KECVS: 一个面向专业文献知识实体的类型标注及可视化系统

温　雯[1]　伍思杰[1]　蔡瑞初[1]　郝志峰[1,2]　王丽娟[1]

[1](广东工业大学计算机学院　广东 广州 510006)

[2](佛山科学技术学院数学与大数据学院　广东 佛山 528000)

摘　要　知识实体的类型标注是专业文献结构化管理和知识脉络挖掘中的一个重要任务。由于专业文献的知识实体具有专业性强、类型多样、随时间变化的特点,如何在无监督的情况下对其进行类型标签抽取、实体类型标注及知识关系挖掘具有重要的意义。设计并实现了一个面向专业文献知识实体的类型标注及可视化系统,提供文献数据的实体识别、实体类型标注、知识实体关系图构建及其可视化等功能,帮助科研工作者更加便捷、直观、准确地把握知识关系和研究热点。

关键词　信息抽取　实体类型标注　知识关系挖掘　数据可视化　专业文献

中图分类号　TP391　　**文献标识码**　A　　DOI: 10.3969/j.issn.1000-386x.2017.08.016

KECVS: A KNOWLEDGE-ENTITY CATEGORIZATION AND VISUALIZATION SYSTEM FOR ACADEMIC LITERATURE

Wen Wen[1]　Wu Sijie[1]　Cai Ruichu[1]　Hao Zhifeng[1,2]　Wang Lijuan[1]

[1](*School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, Guangdong, China*)

[2](*College of Mathematics and Big Data, Foshan University, Foshan 528000, Guangdong, China*)

Abstract　Knowledge-entity categorization is an important task for the structural management of academic literature and knowledge-graph mining. Since knowledge entities are highly specialized, diverse and evolve with time, how to categorize, annotate and analyze the knowledge-entity on unlabeled data is of great significance. In this paper, a knowledge-entity categorization and visualization system are designed and developed for academic literature, which provides knowledge-entity recognition and categorization, as well as generation and visualization of the knowledge-graph. Hence it is able to help researchers effectively analyze the knowledge relations and research hotspots.

Keywords　Information extraction　Entity categorization　Knowledge relationship mining　Data visualization　Academic literature

0　引　言

近年来,随着互联网的快速普及和硬件存储技术的高速发展,人们可以轻松地通过众多的学术数据库或学术搜索引擎获取到所需的专业文献,如 Google Scholar、百度学术、CNKI、万方数据等。由此看来,从互联网上获取海量的电子文献资源的确成为了一件轻松简单的事情。但是随之出现的问题是,现有的知识服务已经无法满足人们对信息“快速、简单、准确”的需求。面对这样的知识服务需求,我们有必要针对这类专业文献进行更深层次的信息抽取与文本挖掘,建立结构化的专业知识体系,以辅助用户进行文献检索。其中,实体类型标注是信息抽取任务中一个重要的子任务,准确的实体类型信息可以实现对知识点的有效分层和分类,还可以基于实体类型进一步挖掘其中的

收稿日期: 2016-06-21。国家自然科学基金项目(61202269,61472089,61502108);NSFC-广东联合基金项目(U1501254);广东省科技计划项目(2015B010108006,2015B010131015);广州市科技计划项目(2014Y2-00027)。温雯,副教授,主研领域: 机器学习,模式识别,信息检索。伍思杰,硕士生。蔡瑞初,教授。郝志峰,教授。王丽娟,讲师。

知识关系，从而构造知识脉络图。已有的研究大多针对互联网上常见的新闻文本[1]、微博[2]、Tweets[3]、Facebook 等文本进行实体抽取，而针对专业文献这类特殊文本的研究还比较少，因此有必要对这类文本进行更深入的研究。此外，实体类型标注是实体识别的一个重要组成部分，对后续实体关系的抽取也具有重要的意义[4]。

专业领域的信息抽取已经有一些相关的研究，如 Yoshida 等针对生物医学领域实体的研究[5]，毛存礼等对有色金属领域实体的研究[6]，还有针对商务领域产品领域实体的研究[7]等。这些研究都是针对不同领域的专业实体进行实体识别，但是却缺少更深一层的实体类型分析与挖掘，也没有设计和实现一个真实可用的实体类型标注及可视化系统。还有一部分工作，主要是通过分析网络上的大量命名实体，实现了上位词/类别挖掘系统[8]。但是，这些工作缺少对专业文献这类蕴含复杂且丰富知识点的数据的针对性研究，并不能切实解决专业领域知识服务的迫切需求。

