多，面对如此庞大的数据量该如何高效、有序的进行关联规则分析任务是该子系统设计的重点。首先按照业务目标把分析任务分为不同的子任务，设置他们之间的优先级和依赖关系，由调度模块控制任务的队列和执行。然后根据数据量的预判，把子系统分别部署到不同的服务器，以保证计算任务执行时有充分的资源，并且能够同时执行多个没有依赖关系的子任务，缩短任务完成所需要的时间。

（4）关联规则分析任务异常监控和处理

整个分析过程会耗费一定的系统资源和时间，虽然进行了数据预处理，但还是很难完全保证在计算过程中遇到系统宕机、资源不足等异常，如果监控和处理这些异常也是该子系统设计的重点。关联规则分析任务的异常大致分为系统宕机、停止执行和执行出错三类。

系统宕机的监控是采用监控子系统定时连接关联规则分析任务所在服务器的方式，一旦监控发现故障会发送报警邮件给系统运维人员，进行人工重启，然后在分析相关的日志文件和排查问题。停止执行的监控也是有监控子系统完成，计算模块会定时的输出心跳日志，监控系统获取心跳日志，判断最后一次心跳日志的时间和当前时间的间隔，如果超过一定范围则认为系统已经没有心跳，发送报警邮件给系统运维人员进行排查和处理。执行出错时计算模块会主动抛出异常并且在数据库中记录本次任务批次的状态，监控子系统通过查询计算任务批次的状态，发现有异常的状态则报警，运维人员处理后再将状态改为已处理。

（5）关联规则结果验证

为了保证关联规则的正确性，需要对计算结果做验证性的功能，以保证对外提供的数据查询接口可以正常的运行。根据每个计算子任务的目标，分别编写测试脚本对每个子任务的数据完整性、对外接口运行的正确性进行验证。

### 4.3.3 用户偏好规则的存储和发布

通过以上的关联规则分析，为智能助教子系统的相关教程主动推送、在线提问智能答复、社交关系智能推荐提供了规则模型，这些关联规则数据经过存储和发布，以服务接口的方式提供给其他子系统。

（1）用户偏好规则存储设计

关联规则分析的结果如“{先导}=>{后继},支持度，置信度，提升度”的形式，无法直接提供给其他子系统应用，需要根据分析结果各个数据的含义，对应的设计数据库表结构，转换格式后存入数据库。数据库表中的属性包含唯一编号、先导、后继、支持度、置信度、提深度、对应分析任务批次号等。在用户偏好分析子系统中，设计了专门的模块进行这部分的工作，把分析结果转换为数据矩阵的方式存储到数据库中，再通过封装一定的访问接口从数据库中获取这些数据。

由于数据量较大，数据分析过程需要耗费较长的时间，为避免由于分析过程中遇到性能问题、异常和各种故障导致分析过程中断，影响其他服务的正常运行存储方案设计为两个独立的数据库，分别是主数据库和备用数据库，保证主数据库正在被使用，同时备用数据库参与分析过程的结果存储，每次分析完成都保存到备用数据库中，校验数据通过后，通过数据库链接管理模块，切换备用数据库为主数据库，相互交换主备关系，为下一次计算做准备。当主备数据库在切换进行过程中，关联分析模块为暂停运行状态。

由于分析过程中需要耗费较多的服务器资源，系统把需要分析的任务分为多个子任务，每个子任务的每次分析都用唯一的批次号来区分。所有的任务排队进行分析，任务队列的调度由专门的关联分析任务模块进行管理，使得计算所耗费的时间对系统的影响减到最小。系统管理人员可以随时查看任务的执行情况、耗费的时间，如果分析任务遇到失败或异常，会主动给相关人员发送通知邮件，以便可以及时排查问题和解决故障。

