基于子图相交的社交账号与知识图谱实体对齐

刘家祝,郭 强,吴碧伟,曾明勇
(江南计算技术研究所,江苏 无锡 214085)

摘 要:社交媒体与知识图谱的数据各具特点,相互之间的数据互通具有较强的现实意义,而社交账号与知识图谱实体的对齐是数据互通的前提。针对社交媒体与知识图谱的特点,提出了一种基于子图相交的对齐方法,旨在给定社交账号的情况下,根据社交账号的相关信息在知识图谱中找到正确的对应条目。该方法在候选实体生成阶段对比实验了不同的生成策略。在目标实体选择阶段提出一种基于子图相交的算法,利用社交账号的社交关系在知识图谱中映射成子图。子图相交算法通过考察子图中候选实体周围顶点的"稠密"程度,确定社交账号所对应的目标实体。由于该领域尚无公开可用的测试数据集,构造了一个基于 Twitter 与 Wikidata 的对齐数据集,使用该数据集对该方法进行评估,对比测试了标题匹配算法和 AGDISTIS 算法,子图相交算法能够达到更好的效果。

关键词:社交媒体;知识图谱;子图;社交关系;对齐
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2020)05-0010-06
doi:10.3969/j.issn.1673-629X.2020.05.003

Subgraph Intersection Based Alignment between Social Media Account and Knowledge Graph Entity

LIU Jia-zhu, GUO Qiang, WU Bi-wei, ZENG Ming-yong
(Jiangnan Institute of Computing Technology, Wuxi 214085, China)

Abstract: The data of social media and knowledge graph have their own characteristics, and the data exchange between them has strong practical significance. The alignment of social accounts and knowledge graph entities is the premise of data exchange. Focused on the characteristics of social media and knowledge graph, an alignment method based on subgraph intersection is proposed to find the correct corresponding entries in knowledge graph under given social media accounts. In the phase of candidate entity stage, different generation strategies are compared and experimented. A subgraph intersection algorithm in the target entity selection stage is presented, which creates subgraphs by using the social relations of social media accounts within knowledge graph. By investigating the "density" of vertices around candidate entities in the subgraph, the target entities corresponding to social media accounts are selected. There is no publicly available data set for testing and evaluation in this field, so an aligned data set based on Twitter and Wikidata is constructed to evaluate the proposed method, and the algorithm based on title comparison and AGDISTIS algorithm are tested. Subgraph intersection method can achieve better results.

Key words: social media; knowledge graph; subgraph; social relationship; alignment

0 引 言

随着互联网基础设施的完善和移动互联网设备的普及,社交媒体越来越深入社会生活的方方面面。特别是对于公众人物和组织,社交媒体是他们与外界保持信息流通的重要媒介。社交媒体的数据具有实时性高,非结构化等特点。知识图谱的概念是由 Google 公司在 2012 年提出[1],是一种揭示实体之间关系的语义网络,可以对现实世界的事物及其相互关系进行形式化描述。知识图谱的数据具有实时性较差[2]、准确度高[3]等特点。

社交媒体和知识图谱互相利用对方数据的特点,将对知识图谱扩充与社交网络分析等领域产生积极的促进作用,而社交账号与知识图谱实体对齐是联通二者的桥梁。文中研究的目的在于将社交账号当作一种实体,利用社交账号相关信息和知识图谱实体相关知识,研究社交账号与知识图谱实体的链接技术,将社交

收稿日期:2019-07-01 **修回日期**:2019-11-06 **网络出版时间**:2020-01-10
基金项目:国家科技部重点研发计划项目(2018ZX01028101)
作者简介:刘家祝(1986-),男,硕士研究生,研究方向为知识图谱、大数据。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20200110.1118.018.html

账号与知识图谱中正确的实体链接起来。

为了对齐账号与实体,提出的方法分为两个步骤:候选实体集生成与目标实体选择。图 1 给出了一个对齐过程的示例。

图 1　Twitter 账号与 Wikidata 对齐

在候选实体生成步骤中,利用知识库提供的搜索服务,在不同的搜索策略下构建候选实体集。在目标实体选择步骤中,提出了一种子图相交算法,旨在利用从社交媒体账户中提取的社交关系,通过搜索服务映射到知识图谱中,形成知识图谱的子图,利用子图存在的相关特性来选择目标实体。