针对以上问题，本文设计并实现了一个面向专业文献知识实体的类型标注及可视化系统 KECVS (Knowledge-Entity Categorization and Visualization System)。该系统能够根据用户查询的实体关键词进行实体类型标注，然后可视化地呈现出知识实体之间的类型关系、层次关系和时序演变模式。系统实施简单，标注准确率高，具有很强的实际价值和现实意义。另外，提出的一种面向专业文献知识实体的类型标签抽取及标注方法可以有效地对专业领域的知识实体进行类型标注。同时也得到比较全面的类型标签集合，解决了人工预定义实体类型的局限性和主观性问题，有助于专业知识网络的结构化实现。

1 系统架构

如图 1 所示，KECVS 系统分为 4 层逻辑结构，自底向上分别是：数据获取层、数据处理层、存储层和应用层。各层的功能依次递进，紧密相扣：1）最底层为数据获取层，它的功能是数据源的获取和存储，主要包括在线爬虫及其管理模块、页面解析模块和本地文献存储模块。2）数据处理层，主要为上面两层提供核心处理算法，包括有知识实体边界识别、类型标签抽取方法及基于多标签加权标签传播的类型标注方法等关键技术的实现。3）中间的存储层主要是把处理后的数据进行数据库存储并建立索引，然后对数据进行知识实体关系图建模，并转换成 JSON 格式数据供应用层实现可视化。4）最上面的应用层主要是与用户进行可视化交互，功能是根据用户的输入反馈出不同的可视化关系图，包括有层次图、关系图、热点图等。以下对部分重要模块进行介绍。

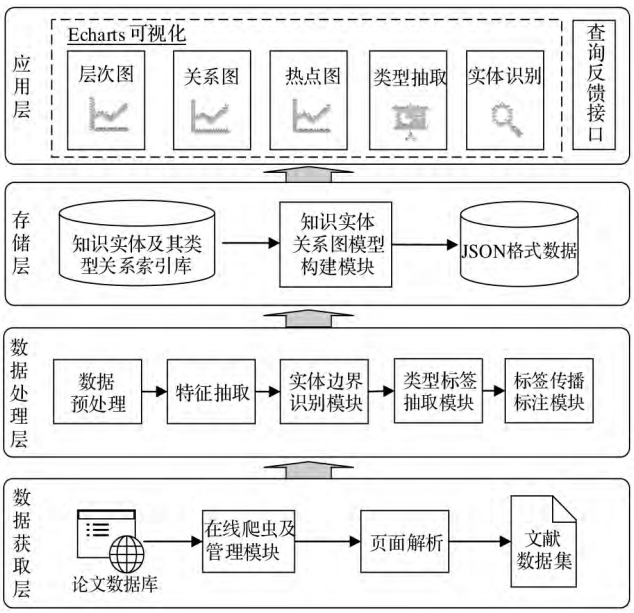

图 1 KECVS 系统架构图

1.1 在线爬虫及其管理模块

管理员可以通过后台指定爬取页面的地址和范围，在线爬虫模块在后台自动化地对文献数据进行爬取并存储在本地，从而实现定向的文献爬取及分析。这样可以简便地把实体类型抽取扩展到其他专业领域或者其他论文数据库，为上面三层提供了充足的数据来源。

1.2 类型标签抽取及类型标注模块

知识实体的类型标签抽取模块主要是对识别到的知识实体进行类型标签抽取，得到类型标签集合和部分标注数据。然后，通过基于多标签加权的标签传播算法对未标知识实体实现进一步的标注，得到的类型标注数据传递给存储层进行本地存储，并建立知识实体及其类型关系索引库，提高检索效率。