（2）用户偏好规则发布设计

用户偏好规则存储到数据库中后，想获得这些关联规则数据需要调用该子系统的对外服务接口，主要包括的接口如表 4-1。

表 4-1 用户偏好规则发布对外接口表

| 接口名称 | |
|---|---|
| getCourseRuleByUserBackgrc | 则 |
| getCourseRuleByCourseId | 则 |
| getCourseClassifyRuleByCourseClassify | 根据教程分类 ID 获得相关教程分类规则 |

1、 表格中的文字为 5 号（包括标题）
2、 表格居中
3、 全文图表边界不要超过正文边界
4、 全文图表统一编号

| | |
|---|---|
| getAnswerRuleByQuestionClassify | 根据问题分类获得相关答案规则 |
| getFriendRuleByUserBackground | 根据用户背景获得新好友规则 |
| getCircleRuleByUserBackground | 根据用户背景获得新圈子规则 |
| getFriendRuleByFriendNetwork | 根据当前好友社交关系获得新好友规则 |
| getCircleRuleByCircleClassify | 根据圈子分类获得新圈子规则 |

## 4.4 智能助教子系统设计

以上的用户偏好分析子系统为智能助教子系统提供了规则依据，智能助教子系统结合实际的需求为系统提供各种基于这些规则的智能功能，主要功能包括相关教程主动推送、在线提问智能答复、社交关系智能推荐等。下面详细介绍智能助教子系统的设计。智能助教子系统与其他子系统模块之间的关系如图 4-9。

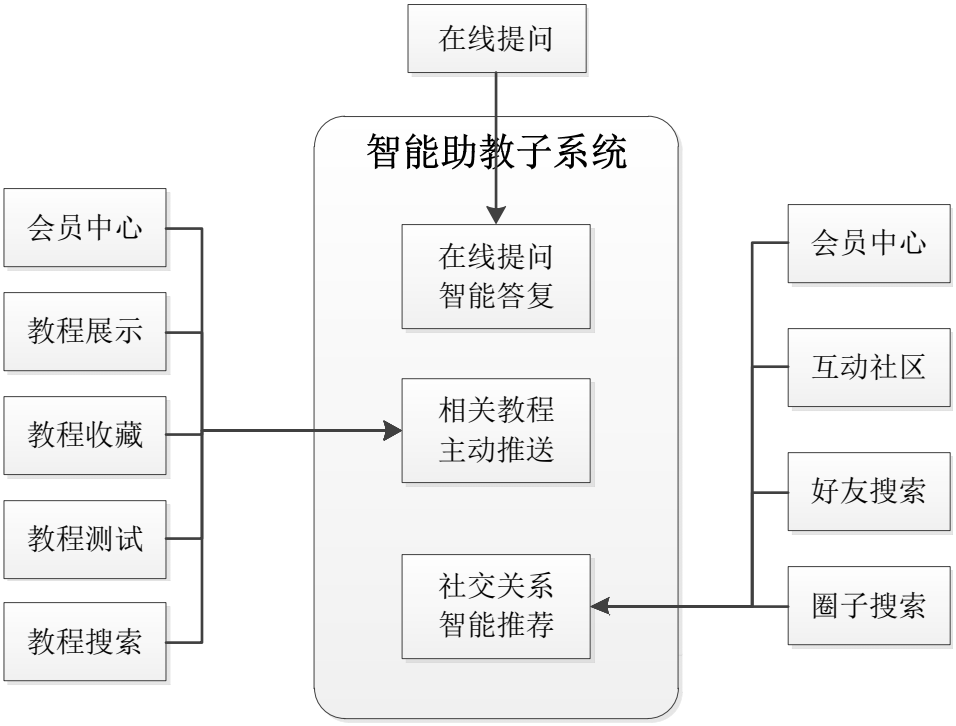



图 4-9 智能助教子系统与其它模块关系图

### 4.4.1 相关教程主动推送

智能助教子系统的相关教程在用户登录到会员中心、教程展示、教程收藏、教程测试、教程搜索几个环节所调用。该模块设计了三个主要接口，分别是根据用户唯一编号获得与用户背景相关的教程、根据教程唯一编号获得与该教程相关