1　相关工作

社交账号与知识图谱实体的对齐问题是近年来在知识图谱研究领域的热点问题。2017 年 Trendo 大学的 Nechaev Y 等人[4]首次提出该问题,他们研究了 Twitter 账号与 DBpedia 之间的链接问题,基于监督学习给出了初步解决方案并提出了 SocialLink 问题,指出跨社交网站的账号链接是其中的难点和重点。文献[5]提出了对 SocialLink 问题的改进,引入了 Social Embedding 的概念,与知识图谱中的 RDF graph embedding 方法配合使用,以提高对齐问题的评估指标。

实体链接是指将文本中的实体提及(entity mention)链向知识图谱实体的过程[6]。文献[7]研究了不同策略在候选实体集构建过程中的作用。文中研究的问题是将社交账号链向知识图谱实体,故实体链接问题在实体链接研究框架、实体链接步骤及各阶段所采用的技术方法等方面[8]对本课题具有较强的参考意义。

Usbeck R 等人[9]发布的 AGDISTIS 系统试图利用构建知识图谱中的子图的方式,完成批量的实体链接工作。在目标实体选择阶段,他们采用 HITS(hyperlink-induced topic search)[10]或 PageRank[11]算法,选取重要程度最高的实体为目标实体。AGDISTIS 系统在本课题的数据集上取得了 0.537 左右的准确率。链接效果不理想的原因在于 HITS 算法依赖于子图中候选实体节点的入度,但这一特征在 Wikidata 实体中不明显。该系统的子图构建方法对本课题具有一定的参考价值。

由于社交媒体通常严密防范网络爬虫,同时提供的 API 也有严格的限制[12],获取社交账号的社交关系并不容易。文献[13]收集了一个小规模的社交图数据集,并证明它可以有效地改善社交账号对齐的结果。知识图谱中的实体通常包含指向其他实体的链接,这些实体可以被看作一种社交关系图。

2　问题定义与方法

文中研究的目的是针对给定的 Twitter 账号 t,在知识图谱 KG 中找出对应的实体 e_t,这里的知识图谱特指 Wikidata[14]。令集合 C 为账号 t 在 KG 中生成的候选实体集,$C=\{c_1,c_2,\cdots,c_n\}$,函数 φ 为根据账号 t 在知识图谱 KG 中生成的候选实体集,函数 ψ 为计算候选实体 c_i 为正确实体的可能性,链接过程可以形式化地描述为如下两个部分:

(1)候选实体集生成:$C=\varphi(t,\mathrm{KG},\mathrm{Tr})$,其中 Tr 为搜索策略集。

(2)目标实体选择:$\tilde{e}_t=c_q$,其中 $c_i\in C$ 且 $\underset{c_q}{\operatorname{argmax}}(\psi(c_q))$ 成立。

2.1　候选实体集生成

为了能够正确地将 Twitter 账号与知识图谱实体对齐,首先需要根据账号信息生成包含正确命名实体的候选实体集 C。这里利用 Wikidata 提供的搜索服务生成候选实体集。

Wikidata 搜索页面提供关键词搜索服务,能够对实体条目的标题和内容进行搜索。使用搜索页面生成候选实体集的关键在于提供合理的搜索关键词:既能生成包含目标实体的候选实体集,又能减少候选实体集实体数量。根据这一目的,测试了三种搜索策略:

用户名策略(S_{user}):使用 Twitter 账号的用户名。用户名为社交网络用户自行设置的字符串,可以更改或重复,国内社交媒体一般称为“昵称”。这个策略中使用原始的用户名作为搜索关键词。

用户名去符号策略(S_{remove}):由于用户名为用户自行设置的字符串,用户名中可能包含标点符号或者图像符号。这个策略将除空格与文字字符以外的所有

符号去掉。

用户名分割策略(S_{split}):这个策略以策略 S_{remove} 中产生的字符串为基础,以空格为分割符,将用户名分割为多个不同的搜索关键词,将多个关键词的搜索结果合并以产生候选实体集。

表 1 给出了一个搜索策略的示例。对于这些策略,可以将多个策略联合使用,即对于某个目标账号,将多条策略生成的多个候选实体集取并集,生成最终候选实体集。这里没有使用 Twitter 账号中更多的信息构造搜索策略,如描述、推文等,因为 Wikidata 的搜索服务对于长字符串的处理并不理想。最后对每一次搜索服务的返回结果取前 k 个,构成最终候选实体集。