1.3 知识实体关系图模型构建及可视化模块

为了更好地将挖掘到的知识脉络实现可视化，我们需要对知识实体及其类型数据进行图模型的构建。根据用户输入的关键词对索引库进行检索，构建出不同的知识实体关系图模型，包括有基于同一类型的实体层次关系树模型(层次图)、基于不同类型分组的知识关系图模型(关系图)和基于时序的知识热点跟踪图模型(热点图)。然后，把得到的关系图模型转换成 JSON 格式的数据，传递到应用层利用 Echarts 进行 Web 可视化实现。

2　关键技术

2.1　基于启发式规则的类型标签抽取方法

通过对专业文献的知识实体内部进行统计分析实验,发现大部分类型词本就存在于实体内部,我们只需要利用启发式规则的方法就可以抽取到大部分的类型标签数据。基于启发式规则的类型标签抽取方法具体步骤如下:

1) 首先,我们需要结合文献摘要信息中知识实体的相关上下文以辅助类型标签抽取。以识别到的知识实体为匹配词,对文献的摘要进行知识实体匹配,再把在摘要中匹配到的知识实体及其后相邻的名词提取出来,添加到知识实体集合中。

2) 利用基于启发式规则的方法对步骤 1) 得到知识实体集合进行类型标签抽取,得到候选类型标签集合,类型抽取的同时获得部分已标注实体,其中启发式规则如下:

启发式规则 1: 设知识实体 $e_i=(w_1,w_2,w_3,\cdots,w_{n-1},w_n)$,$n\geqslant 1$,组成词 w_i 的词性为 c_i。如果 c_i 为名词,则进入规则 2。

启发式规则 2: 设知识实体 $e_i=(w_1,w_2,w_3,\cdots,w_{n-1},w_n)$,$n\geqslant 1$,$w_n$ 是实体 e_i 的最后一个词,且规则 1 中的 $w_i=w_n$,那么把 w_n 加入类型词候选集 T_i。

3) 筛选掉不可靠的类型标签,通过统计类型标签与其所属知识实体共现的频次,然后根据频次特征筛选掉共现频次低且对应知识实体出现频次少的类型标签,筛选后的类型标签集合作为最终输出。

2.2　基于多标签加权的标签传播标注方法

由于基于启发式规则的类型标签抽取方法可以得到绝大部分类型标签集合和一部分的已标注实体,因此我们考虑可以将其转换为一个多标签标注问题,提出一种基于多标签加权的标签传播算法,用于实现剩余未标知识实体的类型标注。

标签传播算法[9] (LPA) 是由 Zhu 等于 2002 年提出,它是一种基于图的半监督学习方法,其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息。节点之间边的权重越大,标签信息越容易在节点间传递。因而,样本节点越相似,它们拥有同样的标签的可能性就越大[10]。我们给出如下定义:

定义 1　转换概率矩阵 T:

$$T_{ij}=P(j\rightarrow i)=\frac{W_{ij}}{\sum_{k=1}^{n}W_{ij}} \tag{1}$$

式中,T_{ij} 表示从节点 x_j 转移到节点 x_i 的概率,也就是知识实体 e_j 转移到知识实体 e_i 的概率。这里转移概率 W_{ij} 可由式(2) 计算得到:

$$W_{ij}=\exp\left(-\frac{S_{ij}^2}{\partial^2}\right) \tag{2}$$

其中,S_{ij} 是知识实体 e_i 和 e_j 的相似度度量,本系统使用编辑距离作为度量方法,∂ 参数用于调整 S_{ij} 的比例,设 ∂ 为 S_{ij} 的均值。

定义 2　类型标签矩阵 Y 设第一层抽取中成功抽出类型词的知识实体个数为 l,未能抽出类型词的知识实体个数为 u,则定义类型标签矩阵 Y 是一个 $(l+u)\times R$ 矩阵,R 为已抽取类型标签的去重个数。设知识实体 e_i 在第一层类型标注后有 K 个类型标签,C_{ik} 是第 i 个实体的 k 标签的出现频次。

$$W_{ik}=\frac{C_{ik}}{\sum_{k=1}^{R}C_{ik}} \tag{3}$$

$$Y_{ij}=\begin{cases}W_{ik} & \text{if } y_i \text{ is label } r_j\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

其中,W_{ik} 是知识实体 e_i 拥有类型标签 k 的权重,以标签 k 在 e_i 中出现的频率来度量。当知识实体 e_i 拥有类型标签 k 时,则 $Y_{ij}=W_{ik}$,否则 $Y_{ij}=0$。