的教程、根据教程唯一编号获得与相关教程分类。主要接口和类的设计如图4-10。

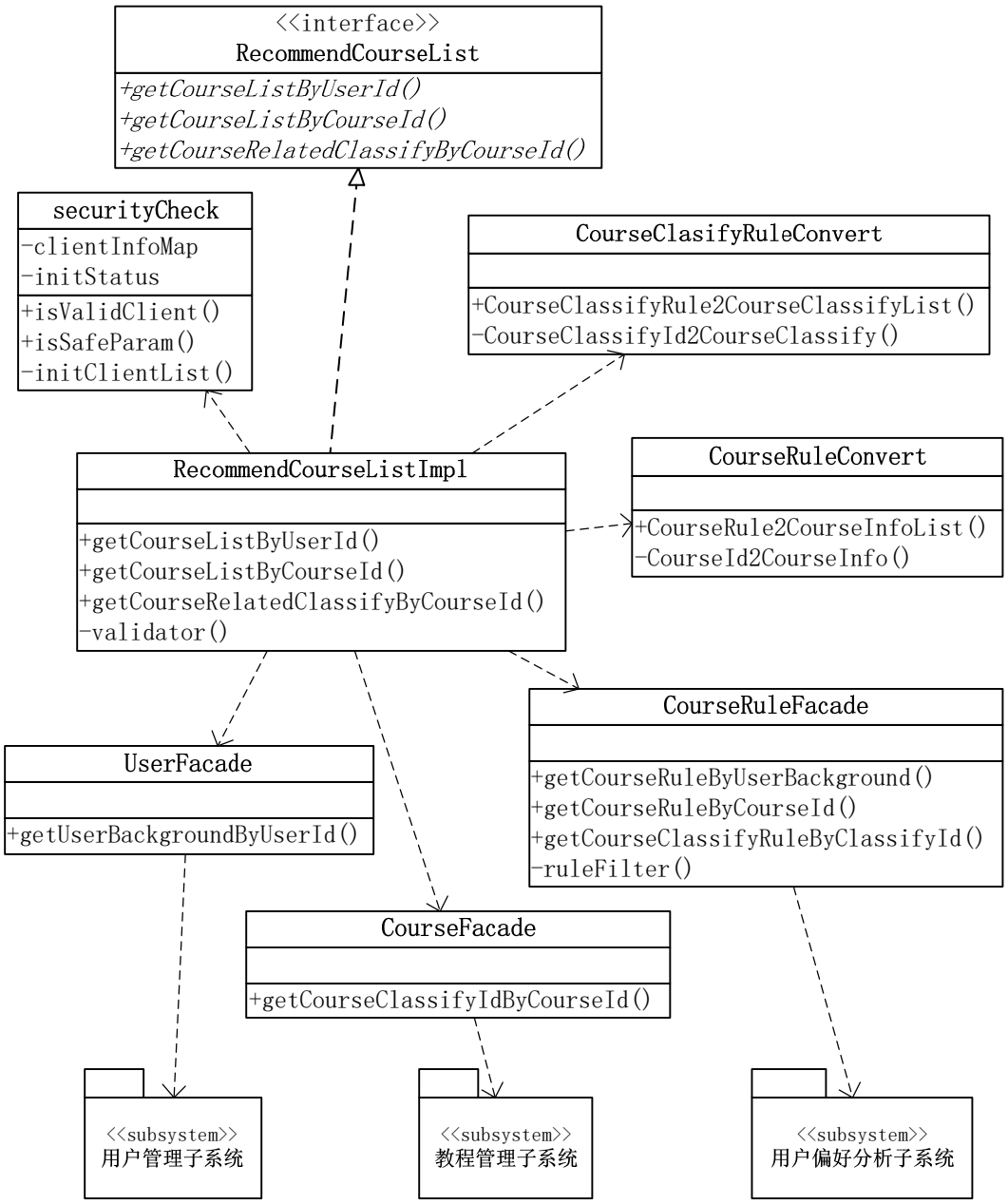



图 4-10 相关教程推荐模块主要类图

（1）推送与用户背景相关的教程

首先在用户登录系统时，系统会调用智能助教子系统的根据用户唯一编号获得与用户背景相关的教程接口，根据用户背景资料与教程的关联规则获得教程列表返回给用户，在这个过程中，系统根据用户唯一编号从用户管理子系统中获得用户背景资料对象，把该对象作为参数调用用户偏好分析子系统获得教程关联规则，关联规则里包含了课程的唯一编号，系统会过滤多余的规则，例如包含用户

