

基于深度神经网络的信息抽取系统教学研究

万 昊

山东科技大学, 中国·山东 青岛 266590

【摘 要】教学信息具有海量、系统特征, 可为师生开展多元化教学模式提供可能。由此产生信息抓取困难问题, 也对信息处理模式提出更高要求。本文基于 Python 语言、C/S 架构、MySQL 数据库, 结合关键节点优化算法构建基于深度神经网络的信息抽取系统, 增强信息抽取质量与效率。

【关键词】深度神经网络; 信息抽取系统; 教学信息共享; 节点; HTML 结构

Teaching Research of Information Extraction System Based on Deep Neural Network

Wan Hao

Shandong University of Science and Technology, Qingdao, Shandong, 266590

[Abstract] Teaching information is massive and systematic, which makes it possible for teachers and students to carry out diversified teaching modes. This leads to the difficulty of information capture, and also puts forward higher requirements for information processing mode. Based on Python language, C/S architecture, MySQL database and key node optimization algorithm, this paper constructs an information extraction system based on deep neural network to enhance the quality and efficiency of information extraction.

[Key words] Deep neural network; Information extraction system; Teaching information sharing; Nodes; HTML structure

【基金课题】 山东科技大学人才引进科研启动基金项目《同行评议与计量评价在科研评价中应用比较研究》(2017RCJJ094)。

对于教育领域来说, 信息技术升级为其提供更多发展的可能性, 具体表现为线上学习信息可为学校开展教育提供更多元、专业的教学模式^[1]。线上教学可有效解决时间与空间的束缚问题, 而速度越来越快的互联网技术可为线上教学提供有效保障。教学信息是学校开展教育的关键资源, 逐渐成为学校重点收录成果之一^[2]。因教学信息可共享、可自由存取, 教学信息管理遵循 DC 标准与 OAIPMH 协议, 以此提高信息存储规范性与质量, 并便于用户搜索。同时, 信息抽取主要是从非结构化文本中提取结构信息, 有利于学校提高现有教学资源利用率, 破解信息孤岛问题^[3]。学生可借助信息抽取系统及时、准确获取相应教学资源。教师可通过该系统及时向学生共享教学信息, 有效扩展教学课堂。利用深度神经网络可优化信息抽取系统中存在的某些问题, 增强信息抽取效率。本文设计并应用具备信息对比、查询、更新、导出等多项功能的基于深度神经网络的信息抽取系统。

1 信息抽取系统教学的研究背景

为提高教学效果、提高优秀教学资源利用率, 教育部发布文件建设精品课程^[4]。经过长期发展, 各学校已经建立起专属于自身的精品课程体系, 甚至构建线上精品课程平台。然而各学校之间的平台接口不开放、信息存储分散等问题削弱了教学信息共享性, 制约学生获取感兴趣课程的资料, 降低有效教学资源利用率。通常情况下, 精品课程平台以导航式搜索模式查询课程信息, 主要由平台管理者建设每个细节^[5]。尽管此模式的运营费用较低, 但容易引发教学信息格式不统一、存储分散、共享难度大等问题^[6]。随后, 教育部颁布计划要求各校开放精品课程教学资源, 并组建多门类精品课程教学资源共享

平台。由此, 各校纷纷关注如何提高教学价值、提升教学资源利用率、增强教学效果等热点话题。从现有精品课程教学平台的构建模式开看, 教学网页通常含有 2 种信息, 一是课程简介信息, 二是课程链接。师生根据自身需求搜索课程时需整合各学校、各学院、各教研室的教学资料库, 以此获取相关课程信息, 这对信息抽取要求极高。同时, 各组织内部平台的设定并不相同, 信息格式与平台接口均不相同, 由此增加信息抽取难度。在没有信息抽取系统的情况下, 师生仅能依靠人工逐个搜索与存储课程信息。而人工抽取信息极易出错且效率不高, 难以为师生开展教学活动提供良好的课程信息基础。相较于人工抽取信息, 信息抽取系统可有效降低信息提取错误率、提高收集效率与信息质量。基于上述关于信息抽取的问题, 本文基于深度神经网络构建信息抽取系统, 增强信息收集、对比、处理与应用等功能, 以期提高学校课程资源利用率与教学信息共享性。

