Journal of Computer Applications ISSN 1001-9081 2016-06-10
计算机应用,2016,36(S1):207-209,227 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2016)S1-0207-03

中文分类体系的构建与查询系统

李金洋,王燕华,樊 艳,汪诚愚,张 蓉,何晓丰*

(华东师范大学 数据科学与工程研究院,上海 200062)

(* 通信作者电子邮箱 xfhe@sei.ecnu.edu.cn)

摘 要:针对中文语言环境中缺少分类体系,无法明确实体类别并建立语义关系的问题,基于维基百科,提出一种混合架构,构建了大规模中文分类体系及其展示查询系统(CTCS2)。CTCS2 包括两个模块:离线模块和在线模块。离线模块又分为 SVM 底层关系抽取子模块、顶层分类树构建子模块两部分。首先,采用 SVM 分类模型抽取语义关系,明确实体类别;然后,通过启发式规则、关联规则挖掘的方式挖掘上层抽象概念关系;其次,使用自底向上的算法从独立的关系中生成完整的中分分类体系,以分类树的形式展现;最后,在线模块分析展示了生成的分类树,并提供语义查询。实验表明,生成的语义关系的准确率高达 95%;为评估分类体系包含中文知识的独特性,使用映射的方法生成 YAGO 的中文版本,YAGO-C,与之相比,CTCS2 中仅有 47.15% 的实体被英文版本覆盖,说明了 CTCS2 的中文独特性。CTCS2 为实体明确了类别类型、在类别类型间建立了语义关系,为构建中文知识图谱提供了基础的语义支持。

关键词:分类体系;知识图谱;支持向量机;启发式规则;维基百科

中图分类号:TP182 文献标志码:A

Chinese taxonomy construction and search system

LI Jinyang, WANG Yanhua, FAN Yan, WANG Chengyu, ZHANG Rong, HE Xiaofeng*

(*Institute for Data Science and Engineering*, *East China Normal University*, *Shanghai* 200062, *China*)

Abstract: Facing the problem that no types or semantic relations are established between items due to the lack of Chinese taxonomy, a hybrid method based on Wikipedia was proposed to develop a Chinese taxonomy construction and search system (CTCS2). There are two modules: offline module and online module. Offline module consists of bottom structure sub-module and top structure sub-module. Firstly, SVM was applied to extract semantic relations indicating entity types. Secondly, high-level concepts were generated by heuristic patterns and association rule mining. Besides, a Chinese taxonomy, presented as a classification tree, was built by a bottom-up algorithm. Finally, online module demonstrated taxonomy and statistical analysis provided a way to search taxonomy and hyponymy relations. Experiments show that that the accuracy of generated sematic relations is 95%. A Chinese version of YAGO, YAGO-C, was generated by mapping process for comparison. Only 47.15% of CTCS2 is covered by YAGO-C, showing that CTCS2 contains unique Chinese knowledge. CTCS2 indicates the type of entities, shows the sematic relations and provides semantic service for constructing Chinese knowledge graphs.

Key words: taxonomy; knowledge graph; Support Vector Machine (SVM); heuristic pattern; Wikipedia

0 引言

知识库、知识图谱通过建立实体之间的语义关联,为知识的组织、管理和展现提供了一种有效途径;其中,分类体系作为知识图谱中的重要部分,对知识图谱中所有的实体进行了严格的类别划分,是知识图谱在语义层次上组织、展现知识的基础。目前已存在的知识库、知识图谱主要集中在英文语言环境下,分类体系的构建方法分为两类:人工构建和自动构建。人工构建虽然质量高,但构建耗时耗力且覆盖率低,如 NELL[1]、DBPedia[2]。另一方面,维基百科包含了海量语义信息,许多自动构建方法将其作为数据源,如 WikiTaxonomy[3]、YAGO[4]。WikiTaxonomy 提取维基百科页面的连通结构以及词法、语法上的特征,将用户生成的页面标签划分为 isA 关系和 notIsA 关系,构建严格的分类体系。YAGO 不仅从维基百科中抽取实体与关系,还结合了 WordNet[5](大型英文词汇语义网),将维基百科中的下层实体与类别映射到 WordNet 中的上层概念。目前最大的知识库 Probase[6],运用 Hearst patterns[7],从网页纯文本中生成 isA 对。显然,数据源必不可少。由于维基百科是多语言的,可以利用不同语言版本间的链接信息,构建跨语言分类体系和知识图谱,如 YAGO3[8]、Xlore[9]。