已经收藏和学习的教程需要删除，之后通过教程管理子系统查询教程的详细信息，例如名称、介绍、发布时间等，然后返回课程列表数据。例如一个用户的专业是计算机，职业是前端软件工程师，工作 3 年，用户登录后，系统经过分析很可能会给他推荐其他的 HTML5 教程和 JavaScript 等一些前端技术的教程。

（2）推送与教程相关的教程

在用户收藏 A 教程、学习完毕 A 教程、完成 A 教程测试时，系统希望会推荐收藏和学习了 A 教程的用户也关心的那些教程。系统根据 A 教程的唯一编号获得与 A 教程相关的教程列表，在这个过程中，系统把 A 教程唯一编号作为参数调用用户偏好分析子系统获得教程关联规则，这些规则里包含了关心 A 教程的用户也关心的所有教程，关心是指曾经收藏和学习了教程，系统过滤了多余的规则之后通过教程管理子系统查询教程的详细信息，返回课程列表数据。例如在关联规则中保存了大多数人学习《Java 编程技术》教程的用户都学习了《Tomcat 的安装和配置》，当用户收藏《Java 编程技术》时就会推荐《Tomcat 的安装和配置》等教程，当然推荐列表中不包含用户已经收藏和学习过的教程。

（3）推送与教程相关的教程分类

在用户搜索关键字时，系统认为用户在寻找某一类教程，这在搜索列表中就可以看到，此时再推荐相关教程会和搜索结果雷同，推荐价值不高，所以把该部分的推荐设计为推荐更多相关的教程分类。在用户正在观看教程中，由于不想影响用户的观看焦点，在排版上推荐的位置较小，此时推荐相关教程会显得数量有限，达不到推送的目的，所以把该部分的推荐设计为推荐相关教程分类。此时系统调用接口查询相关分类，首先通过当前教程唯一编号通过教程管理子系统获得教程的当前分类，使用该分类作为参数查询相关分类并返回结果。例如用户在学习《Spring 入门》的同时，会显示可能感兴趣的教程分类是 Java 基础、服务器配置、网站设计、Eclipse 等教程分类，并显示该分类下面的教程总数。

### 4.4.2 在线提问智能答复

智能助教子系统的在线提问智能答复在用户提交问题时触发，系统通过已有的问题与答案的关联规则，返回该问题的可能匹配的答案提供给用户。主要接口和类的设计与相关教程主动推送类似。

（1）问题分析引擎的设计

由于问题是用户通过文字输入提交文本信息给系统的，类似的问题可能会有不同的文字描述，只有把问题的文字描述归纳总结最终确定问题的类型，以缩小问题查找的范围，才能比较准确的锁定需要匹配的答案。

问题分析引擎就是为满足以上的需求设计的。问题分析引擎主要的功能是对

问题的语意进行分析，把不同描述的问题确定为同一类问题，这一类问题有相同或非常类似的答案。首先，系统的问题分析引擎会按照一定规则对问题进行关键字拆分，去掉干扰的信息，使用主要的关键字组合后进一步得出问题的分类。

问题分析引擎并不是实时的进行数据库扫描进行分析的，因为这样可能会带来很大的性能问题，而是通过定时任务，定期的对增量的问题进行语意分析，获得新的规则和新的分类，也就意味着最近提出的问题不会立即影响到问题分析引擎，需要一定的同步时间，这样的延时是在系统的允许范围内的。