2 结合神经网络的信息抽取系统设计需求分析

在提取或共享课程信息时, 结合神经网络的信息抽取系统需确保信息抽取质量。课程信息一般由学校教师队伍自主建立, 不同于专利、学术论文等资料, 无法仅由一人全权负责信息存储与维护。但若交由第三方机构运营与维护信息抽取系统, 可能因工作量大而掺入脏数据, 降低数据质量。结合神经网络的信息抽取系统可通过信息技术替代部分人工处理工作, 有助于提高信息提取速度、降低信息偏差。为增强信息抽取质量, 结合神经网络的信息抽取系统遵循 DC 元数据标准与 OAIPMH 协议规范获取目标数据。

结合神经网络的信息抽取系统导出检索结果对于信息开放及

质量极具意义。系统导出功能可允许用户将检索结果以任一格式存储,提高信息抽取结果的重复使用性。信息搜索功能可为用户提供信息抽取情况的反馈渠道,提高信息抽取过程的透明度,以便及时更正错误,增强信息抽取质量。

由于教学信息将随课程建设不断丰富,信息抽取目标随之改变。结合神经网络的信息抽取系统需时刻把握信息来源,具备一定的信息检查、更新、删减重复信息等功能,提高抽取信息的完整性与及时性。同时,系统需向用户展示信息抽取更新与更正过程,以便用户查询历史信息、查看信息抽取资料的演变过程。

由此,根据结合神经网络的信息抽取系统需解决的实际问题出发,归纳系统功能。第一,结合神经网络的信息抽取系统需向用户展示信息的直接关系,以便用户对抽取结果进一步做对比、分类等处理。第二,通过信息抽取技术扩大搜索范围,从多种课程进展中获取信息,提高信息抽取质量与效率。第三,设置信息抽取历史查询、更新提醒、操作记录等功能。

3 基于深度神经网络的信息抽取系统教学的设计与实现

3.1 系统架构设计

基于深度神经网络的信息抽取系统构建涉及 Python 语言、C/S 架构、MySQL 数据库,共包含抽取与处理模块、搜索与导出模块、更新与核验模块和用户管理模块。其一,抽取与处理模块是基于深度神经网络的信息抽取系统的关键,也是导出抽取结果与更新信息的前提条件,可有效收集信息。用户可根据自身需求提取目标信息,再设置特定条件处理信息,以便快速获取相关教学信息。其二,搜索与导出模块可为用户提供信息查询与导出的功能。基于深度神经网络的信息抽取系统可根据用户输入的关键字或筛选条件搜索信息,并在界面展示搜索结果,支持用户将结果转为任一文件形式导出。其三,更新与核验模块是对现有信息进行补充,并对新增信息的冗余部分进行删减,使用户直观看到新增信息。在更新抽取信息时,基于深度神经网络的信息抽取系统仅需重点修改新增项,无需对历史信息进行重复工作,提高信息抽取结果的同步质量。其四,用户管理模块是为方便用户管理与使用系统信息,使用户享有相应的用户名、密码设置权限,修改界面显示设置、管理信息抽取队列、新增或删减信息抽取历史记录。

3.2 关键技术

信息抽取处理技术是基于深度神经网络的信息抽取系统的主要技术,可实现教学信息的有效收集。因教学网站页面设置较为固定,信息抽取可通过系统内多个神经网络元胞层提高辨别速度,去除广告控件与噪声。基于深度神经网络的信息抽取系统主要是借助教学网站 HTML 文档与 URL 下载渠道获取信息抽取规则,并通过信息抽取器筛选与关键词或模糊条件相符的信息。未识别待抽取信息,基于深度神经网络的信息抽取系统需建立对特殊内容识别的分割规则。信息抽取模板可依据教学网站 HTML 结构解析结果获得 DOM 模型。DOM 模型是万维网联盟提出的网页信息展示模型,可把 HTML 文档转化成含有各标签节点的树形结构,与深度神经网络相似。目标教学网页中 DOM 树节点由各标签组成,而标签间的关系由节点连接方式表示。