但是,以上自动的方法无法直接用于构建大规模中文分类体系[10],因为:1)中文数据源的缺乏。构建分类体系十分依赖数据源,例如,YAGO[11] 中使用了 WordNet 的概念构建上层体系,Freebase 为 Google Knowledge Graph 提供了数据支撑。2)无法直接应用英文的模式。例如,英文中使用复数形式抽取实体,而中文没有单复数的形式。3)不同语言版本间覆盖率低。据统计,在中文维基百科中,只有 34.66% 的文章和

收稿日期:2015-09-15;修回日期:2015-10-15。 基金项目:国家自然科学基金资助项目(61232002,61402180,61332006)。

作者简介:李金洋(1991—),女,辽宁铁岭人,硕士研究生,主要研究方向:知识图谱构建、数据挖掘; 王燕华(1990—),男,山西太原人,硕士研究生,主要研究方向:大数据管理; 樊艳(1994—),浙江杭州人,主要研究方向:Web 数据管理和数据挖掘; 汪诚愚(1991—),男,江苏苏州人,博士研究生,主要研究方向:信息检索、Web 数据挖掘和知识图谱构建; 张蓉(1978—),女,上海人,副教授,主要研究方向:知识管理、分布式数据管理; 何晓丰(1969—),男,四川成都人,教授,主要研究方向:机器学习、数据挖掘、信息检索。

15.60%的标签被英文版本覆盖[12]。

因此,本文针对以上问题,将中文维基百科页面及用户生成类别标签作为数据源,构建大规模中文分类体系,并设计实现了展示搜索系统,完成统计分析,支持分类体系、上下文语义关系的展示和搜索,为后续处理提供了语义支持。

综上所述,中文分类体系的构建与查询系统(Chinese Taxonomy Construction and Search System, CTCS2)的主要贡献有:1)训练支持向量机(Support Vector Machine,SVM)分类模型,抽取isA关系(准确率高达95%以上),发掘推断规则、计算置信度拓展隐含isA关系;2)采用自底向上的算法将单独的isA关系整合成完整的分类树(实体数目:581 616,类别标签数目:79 470,isA关系数目:1 317 956);3)提供分类树、上下位关系的展示和查询,为其他工作提供语义支持。

1 系统架构

系统架构如图1所示,CTCS2分为两个模块:1)离线模块,构建大规模中文分类体系。其中,又分为两个子模块:SVM底层关系抽取,使用SVM分类模型,抽取实体关系;顶层分类树构建,制定推断规则、计算置信度,采用自底向上的算法构建分类树。2)在线模块,展示分类树、统计分析结果、上下位关系,并支持迭代查询,返回结果。