（2）智能答复的答案评分设计

系统给出问题的匹配答案列表后，会按照历史的匹配度进行降序排序后显示给用户，用户可以通过翻页查看所有的系统认为匹配的答案，如果用户认为没有找到答案，可以等待教程发布人和其他用户的人工答复，如果用户在答案列表中找到了最佳答案，那么需要对答案进行评分，否则该问题总是显示为未解决。用户对答案的正确性给予评价反来过不停的优化和改进智能答复的准确性，所以评分的设计是否合理直接影响智能答复的效果。

评分项包含正确度、匹配度、权重三部分，其中总分和匹配度由用户给出，权重是系统动态计算的，每个问题用户可以给三个备选答案进行评分。其中，正确度分为三档：非常满意、满意、一般；匹配度分为五档以星级的方式选择，最匹配为五星，至少为一星；权重是系统根据教程分享人答复、用户答复、智能答复这三种类型，结合当前已经答复的次数进行计算的，教程分享人的权重大于用户答复的权重，用户答复的权重大于智能答复的权重，问题刚开始有匹配答案时权重偏高，随着匹配答案越来越多用户评分对答案的权重也会有所降低。

（3）智能答复的应用结果设计

智能答复的的操作是否简单、评分是否便于操作、结果显示是否清晰对该功能的效果起到决定性作用，因此对智能答复的交互设计至关重要。

首先，提问的便捷性体现在用户可以在观看教程的时候快速提问。页面采用AJAX 技术防止页面刷新，采用伸缩图层的方式使教程内容和问答内容同时显示，采用增量载入的方式提高响应速度。再者，系统的评分不采用数字，避免了用户由于标准差异导致的评分偏高或者偏低，采用单选较少的选项来减少操作的难度，采用使用鼠标完成评分来代替键盘更好的兼容了个人电脑和移动便携设备两种不同终端。

系统以智能答复用户提问的方式代替了过去等待人工答复的方式，使得用户很有可能就能立即获得答案而继续学习，增强了学习的连续性。同时也减少了人工答复的工作量，特别是系统官方发布的教程，答复即时和工作量的减少，降低了运营成本、改善了用户体验。智能答复在系统中的应用结果设计如图 4-11。


图 4-11 在线提问智能答复应用结果截图

### 4.4.3 社交关系智能推荐

智能助教子系统的社交关系智能推荐是通过分析用户参与社交关系的特征，结合用户档案资料，学习轨迹，找到与用户兴趣特征相关联的好友、圈子，从中挑选出人气指数高、活跃度高的好友和圈子推荐给用户。当用户在搜索好友、添加好友、搜索圈子、加入圈子、新建圈子时会触发社交关系智能推荐。接口和类的设计与相关教程主动推送类似，主要包括接口：根据用户背景推荐新好友、根据用户背景推荐新圈子、根据当前好友的社交关系推荐新好友、根据圈子分类推荐新圈子。

（1）根据用户背景推荐新好友

首先在用户登录系统时，系统会调用智能助教子系统的根据用户背景推荐新好友接口，根据用户背景资料与其他用户的关联规则获得同背景、相关背景的所有用户列表返回给用户，在这个过程中，系统根据用户唯一编号从用户管理子系统中获得用户背景资料对象，把该对象作为参数调用用户偏好分析子系统获得关联规则，再通过查询这些有关联的背景下的所有用户基本信息作为返回结果。例如经过分享，登录用户的背景是属于分类 A，那么会返回分类 A 的用户的当前好友列表里关联度最高的一些用户背景分类 B、C、D，再通过这些分类找出分类下

的所有用户列表。

（2）根据用户背景推荐新圈子

用户登录系统时，系统还会调用根据用户背景推荐新圈子接口。与根据用户背景推荐新好友类似，该接口根据用户背景资料找出同背景的用户都加了什么圈子，之后还会根据圈子的人气指数高、活跃度高排序后推荐给用户。圈子的人气指数主要是指加入的人数多，活跃度指该圈子的交流平凡、圈子的成员登录平凡，因为人数少或成员大部分都不常常登录的圈子对用户的意义不大，所以需要这两项指标进行排序，使用户能找到有用的圈子。