基于多标签加权的标签传播算法主要分两部分迭代过程: 第一次迭代是以初始矩阵 Y_L^0 中所有的已标数据作为输入; 第二次迭代开始,都是以新增已标矩阵 Y_N^t 作为输入,只对新传播得到的标签进行下一轮的传播,这样可以有效地提高算法迭代收敛的速度。在循环迭代的过程中,对于取出的每一个 $Y_j\in Y_U^t$ (e_j 的类型标签),循环遍历 Y_N^t 中的 Y_i,分别计算知识实体 e_j 与 Y_N^t 中每一个 Y_i 对应的知识实体 e_i 的转换概率 T_{ij},若 T_{ij} 大于阈值 ξ (阈值 ξ 取值为 $\sum_{k=0}^{N}T_{kj}/N$,N 为 Y_N^t 的行数),将进行多标签加权的标签传播。如此类推,循环迭代进行标签传播直到没有未标标签矩阵不再改变,迭代结束,输出标签传播后的输出标签矩阵 Y_L^{t+1}。其中,多标签加权传播过程见图 2。

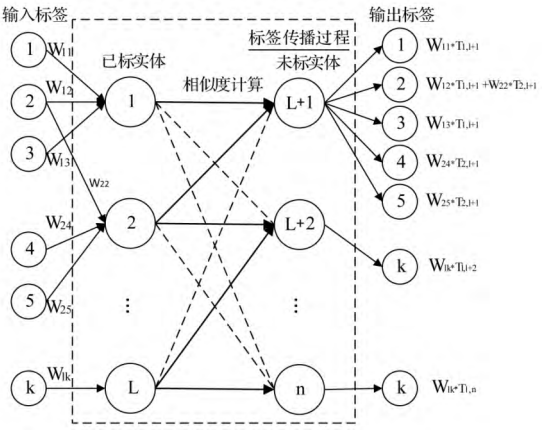

图 2　基于多标签加权的标签传播

2.3 基于知识实体类型关系图模型构建方法

基于知识实体类型关系图模型的构建方法主要包括以下 5 个步骤:

1) 根据用户输入的关键词从知识实体索引库中提取出与该关键词相关的知识实体集合,相关关系包括标题中和摘要中的共现关系、包含关系,以及扩展关系。

2) 构建基于同一类型的实体层次关系树模型,验证知识实体集合中两两个实体之间的扩展或包含关系,如果实体 e_i 包含实体 e_j,则建立树图模型中父子关系,表示 e_i 是 e_j 的父节点,依次类推,建立层次关系模型。

3) 构建基于类型分组的知识关系图模型,对知识实体集合中的知识实体按类型进行分组,统计每个类型分组的权值,分组内的知识实体也按照实体权重降序排序; 筛选出权值最高的 N 个分组,每个分组筛选出排在前 M 个的知识实体(N 和 M 可由用户选择指定) ,按照关键词—类型分组—实体的次序构造三层的图模型。

4) 构建基于时序的知识热点跟踪图模型,根据知识实体的时间进行排序,构建按照半年为周期的时间段分组,分别统计每个时间段出现的相关的知识实体数量,各个时间段分组内的知识实体按照实体权重进行排序,最后以时间分组和对应实体列表构建热点跟踪图模型。

5) 把步骤 2) ~4) 所述的模型转换成 JSON 形式的数据并输出到数据可视化模块。

3 实验及结果分析

3.1 数据使用情况

本文设计实现网络爬虫对 CNKI 中国知网(http://www.cnki.net/) 的计算机类专业论文进行爬取,并以爬取到的论文题目、论文关键词,以及论文摘要作为实验数据,共包含 56 462 篇计算机类核心期刊论文。我们以论文关键词为分词用户词典,对论文标题进行中文分词及知识实体抽取,共抽出 77 364 个知识实体。其中,随机抽取出 500 个知识实体并进行人工类型标注,以标注后的知识实体作为测试集。