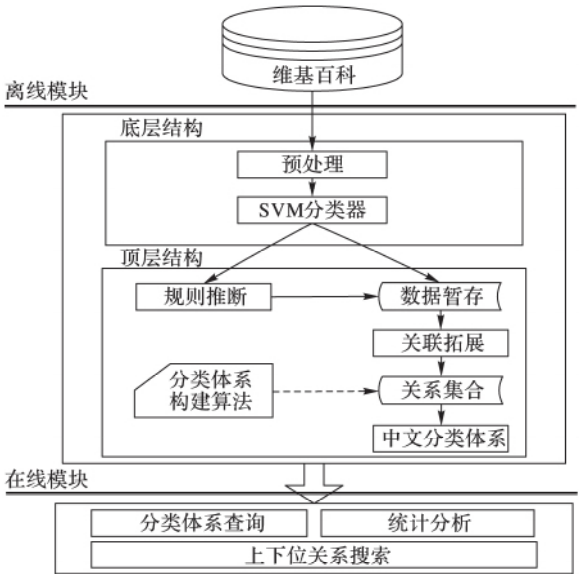

图1 系统架构

维基百科标签分为四类:管理标签、主题相关标签、属性相关标签和类别标签。举例说明CTCS2处理流程:经过预处理保留实体页面,如页面“马云”,标签有“1964年出生”“中国企业家”“阿里巴巴集团”等,但只有“中国企业家”是需抽取的类别标签,故提取有效特征训练SVM分类模型,抽取isA关系“马云 isA 中国企业家”;属性相关标签可以用于推断实体的类别信息,由“1964年出生”推断出“马云 isA 人物”;由于类别标签的抽象程度不同,如“人物”“中国企业家”,前者抽象层次更高,需将其关联起来,通过关联规则挖掘,计算置信度,发掘隐含isA关系“中国企业家 isA 人物”。最后,通过分类树构建算法自底向上地将独立的关系整合起来,用分类树的形式展现。获得分类树后,进行统计分析,由在线模块提供展示、查询等语义支持。

2 离线模块

离线模块完成了中文分类体系的构建,包含实体581 616项,类别标签79 470个。为了评估关系的正确性,随机抽取2 000条关系,计算正确率置信度为0.95的置信区间,为97.60% ± 0.71%。

2.1 SVM底层关系抽取子模块

首先,预处理数据,中文繁简转化,过滤维基百科中的无用页面,如模板页面、管理页面、重定向页面等。然后,将isA关系抽取问题看作分类问题,提取7个有效特征,训练SVM模型抽取实体和isA关系。SVM是一种二分类模型,其基本模型定义为特征空间上的间隔最大的线性分类器,此模块根据特征将维基标签划分为正例和负例,抽取isA关系。

特征可以分为两大类:实体无关特征(特征1~4),实体相关特征(特征5~7)。

特征1 标签长度。过长或过短的标签都不能很好地描述实体。

特征2 词性标注。有效标签通常是名词或名词短语。使用分词和词性标注技术,将标签中心词的词性作为特征。

特征3 主题相关标签。一些过于抽象化的词,如经济、政治、历史等,是主题相关的。将它们收集成词表,把标签中心词是否属于词表作为特征。

特征4 语言模式。英文中,“前修饰词+中心词+后修饰词”是有效的模式[4]。与之类似,中文中,模式“修饰词+的+中心词”可以用来匹配有效标签。

特征5 实体与标签的重复序列。将实体与标签是否出现重复序列作为特征。例如,“政党”是“共产党”的类别标签。

特征6 中心词匹配。与特征5相似,若重复序列为标签中心词,则该标签更可能是有效的。

特征7 标签混杂度。标签用于标识多个实体,应用NER技术识别实体,定义混杂度为所有实体中比例最大的命名实体类别的占比。设定阈值,将混杂度是否大于该值作为特征。

2.2 顶层分类树构建子模块

对于SVM分离出的负例标签进一步处理,制定推断规则,通过关联规则挖掘,计算置信度,抽取上层标签,最终采用自底向上的算法构建整个分类树。

表1 推断规则示例

类别	正则表达式	推断数量	准确率/%
城市	(.* 省)市镇	32 091	100
人物	(.* ?\d{1,4}年)逝世	10 148	99
人物	(.* ?\d{1,4}年)出生	4 801	99
君主	(.* ?)(君主|国王)	3 649	100

isA关系推断 针对维基属性相关标签制定推断规则,采用正则表达式匹配,抽取属性相关标签中的isA关系。表1列举了推断规则和正则表达式示例,如属性标签“1964年出生”,通过正则表达式“(.* ? \d{1,4}年)出生”推断出类别“人物”,且此项规则的推断数据有4 801条,准确率为99%。部分推断规则见表1。