（3）根据当前好友的社交关系推荐新好友

用户在搜索新好友和添加新好友时，系统会调用根据当前好友的社交关系推荐新好友接口。该接口根据用户偏好分析子系统的关联规则，获得已有好友大都添加了哪些还不在用户好友列表中的用户，作为新好友列表返回给用户。系统认为如果你有多个好友同时拥有一个用户作为好友，那么很有可能你对该用户也认识或者感兴趣。系统同样会去除哪些长期不登录的用户，保留那些活跃度较高的用户进行推荐。

（4）根据圈子分类推荐新圈子

用户在搜索新圈子、新建圈子、加入圈子时，系统先经过查询获得该圈子的分类，然后调用根据圈子分类推荐新圈子接口，根据用户偏好分析子系统的关联规则，获得该圈子相关的圈子信息列表返回给用户。这里不做同分类圈子的推荐，因为同一分类的圈子往往只有少数具有较高的活跃度，大部分用户都会添加那几个活跃度较高的圈子，所以同一类型的圈子推荐意义不大。有相关度的圈子也是按照人气指数和活跃度排序后推荐给用户的。

## 4.5 与同类系统比较

与其他在线培训系统相比，海策智能在线培训系统保留了传统培训系统的核心功能，又结合了目前社区网站、视频网站、协同工作系统的优点，该系统在设计和功能开发上具有其明显的自身特点。

（1）采用数据挖掘技术实现了个性化内容的优势

与其他系统依靠用户搜索、分类查找教程不同，该系统使用了关联分析技术，在教程学习核心功能之外，利用用户的资料数据、学习数据进行数据挖据，为用户推荐感兴趣的相关教程，体现了系统提供个性化内容的优势。

（2）多人协同制作教程比传统培训网站具有更好的协作机制

该系统还借鉴了协同工作系统的原理，除了支持个人账号发布教程之外，还设计了组织，即多人一起组成一个团队，进行教程的制作，为复杂的教程制作、

多人分工、多人并行维护提供了可能。多人协作的设计包含了版本控制、迁入迁出、冲突解决等复杂的功能，同时也提供了很多便捷的工具管理团队，例如邀请成员加入、设置团队所有

1、本章最后应补充同类系统的比较（架构、核心技术应用、集成方法或者其他方面与同类系统的差别，突出“创新”，否则难以高分或通过
2、本章最后补充系统的应用效果（不仅和同类系统有不一样的地方，而且效果也较好）

（3）自动答复问

系统通过对用户提

复，在人机交互的智能性上走在了同类系

得反馈，学习过程不易被问题打断，增强

（4）采用了面向服

系统采用了面向服务的架构，多个子系统之间已服务的方式提供给对方调用，系统是一组松耦合的服务组成，每一个服务的建立和替换都是相对简单的，增强了扩展性，能更快速、更有效地适应业务需求的变化

## 4.6 海策智能在线培训系统应用效果

系统基于关联分析的智能模块设计和实现之后，让系统的智能性得到了很大的提高，使系统从传统的互动点播模式转向了智能互动模式，特别是以下几个应用效果是本次的主要成果。

（1）智能推荐相关教程增加了系统流量和延长了用户停留时间

系统的相关教程推荐在系统首页、用户中心、教程搜索、教程收藏、教学学习的栏目中都得到了应用。从系统运行的效果来看，用户收藏的教程数量明显增加，连续学习的比重也有明显增多的趋势，系统的流量因此有了一段连续性的增长，用户登录系统后的停留平均时间增加。

（2）智能答复问题改善了用户体验和增强了学习的连续性

系统的智能答复是在用户提交问题后触发的，该功能的上线使得用户学习的连续性得到了保证，用户体验也得到了提醒，虽然系统中还有很多提问是重复的，但是教程分享人进行人工回答的比率在下降，用户对该功能给予了好评，同时也降低了教程的运营成本。同时，用户之间的问题的答复由于有答案评分机制，使得答案的匹配选择更广，并且问题的正确答案有了一个自由、良性的扩展方式。