表 1 搜索策略示例

策略	搜索关键词
S_{user}	Universität Kiel CAU⊙
S_{remove}	Universität Kiel CAU
S_{split}	Universität, Kiel, CAU

2.2 目标实体选择

在产生候选实体集以后,需要采用一定的方法计算候选实体为目标实体的可能性,选择最有可能的候选实体作为目标实体。这里对三种算法进行实验,分别为标题匹配算法、AGDISTIS 算法、子图相交算法。

2.2.1 标题匹配算法

标题匹配法以 Twitter 账号用户名与候选实体标题字符串的相似度为选择标准,选择第一个与 Twitter 账号用户名完全相同的候选实体为目标实体。

2.2.2 AGDISTIS 算法

AGDISTIS 算法[9]是一种基于图的实体链接方法。该方法利用目标实体提及(entity mention)与上下文中的实体提及,在知识图谱中生成有向子图,以 HITS (hyperlink - induced topic search)[10] 或 PageRank[11]算法计算图中顶点的重要程度,选择重要程度最高的顶点作为预测的目标实体。

在该实验中,将目标账号中存在的社交关系视为上下文实体提及用于生成子图。AGDISTIS 方法生成子图的过程,是提出子图相交算法的基础。

2.2.3 子图相交算法

文中在 AGDISTIS 算法[9]的基础上提出了子图相交算法。AGDISTIS 算法在目标实体选择阶段使用的 HITS[10]或 PageRank[11]算法,其权值的计算依赖于候选实体节点的入度。但是在知识图谱的结构中,发现大量的候选实体入度为 0,进而导致权值计算结果为 0。基于这样的观察,引入了基于子图相交来衡量节点重要程度的算法。

首先从给定社交账号 t 的社交关系中涉及的相关账号生成扩展实体集 C_{E} , C_{E} 中的元素为知识图谱中的实体条目。这里提取了 Twitter 页面中出现的转发(retweet)、提及(mention)、引用(quote)以及关注(following)中的用户作为账号 t 的相关账号集 $\text{RA} = \{\text{ra}_0, \text{ra}_1, \cdots, \text{ra}_n\}$,取 RA 的前 m 个账户生成 C_{E} :

$$C_{\text{E}} = \bigcup_{i=0}^{m} \varphi(\text{ra}_i, \text{KG}, \text{Tr}) \tag{1}$$

针对候选实体集 C 中的每个实体 c_i ,生成一个搜索深度为 d 的扩展图 G_{c_i} 。在这里将知识图谱 KG 认为是一个有向图 $G_{\text{KG}} = (V_{\text{KG}}, E_{\text{KG}})$,其中,顶点 V 是 KG 中的实体,边 E 为 KG 中的关系,那么 $x, y \in V$, $(x,y) \in E \leftrightarrow \exists p: (x,p,y)$ 为 KG 中的一个 RDF 三元组。

扩展图 G_{c_i} 的建立从初始化 $G_0 = (V_0, E_0)$ 开始,其中 V_0 为 $\{c_i\}$, E_0 为空,然后采用广度优先搜索扩展 G_0。定义扩展操作为 ρ ,第 i 步扩展结果为 $G_i = (V_i, E_i)$, $i = 1, 2, \cdots, d$,扩展过程为:

$$G_{i+1} = \rho(G_i) = (V_{i+1}, E_{i+1}) \tag{2}$$

$$V_{i+1} = V_i \cup \{y: \exists x \in V_i \wedge (x,y) \in E_{\text{KG}}\} \tag{3}$$

$$E_{i+1} = \{(x,y) \in E_{\text{KG}}: x, y \in V_{i+1}\} \tag{4}$$

对 G_0 执行 d 次 ρ 操作即得到扩展图 G_{c_i} 。取每个扩展图 G_{c_i} 的顶点集 V_{c_i} ,令交集元素个数:

$$\text{ic}_i = |V_{c_i} \cap C_E| \tag{5}$$

保留得到 ic_i 值最大的实体 c_i 为预测目标实体 $\tilde{e}_t$ 。

算法过程描述如下:

算法:子图相交算法。

输入:目标账号 t ;候选实体集 C ;搜索结果保留数 k ;相关账号保留数 m ;图搜索深度 d ;知识图谱 KG;搜索策略集 Tr;

输出:预测目标实体 $\tilde{e}_t$ 。

步骤:

1. $C_E \leftarrow \varnothing$
2. RA $\leftarrow$ getRelateAccount(t,m)
3. FOR $\text{ra}_i \in$ RA
4. $C_{\text{E}} \leftarrow C_{\text{E}} \cup$ KGSearchService(ra_i,Tr,k,KG)
5. END FOR
6. $\tilde{e}_t \leftarrow$ null
7. ic $\leftarrow$ 0
8. FOR $c_i \in C$
9. $G \leftarrow$ breadthFirstSearch(c_i,d,KG)
10. $V \leftarrow$ getVertexesFromGraph(G)
11. tc $\leftarrow |V \cap C_{\text{E}}|$
12. IF tc $\geqslant$ ic THEN
13. $\tilde{e}_t \leftarrow c_i$
14. ic $\leftarrow$ tc
15. END IF
16. END FOR
17. RETURN $\tilde{e}_t$

3　实验与分析

无论是社交网络分析还是基于知识图谱的实体链接工作,先前的研究者们已经发布了大量的数据集用于研究和测试。文中研究内容为 Twitter 账号与 Wikidata[14]知识图谱实体的对齐工作,无法直接使用这些先前发布的数据集。因此将使用自行构造的数据集对文中方法进行测试和评估。

实体对齐旨在从候选实体集中选择最有可能的实体作为目标实体,故最终的结果只有“成功”或“失败”两种结果。因此,文中衡量算法性能的指标为准确率(Accuracy)。在 2.2.1 中提到的标题匹配算法将作为本次实验的基准算法。这个算法实际上是利用 Wikidata 自身搜索服务来实现 Twitter 账号实体的对齐。

评估将分为三个部分进行。首先对候选实体集生成部分与目标实体选择部分分开评估,然后对总体方法进行评估。

3.1　实验数据

针对文中提出的研究工作,需要一批特殊的 Twitter 账号,这些账号在 Wikidata 中存在相应的实体。同时需要这些 Twitter 账号的基本信息,如账号名、昵称等。由于文中方法主要从 Twitter 用户的社交关系入手,所以 Twitter 账号发表的推文以及关注的账号也是需要的信息。

文中研究所涉及的知识图谱 Wikidata 由维基媒体德国分会首先提出[14]。为了完成数据集的构造,首先通过 Wikidata Query Service,利用 SPARQL[15]语言获取了 3 024 条具有 Twitter 账号的 Wikidata 实体,其中包含 1 379 个人物账号,1 645 个组织账号。

获取的结果格式如表 2 所示,其中 Item 为 Wikidata 中的实体链接,ItemLable 为 Wikidata 中的实体标签,Twitter 为实体所对应的 Twitter 账号名。根据 Twitter 账号名,利用网络爬虫技术[16-17],爬取相关账号的基本信息、推文及关注账号列表。

表 2　Wikidata Query Service 结果示例

Item	ItemLabel	Twitter
http://www.wikidata.org/entity/Q69319	John Kasich	johnkasich
http://www.wikidata.org/entity/Q3956858	Tom Rice	RepTomRice
http://www.wikidata.org/entity/Q3956999	Richard Hudson	RepRichHudson

为了保证能够获取较为可靠的社交关系,去除了推文总数在 300 条以下且关注总数在 100 以下的账号,最终保留账号 2 281 个,其中人物账号 1 086 个,组织账号 1 195 个。表 3 列出了推文和关注相关的统计信息,平均推文数 6 129.63 条,关注账号数 2 937.66 个。

表 3　Twitter 账号数据统计

	最大值	最小值	均值	中位数
推文	70 746	310	6 129.63	3 981.50
关注	614 157	101	2 937.66	1 135.50

3.2　候选实体集生成

在实验数据集上对搜索策略逐条进行评估。评估针对 Twitter 账号类型按照人员、组织和综合(人员+组织)分为三大类。评估的指标为:

(1)候选实体集大小均值(average number of candidate entities,ACE):候选实体集元素个数的平均数量。

(2)非空率(non-empty rate,NER):候选实体集不为空的比率。

(3)覆盖率(coverage rate,CR):在候选实体集中包含目标实体的比率。

在评估候选实体集生成性能之前,需要确定搜索结果保留数 k。图 2 描述了三种搜索策略在测试数据集上的性能表现。为了保证目标实体覆盖率与候选实体集大小之间的平衡,最终选择 $k=5$。