关联规则挖掘 通过计算置信度,关联抽象层次不同的上层标签,发掘隐含isA关系。$S(x)$表示标签x对应的实体集合,置信度计算公式如下:

$$conf(c_1, c_2) = S(c_1) \cap S(c_2) / S(c_1) \quad (1)$$

例如计算“中国企业家”“人物”的置信度,即同时被分类为“中国企业家”和“人物”的实体数目与分类为“中国企业家”实体数目的比值,该值大于阈值,则建立关系“中国企业家 isA 人物”。

分类树构建算法　最后采用自底向上的算法整合单独的 isA 关系,如算法 1 中所示,输入为 isA 集合,初始化时,子树集合 T^{sub} 即为 isA 关系集合,通过以下三种操作构建分类树:

1) 子树水平融合。判断子树根节点相同,进行水平融合,扩展子树宽度(行 3 ~ 7)。

2) 去环操作。针对可能形成有环图,进行去环操作。使用深度优先搜索(Depth First Search,DFS)算法检测图是否连通,在每个连通分量上去除不包含在 DFS 生成树上的边,达到去环的目的(行 8 ~ 12)。

3) 子树垂直融合。判断各个子树的根节点和叶子节点相同,子树垂直融合,增长子树深度(行 13 ~ 17)。

重复迭代直至无子树生成、改变; 最后构建顶层概念,将子树连接到 root 根节点,返回分类树 T。

算法 1　分类树构建算法。

输入: isA 关系集合 S;

输出: 中文分类树 T。

算法描述如下:

1. T^{sub} 表示子树集合,G 表示由 S 生成的图;
2. $T^{sub} \leftarrow S$;
3. for each $T(x) \in T^{sub}, T'(x) \in T^{sub}$ do
4. if Identical ($T(x), T'(x)$) then　//根节点相同
5. HorizontalMerge ($T(x), T'(x)$);　//子树水平融合
6. end if
7. end for
8. for each $g \in G$ do
9. while ContainCycle (g) do
10. RemoveCycle (g);　//基于 DFS 的去环操作
11. end while
12. end for
13. for each $T(b) \in T^{sub}, T(x) \in T^{sub}$ do
14. if ShareNode ($T(b), T(x)$) then　//节点有重叠
15. VerticalMerge ($T(b), T(x)$);　// 子树垂直融合
16. end if
17. end for
18. build top concepts for T^{sub} to construct taxonomy;
19. return T;

3　在线模块

离线模块负责构建分类树,在线模块负责展示系统、分析结果,提供查询功能。本系统使用 Java 语言开发实现,数据库使用 MySQL,运行平台为 Tomcat。

系统分为三个部分: 分类树展示与查询页面、统计分析展示页面和上下位关系搜索与展示页面。系统提供了分析、展示功能,并支持语义查询。

1) 分类树的展示与查询。如图 2 所示,分类树展示查询页面,展示静态中文分类树; 动态查询,输入树的层数 n,默认从根节点查询展示相应子树。

2) 统计分析展示。如图 3,分类树分析结果页面,分析分类树,包括每层实体集大小的统计分析,典型分类标签、实体示例和树顶层类别占比统计分析,帮助用户了解分类树特点。

3) 上下位关系搜索。如图 4,上下位关系搜索与展示页面,用户输入一个词条,系统返回其在分类体系中的上下位关系,并且支持关联词条迭代查询。例如,输入“中国动物”,系统搜索其在分类树中的位置,返回上位词“动物”,下位词“藏狐”“亚洲豹”“大熊猫”等; 点击上下位节点,进行迭代查询,如点击“动物”,系统迭代查询“动物”,返回其上位词“生物”,下位词“中国动物”“美国动物”“泰国动物”等。