信息抽取模板是在解析教学页面内容后生成信息抽取规则,

形成特有信息抽取生成器。从基于深度神经网络的 DOM 树节点来看,信息待抽取部分是一个节点,对应着一条来自根节点的连接路径。而每个节点的具体位置均可转换为 XPath 表达式,由此组建信息抽取模板。然而, XPath 表达式会随着节点距离根节点的分支增加而变得复杂,会降低信息搜索与抽取效率。为化解此问题,基于深度神经网络的信息抽取系统引入关键节点简化 XPath 表达式,使每个节点标签具有唯一性。此特征可去除对节点的父节点查询过程,直接跳跃到节点位置,抽取相关教学信息。信息抽取模板构建需先依据教学网页 HTML 结构形成 DOM 树,再获取节点位置搜索路径与相应 XPath 表达式,随后从上到下搜索节点集合中的同名节点,将唯一存在的目标节点抽取到指定地方。将抽取出的节点组成列表,并由最后一个节点为关键节点再次重复筛选唯一节点,直到列表中无节点。结合关键节点列表与最后节点的子树 XPath 表达式,组成信息抽取模板。例如,构建教学课程简介页面的信息抽取模板,找到待抽取信息途径的各个节点,并确定具有唯一性的节点,将其添加到关键节点列表。在关键节点列表中再次将具有唯一性的节点加入新关键节点列表,并结合 XPath 表达式,从而得到专属于该教学页面的信息抽取模板。并借助此模板构建方式依次组成教学网页中不同教学课程的简介信息抽取模板。各信息抽取模板均放置到抽取与处理模块,并通过 Python 编写待处理信息的抽取与处理路径。

4 结语

为增强教学信息共享效果,本文设计并应用基于深度神经网络的信息抽取系统。该系统具备信息对比、查询、更新、导出等多项功能,方便用户及时获取教学网页信息,定期获取最新教学信息。基于深度神经网络的信息抽取系统借助 HTML 结构、关键节点优化后的 XPath 表达式构建信息抽取模板,加快信息抽取系统处理速度,增强系统性能与实践效果。然而,此次设计并未针对教学网页代码编写不规范的情况进行针对性处理,故下一步研究可就此进行扩展研究,以便提高系统抽取信息的精准率与实用性。

参考文献:

- [1] 宇云飞,李业丽,孙华艳.基于深度神经网络的个性化推荐系统研究[J].电子技术应用,2019,45(01):14-18+22.
- [2] 赖娟,洪艳伟.基于规则约束的深度学习网络用于文本信息抽取[J].计算机工程与设计,2021,42(12):3548-3554.
- [3] 王东波,高瑞卿,苏新宁,朱丹浩.面向情报学课程设置的数据科学技能素养自动抽取及分析研究[J].情报理论与实践,2018,41(12):61-66.
- [4] 陈洪华.基于教育信息化的大学课程实训平台系统设计[J].现代电子技术,2021,44(10):52-56.
- [5] 张敏军,华庆一,贾伟,陈锐,姬翔.基于深度神经网络的个性化推荐系统研究[J].西南大学学报(自然科学版),2019,41(11):104-109.
- [6] 胡悦,罗小楠,王彬,张伟.基于深度神经网络的继续教育信息智能推送技术研究[J].电子设计工程,2021,29(14):42-46.

作者简介:

万昊(1989.5-)男,山东济宁市人,中国科学院大学博士,山东科技大学馆员,研究方向:科技评估与情报分析。