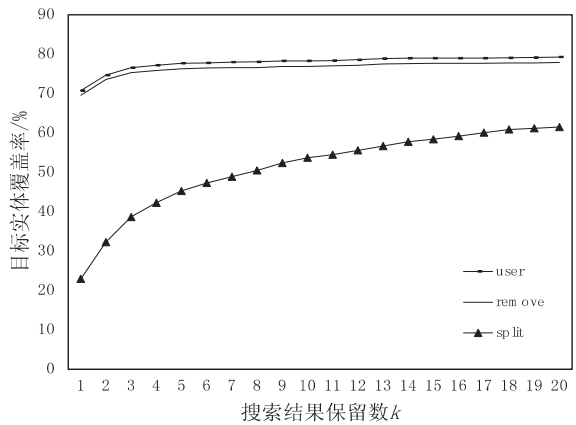

图 2　k 取值与覆盖率的关系

表 4 显示了针对不同搜索策略的实验结果。可以看出 S_{remove} 相对于 S_{user} 虽然去除了非文字符号,但是性能并没有提升。S_{split} 对于单个 Twitter 账号来说增加了搜索次数,提高了候选实体集非空率(NER),但是字符分割带来的语义破坏,影响了目标实体覆盖率(CR)。最终选择使用 S_{user} 策略。

表 4 搜索策略实验结果

策略	人员			组织			综合		
	ACE	NER	CR	ACE	NER	CR	ACE	NER	CR
S_{user}	2.676	0.883	0.837	3.578	0.856	0.743	3.143	0.869	0.788
S_{remove}	2.627	0.868	0.827	3.313	0.804	0.705	2.987	0.835	0.763
S_{split}	4.924	0.991	0.357	4.331	0.928	0.542	4.613	0.958	0.454

3.3 目标实体选择

目标实体选择部分的评估将在包含正确目标实体的候选实体集中进行,候选实体集的生成策略采用 S_{user}。将对三种方法在实验数据集上的准确率进行计算,三种方法分别为:标题匹配算法(A_{title})、AGDISTIS 算法[9](A_{HITS})和子图相交算法(A_{subg})。

AGDISTIS 算法在构建子图时使用的宽度优先搜索深度(d)为 2[9]。子图相交算法在构建子图时,基于子图规模与算法运行时间的平衡,选择宽度优先搜索深度(d)为 3;保留了爬取数据中所有的相关账号,平均个数为 311.788。

表 5 显示了不同方法针对实验数据集中不同类型数据的准确率。可以看出,子图相交算法在针对人员类型账号数据时,相对于其他两种方法性能较好,但是对于组织类型账号数据性能弱于 AGDISTIS 算法。这样的差别在于,在知识图谱中,组织类型的实体往往有多条相关的实体[18]。特别是对于 Wikidata 这类由社区众筹维护的知识图谱来说,针对某个组织往往存在多个实体。

表 5 目标实体选择实验结果

算法	人员	组织	综合
A_{title}	0.626	0.454	0.541
A_{HITS}	0.636	0.728	0.681
A_{subg}	0.94	0.673	0.808

子图相交算法使用相交子图的顶点个数— ic 值作为目标实体的选择依据。这一简单的指标在实验中取得了最高 0.94 的准确率。表 6 显示了目标实体与非目标实体之间平均相交子图顶点数值(ic)上的对比。

表 6 目标实体与非目标实体的平均 ic 值对比

实体类型	人员	组织	综合
目标实体	34.79	24.49	28.8
非目标实体	11	13	12.26

从表 6 可以看出,目标实体与非目标实体在平均 ic 值上具有较大的差异。这充分显示了社交账号中的社交关系映射到知识图谱中形成的子图,在结构上呈现出了聚集的特性,在正确的候选实体附近较为“稠密”,ic 值为描述这一结构特性较为直观的指标。

3.4 总体方法

总体评估针对整个数据集,对于候选实体集为空的数据将标记为对齐失败。表 7 显示了三种方法的总体评估结果。

从表中可以看出,总体性能相对目标实体选择部分有很大下降,这主要是由候选实体集生成部分目标实体覆盖率(HTVC)引起的。对于未包含正确实体的候选实体集,目标实体选择算法是无法找到正确对齐实体的。总体性能特点与目标实体选择部分类似,组织类型数据对齐正确率较低,子图相交算法取得了最好的综合性能。