3.2 实验结果及分析

为了方便统计和实验,我们对 500 个知识实体进行类型分组,我们把类型样本个数占比最高的前 5 个类型独自作为类型分组,剩下的其他类型由于样本占比较少,合并作为一个类型分组,并统一标注为其他。因此,测试集中共标注 6 种类型,分别是方法、算法、系统、模型、技术和其他。以准确率(Precision) 、召回率(Recall) 和 F1 系数(F1-Measure) 作为评价指标,F1 系数计算方法见式(5) 。

$$F1\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

图 3 是系统分别对这 6 种类型进行知识实体类型标注实验的结果。可以看出,各个分组的准确率都较高,最高的“技术”分组准确率达到 94.11%,说明系统判断的正确率较高,能较好地对知识实体进行类型标注。而召回率相对低一点,原因是有较多知识实体专业性较强,出现频次过少而无法根据实体相似性传播类型标签所导致的,因此这类出现频次较少的知识实体大多会被判定为其他。改进的方法是扩大爬虫爬取范围,增加训练数据,解决部分知识实体训练样本不足的问题。总体而言,实验中的各个类型分组结果都表现较好,F1 值最低的“其他”分组也达到 72.29%,各组平均 F1 值约 77%,说明系统可以有效地对专业知识实体进行类型标注,并具有较高的指导意义。

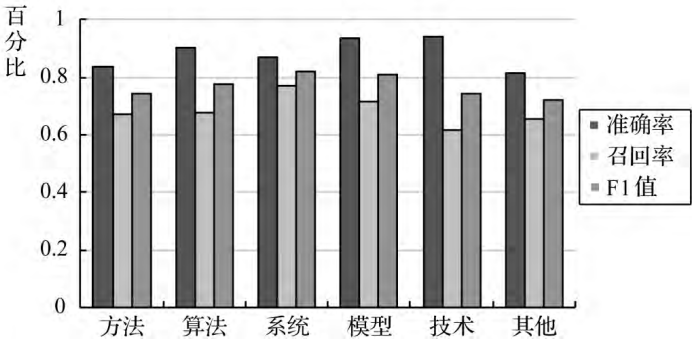

图 3 不同分组的结果对比

在另外一组实验中,将本文的标注方法与传统的 CRF 标注方法进行了实验对比。从图 4 以看出,本文提出的基于多标签加权的标签传播方法效果较好,在各个分组的测试结果中 F1 系数值均比 CRF 方法要高,各分组平均 F1 值提升了 7.61%。其中在“方法”分组中,本文方法的 F1 值相对 CRF 方法提升了 19.66%。由此,我们可以看出 CRF 这类方法并不适合于这类专业文献知识实体的类型标注,可能的原因是这类知识实体长度较短,上下文信息特征不足,导致无法准确判断标注。另外,由于知识实体类型种类较多,类型特征高达 3 000 多维,导致 CRF 模型在训练的过程中十分缓慢,性能较差。由此可以看出,本文提出的方法能更加有效地解决专业文献知识实体的类型标注问题,无论是准确率还是性能都比传统 CRF 方法有较大的提升。

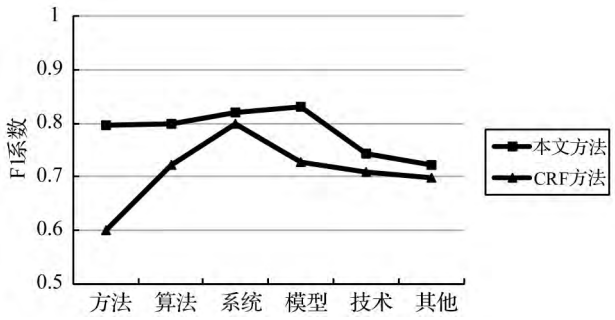

图 4　不同方法的结果对比

4　应　用

4.1　系统环境需求

KECVS 系统的服务器硬件配置要求为: CPU: Intel Core i3 以上,内存 1 GB 以上,操作系统: Windows 7; 环境配置要求: JDK1. 7. 0 及以上版本, PHP5. 5. 12, Apache2. 4. 9, MySQL5. 6. 17; 用户浏览器要求: IE10. X 及以上版本 IE 内核浏览器、Firefox、Chrome 浏览器。