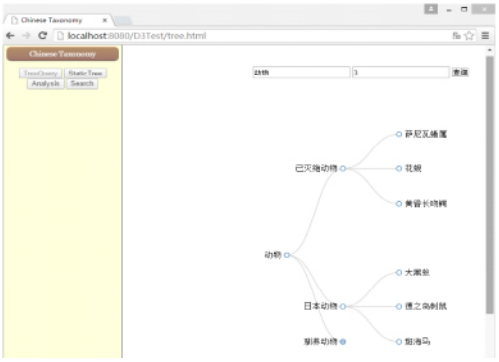

图 2　分类树的展示与查询页面

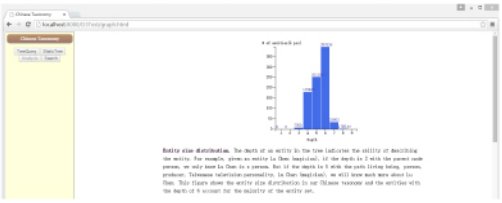

图 3　分类树分析结果页面

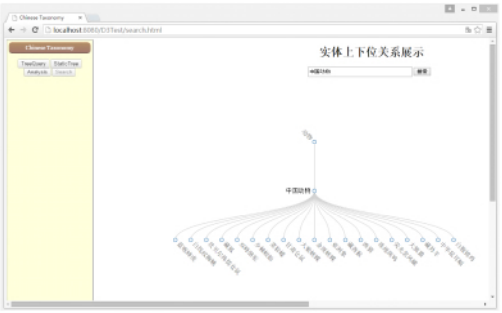

图 4　上下位关系搜索与展示页面

4　结语

本文基于维基百科,采用混合架构构建了大规模中文分类体系,及其展示查询系统(CTCS2),使用有效特征训练 SVM 分类模型抽取 isA 关系,制定推断规则,通过关联规则挖掘,计算置信度,拓展隐含 isA 关系,采用自底向上的算法构建分类树。并在此基础上进行统计分析,提供树、上下位关系的查询和展示功能,为其他工作提供了语义支持。

维基百科中包含的知识不仅局限在实体与类别上,从纯文本中可以挖掘更多的关系。考虑将抽取更丰富的关系作为未来工作的重点,同时增加中文数据源如百度百科,拓展中文分类体系。

参考文献:

[1] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]// Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2010: 529 – 573.

(下转第 227 页)

一方面由于学习者大多已经工作,只能利用业余时间来完成课程学习,因此学习者的学习行为既不同于全日制学生又不同于 MOOC 这类开放型网络教学平台的用户,粗粒度的度量指标已不适用于网络学院学习者的行为风格分析,为此本文基于细粒度的指标对网络学院学习者的学习行为进行了深入分析,发现了网络学院学习者的一些学习行为特征,并研究了这些特征和成绩的相关性,这为教学者改进教学、有针对性地进行指导、更准确地对学习者进行评测提供了依据。

参考文献:

[1] NESTERKO S O, DOTSENKO S, HAN Q, et al. Evaluating the geographic data in MOOCs[C/OL]// Proceedings of Neural Information Processing Systems Workshop on Data Driven Education. [2015 - 10 - 15]. http://nesterko.com/files/papers/nips2013-nestero.pdf.

[2] BALAKRISHNAN G, COETZEE D. Predicting student retention in massive open online courses using hidden Markov models[R]. Berkeley: University of California at Berkeley, Department of Electrical Engineering and Computer Sciences, 2013.

[3] KIZLCEC R F, PIECH C, SCHNEIDER E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses[C]// Proceedings of the Third International Conference on Learning Analytics and Knowledge. New York: ACM, 2013: 170 - 179.

[4] YANG D, SINHA T, ADAMSON D, et al. "Turn on, tune in, drop out": Anticipating student dropouts in massive open online courses [C/OL]// Proceedings of the 2013 NIPS Data-Driven Education Workshop, 2013. [2015 - 10 - 15]. http://lytics.stanford.edu/datadriveneducation/papers/yangetal.pdf.

[5] ZHU H, ZHANG X, WANG X, et al. A case study of learning action and emotion from a perspective of learning analytics[C]// Proceedings of the 17th International Conference on Computational Science and Engineering. Washington, DC: IEEE Computer Society, 2014: 420 - 424.