表 7 总体方法实验结果

方法	人员	组织	综合
A_{title}	0.524	0.337	0.426
A_{HITS}	0.532	0.541	0.537
A_{subg}	0.787	0.5	0.637

4 结束语

提出了一种将社交账号与知识图谱实体进行对齐的算法—子图相交算法。算法通过将目标账号的社交关系图映射到知识图谱中形成子图,以衡量候选实体所在子图位置的实体聚集程度,来选择目标实体。该研究揭示了基于社交关系映射的知识图谱子图,在目标实体“附近”存在聚集特性,利用这一特性预测目标实体取得了较高的准确率。

通过 Wikidata 提供的搜索服务构建了用于测试和评估的对齐数据集,文中方法在该数据集上实现了 0.637 的准确率。

子图相交算法所利用的社交媒体的社交关系图以及知识图谱的图结构等信息,是普遍存在于社交媒体和知识图谱中的。所以该对齐方法可以应用于其他的社交媒体和知识图谱。

下一步的工作可以从三个方面开展。首先是增加方法对不存在目标实体的空链接(NIL)项的处理功能,可以考虑引入更多的特征,采用机器学习方法,设计一个基于监督方法的实体对齐系统。

其次是将映射子图的聚集特性应用于更多的领域,例如其他社交媒体的实体链接、社群分析、话题识别等方面。最后是扩充数据集,添加更多职业和类别的账号,同时加入 NIL 类账号用于机器学习算法的训练和测试。

参考文献:

[1] 黄恒琪,于　娟,廖　晓,等. 知识图谱研究综述[J]. 计算机系统应用,2019,28(6):1-12.

[2] FETAHU B,ANAND A,ANAND A. How much is Wikipedia lagging behind news? [C]//Proceedings of the ACM web science conference. Oxford,England:ACM,2015:28.

[3] HOFFART J,SUCHANEK F M,BERBERICH K,et al. YAGO2:a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence,2013,194:28-61.

[4] NECHAEV Y,CORCOGLIONITI F,GIULIANO C. Linking knowledge bases to social media profiles[C]//Proceedings of the symposium on applied computing. Marrakech,Morocco:ACM,2017:145-150.

[5] NECHAEV Y,CORCOGLIONITI F,GIULIANO C. SocialLink:exploiting graph embeddings to link DBpedia entities to Twitter profiles[J]. Progress in Artificial Intelligence,2018,7(4):251-272.

[6] 陆　伟,武　川. 实体链接研究综述[J]. 情报学报,2015(1):105-112.

[7] 杨紫怡,盛　晨,孔　芳,等. 多策略候选集构建与实体链接[J]. 计算机工程与科学,2018,40(12):2224-2233.

[8] 李茂林. 英文实体链接系统的研究与实现[D]. 北京:北京邮电大学,2016.

[9] USBECK R,NGOMO A C N,RÖDER M,et al. AGDISTIS-graph-based disambiguation of named entities using linked data[C]//International semantic web conference. Trentino,Italy:Springer,2014:457-471.

[10] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM,1999,46(5):604-632.

[11] PAGE L,BRIN S,MOTWANI R,et al. The PageRank citation ranking:bringing order to the web[R]. Stanford,California:Stanford InfoLab,1999.

[12] 游　彬,刘晓然,李　宁,等. 社交网络 Twitter 的推文抽取技术研究[J]. 舰船电子工程,2012,32(9):113-115.

[13] LU C T,SHUAI H H,YU P S. Identifying your customers in social networks[C]//Proceedings of the 23rd ACM international conference on conference on information and knowledge management. Shanghai,China:ACM,2014:391-400.

[14] 贾君枝,薛秋红. Wikidata 的特点、数据获取与应用[J]. 图书情报工作,2016(17):136-141.

[15] BIZER C,SCHULTZ A. The Berlin SPARQL benchmark[J]. International Journal on Semantic Web and Information Systems,2009,5(2):1-24.

[16] 王铁刚. 社交媒体数据的获取分析[J]. 软件,2015,36(2):86-91.

[17] 韩　贝,马明栋,王得玉. 基于 Scrapy 框架的爬虫和反爬虫研究[J]. 计算机技术与发展,2019, 29(2):139-142.

[18] 刘申凯,周霁婷,朱永华,高洪皓. 融合知识图谱和 ESA 方法的网络新词识别[J]. 计算机技术与发展,2019, 29(3):12-17.