4.2　系统功能及应用

KECVS 系统(http: //kecvs. dmirlab. com/) 首页如图 5 所示,左侧主菜单包括 5 部分内容: 类型抽取、层次图、关系图、热点图和实体识别。首页同时也是类型抽取页面,在右上角检索框输入实体关键词(或者直接点击检索框下的快捷提示词) 后,类型抽取页面会生成类型结果说明,同时在说明下方生成该检索关键词的类型分布饼状图。例如我们输入“条件随机场”进行类型抽取,可以看到“条件随机场”的最符合类型标签是“模型”。

图 5　KECVS 系统首页

知识实体关系图分别包括层次树图、知识关系图及时序热点图三部分。其中,图 6 展示的是基于类型分组的知识关系图的可视化界面,用户输入关键词“条件随机场”检索后,获得以“条件随机场”为中心的不同类型分组下的知识关系图,包括有模型、方法、算法、协议、系统等多个类型下的知识点。例如,从图中可以看到,与“条件随机场”相关的“方法”有“视频分割方法”、“中文词性标注方法”、“机器学习方法”等。

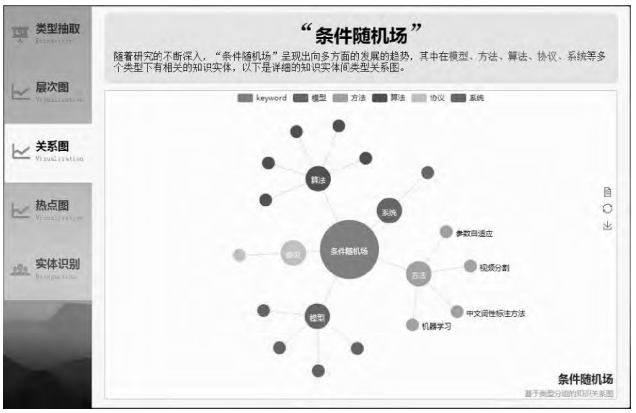

图 6　基于类型分组的知识关系图可视化界面

知识实体识别及类型标注的界面如图 7 所示。用户可以在上方输入框输入想要进行识别的文本或者点击“随机获取摘要”按钮获取文献摘要文本进行识别测试。点击“开始识别”按钮后,可以在下方得到知识实体的类型标注结果,结果中不同的颜色背景代表不同的类型,在最下方有对应颜色的类型图示。例如,从图 7 中可以看到,“条件随机场”、“LDA”、“CRF”等知识实体都能识别出来,并将类型标注为“模型”,而“目标检测”、“机器学习”等则标注为“方法”。

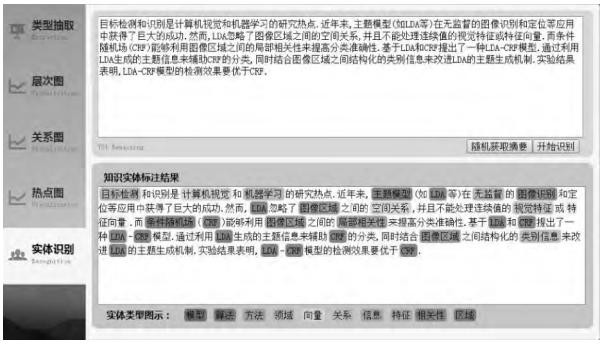

图 7　知识实体识别及类型标注界面

目前,KECVS 系统所有图表采用 Echarts3 实现可视化,均提供数据视图及图表下载功能。

5　结　语

本文设计实现了面向专业文献知识实体的类型标注及可视化系统 KECVS,提供对文献数据的数据爬取、数据清理、实体识别、类型标签抽取、类型标注及知识实体关系图构建等功能,并通过 Web 数据可视化技术呈现给用户。同时,本文通过对比实验验证了本文的方法比传统的标注方法更加适用于专业文献知识实体的类型标注问题,最终实验结果也表明本文的方法

(下转第 144 页)

HTTP [C] //Proceedings of the Second Annual ACM SIGMM Conference on Multimedia Systems, MMSys 2011, Santa Clara, CA, USA, February 23-25, 2011. 2011: 157-168.

[4] De Cicco L, Caldaralo V, Palmisano V, et al. Elastic: a client-side controller for dynamic adaptive streaming over http (dash) [C] //Packet Video Workshop (PV), 2013 20th International. IEEE, 2013: 1-8.