[6] 蒋卓轩, 张岩, 李晓明. 基于 MOOC 数据的学习行为分析与预测[J]. 计算机研究与发展, 2015, 52(3): 614 - 628.

[7] JOVANOVIC M, VUKICEVIC M, MILOVANOVIC M, et al. Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study[J]. International Journal of Computational Intelligence Systems. 2012, 5(3): 597 - 610.

[8] WEN M, ROSE C P. Identifying latent study habits by mining learner behavior patterns in massive open online courses[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM, 2014: 1983 - 1986.

[9] SINHA T. "Your click decides your fate": Leveraging clickstream patterns from MOOC videos to infer students´ information processing & attrition behavior[J]. arXiv preprint arXiv, 2014: 1407.7143.

[10] DIEZ J, LUACES O, ALONSO-BETANZOS A, et al. Peer assessment in MOOCs using preference learning via matrix factorization [C]// Proceedings of the 2013 NIPS Workshop on Data Driven Education, 2013 [2015 - 10 - 15]. http://lytics.stanford.edu/datadriveneducation/papers/diezetal.pdf.

[11] SHAH N B, BRADLEY J K, PARECH A, et al. A case for ordinal peer-evaluation in MOOCs[C]// Proceedings of the 2013 NIPS Workshop on Data Driven Education. [2015 - 10 - 15]. http://lytics.stanford.edu/datadriveneducation/papers/shahetal.pdf.

[12] ANDERSON A, HUTTENLOCHER D, KLEINBERG J, et al. Engaging with massive online courses[C]// Proceedings of the 23rd International Conference on World Wide Web. New York: ACM, 2014: 687 - 698.

(上接第 209 页)

[2] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a Web of open data[C]// Proceedings of the 2008 Internatioanl Semantic Web Conference, LNCS 4285. Heidelberg: Springer Berlin, 2008: 11 - 15.

[3] PONZETTO S P, STRUBE M. Deriving a large-scale taxonomy from Wikipedia[C]// AAAI'07: Proceedings of the 22nd National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2007: 1440 - 1445.

[4] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge [C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 697 - 706.

[5] BARBU E. Property type distribution in WordNet, corpora and Wikipedia[J]. Expert Systems with Applications, 2015, 42(7): 3501 - 3507.

[6] WU W, LI H, WANG H, et al. Probase: a probabilistic taxonomy for text understanding[C]// SIGMOD'12: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2012: 481 - 492.

[7] HEARST M A. Automatic aquisition of hyponyms from large text corpora[C]// Proceedings of the 14th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1992: 539 - 545.

[8] SHCHANEK F M, KASNECI G, WEIKUM G. Yago: a large ontology from Wikipedia and WordNet, DELIS-TR-594[R/OL]. [2015 - 09 - 01]. http://delis.upb.de/paper/DELIS-TR-0594.pdf.

[9] WANG Z, LI J, WANG, et al. XLore: a large-scale English-Chinese bilingual knowledge graph[C]// Proceedings of the 2013 International Semantic Web Conference (Posters & Demos). Zurich: [s.n.], 2013: 121 - 124.

[10] WANG C, GSO M, HE X, et al. Challenges in Chinese knowledge graph construction[C]// Proceedings of the 31st IEEE International Conference on Data Engineering Workshops. Piscataway: IEEE, 2015: 59 - 61.

[11] SUCHANEK F M, HOFFART J, KUZEY E, et al. YAGO2s: modular high-quality information extraction with an application to flight planning[C/OL]// Demo at the German Computer Science Symposium (BTW 2013). 2013: 515 - 518. [2015 - 09 - 01]. http://www.btw-2013.de/proceedings/YAGO2s%20Modular%20HighQuality%20Information%20Extraction%20with%20an%20Application%20to%20Flight%20Planning.pdf.

[12] LI J, WANG C, HE X, et al. User generated content oriented Chinese taxonomy construction[C]// Proceedings of the17th Asia-Pacific Web Conference, LNCS 9313. Zurich: Springer International Publishing, 2015: 623 - 634.