（3）社交关系的智能推荐使得社交网络更壮大而且活跃度更高

社交关系的智能推荐让用户很容易就能找到自己兴趣相投的同学、圈子，使得用户参与的社交关系越来越多，社交网络逐渐壮大，用户和圈子的活跃度有了明显的提高，系统也因此增了流量、提高了系统对用户的粘性，使得之后的多人协作开发教程有了更多的成果。

# 第五章 结 论

本文在针对海策智能在线培训系统中出现的一些不足和问题，例如不能很好地让用户找到感兴趣的教程、问题答疑回复速度慢且问题大量重复、系统的内容更新和优化渠道单一且没有发挥系统的群体智能优势。针对这些问题在系统的智能性设计上进行了新的尝试，采用基于关联分析将数据挖掘技术运用于系统的智能功能模块的设计，从而提高了系统的智能度，增强系统的产品竞争力。

## 5.1 海策智能在线培训系统的特点

1、结论部分可以简要总结全文，把本文做的工作与同类工作比较，以说明本文工作的价值。
2、两个 2 级标题

海策智能在线培训系统在系统设计和功能

（1）主动的推荐用户感兴趣的教程

系统除了提供传统的目录检索、全文搜索

荐功能，在用户收藏教程、查看教程的同时，会用

趣的教程，该智能功能的增加，使得用户停留在网站的时间更长，用户粘性增加，用户学习的连续性得到增强。

（2）自动答复用户的问题

以一般的答疑方式不同，海策智能在线培训系统采用了问题自动答复的功能代替了大部分的人工答复，用户在提问之后系统就自动给出相关可能正确的答案，只有在该答案不满意的时候，才需要人工答复。往往有过正确的人工答复的问题，一般都能够让用户的问题命中对应的答案。该设计使得系统的大部分问题答复都做到了实时响应，提高了用户体验，减少了人工答复的工作量。

（3）社交关系智能推荐

海策智能在线培训系统与传统的依靠论坛作为社交网络的同类系统不同，吸取了及时通讯软件、论坛、协同办公软件的有点，开发了支持好友、圈子、聊天、协同制作教程等培训系统特有的社区模式，并智能分析用户社交行为，为用户只能推荐与用户兴趣匹配的圈子，增加了圈子的作用，形成了良好的用户社交网络，增加了用户对系统的粘性和为发挥用户群体智能提供了基础。

（4）多人协作的教程发布

很多在线培训系统都支持个人分享和发布教程，但海策智能在线培训系统是目前为数不多的支持多人共同协作制作和发布教程的系统，这得益于互动社区的良好设计和用户社交网络的形成。该设计体现了系统的群体智能，通过这样的方式，使得系统的内容有了一个良性的增长和优化的渠道。

## 5.2 不足与展望

1、注意在总结全文后，还要给出不足以及将来需要继续做的工作。
2、本章的内容可以简单一些，篇幅限制在2页。

在本系统的开发过程中，对系统做出了创新[illegible]技术来提高产品的核心进整理，但仍有许多问题有待

（1）功能还不够完善

本系统改进后的功能还不够完整，目前基本满足了业务需要，用户交互还没有经过专门的优化，还需要交互设计师进行改进，功能细节处理得还不够理想，需要根据更多的用户反馈优化和改进系统。系统的移动版本也仅限于一些基础的功能，大部分都采用嵌入网页的方式实现，对于各种分辨率的手机支持得还不够好，手机的访问速度需要再优化。

（2）系统稳定性不够高

系统目前的访问量还不算高，对于大量的并发访问还没有进行优化，下一步需要对访问量较大的几个子系统增加缓存机制、负载均衡等。对于数据库的备份目前还是采取每天一次的冷备份，需要改进为热备，防止系统故障导致数据丢失。监控系统的覆盖面还不够高，需要增加对系统的各个重要节点进行监控，主动发现系统问题，对系统日志、系统负载进行分析，找到系统运行时的瓶颈，纳入下一次的改进需求中。