[5] Evensen K, Kupka T, Kaspar D, et al. Quality-adaptive scheduling for live streaming over multiple access networks [C] //Proceedings of the 20th international workshop on Network and operating systems support for digital audio and video. ACM, 2010: 21-26.

[6] Huang T Y, Johari R, McKeown N, et al. A buffer-based approach to rate adaptation: Evidence from a large video streaming service [C] //Proceedings of the 2014 ACM conference on SIGCOMM. ACM, 2014: 187-198.

[7] Tian G, Liu Y. Towards agile and smooth video adaptation in dynamic HTTP streaming [C] //Proceedings of the 8th international conference on emerging networking experiments and technologies. ACM, 2012: 109-120.

[8] Xiao M, Swaminathan V, Wei S, et al. DASH2M: Exploring HTTP/2 for Internet Streaming to Mobile Devices [C] //Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016: 22-31.

[9] Zhu X, Li Z, Pan R, et al. Fixing multi-client oscillations in HTTP-based adaptive streaming: A control theoretic approach [C] //Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on. IEEE, 2013: 230-235.

[10] Jianhui Z, Bin W. Performance analysis of available bandwidth estimation algorithm based on EWMA and Kalman filter [C] //Proceedings of the 2009 International Conference on Multimedia Information Networking and Security. IEEE, 2009, 1: 604-608.

[11] Xiang S, Cai L, Pan J. Adaptive scalable video streaming in wireless networks [C] //Proceedings of the 3rd multimedia systems conference. ACM, 2012: 167-172.

[12] Li Z, Zhu X, Gahm J, et al. Probe and adapt: Rate adaptation for http video streaming at scale [J]. Selected Areas in Communications, IEEE Journal on, 2014, 32(4): 719-733.

[13] Yin X, Sekar V, Sinopoli B. Toward a principled framework to design dynamic adaptive streaming algorithms over http [C] //Proceedings of the 13th ACM Workshop on Hot Topics in Networks. ACM, 2014: 9.

[14] Xie X, Zhang X, Kumar S, et al. piStream: Physical Layer Informed AdaptiveVideo Streaming Over LTE [C] //Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. ACM, 2015: 413-425.

[15] Cherif W, Fablet Y, Nassor E, et al. Dash fast start using http/2 [C] //Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. ACM, 2015: 25-30.

(上接第 91 页)

拥有更高的准确率和更佳的性能。因此，通过 KECVS 系统可以简单便捷地获取到所关注知识点的层次树图、知识关系图及热点跟踪图等，从而为科研工作者在科研方向上提供有价值的参考和启发。未来进一步的工作包括继续完善系统功能，提高系统后台处理性能，为用户提供更便捷、准确和高效的知识服务系统。

参 考 文 献

[1] 吴共庆，胡骏，李莉，等. 基于标签路径特征融合的在线 Web 新闻内容抽取 [J]. 软件学报，2016，27(3)：714-735.

[2] 郑影，李大辉. 面向微博内容的信息抽取模型研究 [J]. 计算机科学，2014，41(2)：270-275.

[3] Liu X, Li K, Zhou M, et al. Collective semantic role labeling for tweets with clustering [C] // International Joint Conference on Artificial Intelligence. AAAI Press, 2011: 1832-1837.

[4] 陈宇，郑德权，赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取 [J]. 软件学报，2012，23(10)：2572-2585.

[5] Yoshida K, Tsujii J. Reranking for biomedical named-entity recognition [C] //Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Association for Computational Linguistics, 2007: 209-216.

[6] 毛存礼，余正涛，沈韬，等. 基于深度神经网络的有色金属领域实体识别 [J]. 计算机研究与发展，2015，52(11)：2451-2459.

[7] 刘非凡，赵军，吕碧波，等. 面向商务信息抽取的产品命名实体识别研究 [J]. 中文信息学报，2006，20(1)：7-13.

[8] 付瑞吉. 开放域命名实体识别及其层次化类别获取 [D]. 哈尔滨工业大学计算机科学与技术学院，2014.

[9] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation [R]. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[10] Chen J, Ji D, Tan C L, et al. Relation extraction using label propagation based semi-supervised learning [C] //Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 129-136.