# 参考文献

1、所有参考文献要在正文引用处按引用顺序标注；
2、参考文献数量在 15—25 篇之间为宜。且英文文献不要少于 1/3，注意不能仅仅引用网站、书籍资料，注意学术类（学报、会议等）的资料不少于 1/3。
3、尽量引用近 3-5 年的资料。
4、格式严格按照参考文献标准编排！

[1] 李世杰.IT 巨头争抢在线培训地盘满足个性化需求是根本[N].通信信息报，2013-08-21(B03).
[2] 韦夏怡.千亿在线教育市场“战国 考报，2013-07-26(003).
[3] 冉兆春.英国在线教育发展特色及启 3(22)：106-107.
[4] 李立勋.启用“淘宝同学”淘宝将 商报，2013-07-08(D01).
[5] 钱玲，张小叶.美国高校在线教育面临 育研究，2011(02)：65-69.
[6] 张蕊.在线教育,高等教育界的沃尔玛[J -42.
[7] 李雁争.在线教育将成互联网下 券报，2013-08-19(003).
[8] 潘雪峰，张宇晴，毛敏，崔鹤.在线教育产业发展现状及产品设计研究[J].科技和产业，2013(08)：13-16.
[9] 赵卫东.商务智能[M].北京：清华大学出版社，2011：135-140.
[10] 陈志泊.数据仓库与数据挖掘[M].北京：清华大学出版社，2009：212-220.
[11] Philipp K. Janert.Data Analysis with Open Source Tools[M].USA：O'Reilly Media，2010：112-118.
[12] Ian H. Witten.Data Mining:Practical Machine Learning Tools and Techniques[M].Burlington：Morgan Kaufmann，2011：523-530.
[13] Jiawei Han.Data Mining:Concepts and Techniques[M].Burlington：Morgan Kaufmann，2011：233-240.
[14] Yaser S.Abu-Mostafa.Learning From Data[M].USA：AMLBook，2012：67-69.
[15] Giovanni Seni.Ensemble Methods in Data Mining[M].California：Morgan and Claypool Publishers，2010：34-37.
[16] Xindong Wu，Vipin Kumar.The Top Ten Algorithms in Data Mining[M].USA：Chapman and Hall/CRC，2009：101-103.
[17] Haralambos Marmanis，Dmitry Babenko. Algorithms of the Intelligent Web [M].USA：Manning Publications，2009：141-143.
[18] 赵洪英，蔡乐才，李先杰.关联规则挖掘的 Apriori 算法综述[J].四川理工

学院学报，2011(01)：66-70.

[19] Bošt jan Kaluža. Instant Weka How-to[M]. Birmingham：Packt Publishing，2013：20-23.

[20] 王彦增，曹正.基于 WEKA 数据挖掘中关联规则的分析及应用举例[J].经济论坛，2013(01)：165-167.

[21] 李强，周贤娟，韩树人.基于 Weka 的数据挖掘技术在学生管理中的应用[J].科技广场，2011(01)：171-173.

# 致 谢

本次论文写作是在我的导师专业、耐心的指导下帮助我完成的，在此，我向导师表示衷心的感谢，感谢导师从开题报告、资料收集、论文初稿、格式调整到最终成文一直给我的帮助和支持，他严谨的治学态度、专业的知识、强烈的责任心给我留下了很深的印象，真诚的感谢他的大力支持和悉心教诲。

在本次论文还得到了软件学院的老师、同学的支持，感谢学院老师课堂上的精心讲解为我的论文打下了坚实的基础，感谢一起共同努力的同学给我的帮助让我开心而又难忘的度过了硕士研究生的学习生活。

最后感谢我的家人给我的支持以及背后的默默奉献。

# 论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名：_____ 日期：_____

# 论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____