北京大学学报(自然科学版) 第 54 卷 第 2 期 2018 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 2 (Mar. 2018)
doi: 10.13209/j.0479-8023.2017.157

利用URL-Key领域术语识别方法

吕书宁[1] 董志安[2,†]

1. 北京工业大学软件学院，北京 100124; 2. 北京信息科技大学网络文化与数字传播北京市重点实验室，北京 100101;
† 通信作者, E-mail: dong.zhian@163.com

摘要 首次提出利用 URL-Key 进行领域术语识别的方法。以 URL 作为媒介，借助已知 URL-Key 的领域性来判断未知领域候选术语的领域性。首先，借助互联网中已有的人工分类领域 URL，根据 URL-Key 在各领域汇总使用的频度，采用基于方差的领域 URL-Key 识别方法，构建领域 URL-Key 词表；然后，利用伪反馈技术，收集候选领域词检索得到的 URL 结果集，根据 URL 结果集构建候选领域术语的 URL-Key 特征向量；最后，利用 SVM 对候选领域术语进行提取。在 4 个领域进行实验，都取得不错的效果。新提出的方法可以有效地解决低频术语识别问题，为低频术语的识别提供新的思路。

关键词 URL; URL-Key; 领域术语; 低频术语; SVM
中图分类号 TP391

Domain Term Extraction Using URL-Key

LÜ Shuning[1], DONG Zhian[2,†]

1. School of Software Engineering, Beijing University of Technology, Beijing 100124; 2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101;
† Corresponding author, E-mail: dong.zhian@163.com

Abstract A new approach was presented for domain term extraction using URL-Key. With the help of known URL-Key's domain, unknown URL-Key's domain can be identified. First, according to the frequency of URL-Key appearing in various fields, a method based on the variance was proposed to identify the domain URL-Key and build the dictionary of domain URL-Key. Then, the pseudo related feedback was used to construct the URL-Key vector of candidate domain terms. Finally, SVM was applied to extract terms. Experiment was conducted on four different domains for Chinese term extraction. Experimental results indicate that the proposed method is quiet effective. In addition, it can effectively solve the recognition problem of low frequency terms, and provides a new way for the identification of low frequency terms.

Key words URL; URL-Key; domain term; low-frequency term; SVM

随着技术的发展，互联网领域已经发生巨大的变化，人们不再局限于通过网络获取数据，还是互联网数据的创造者。目前已经进入“大数据”时代，数据信息不仅规模大，而且错综复杂。新的理论、新的方法和新的概念不断涌现，同时产生大量新的领域术语。人工构建领域术语不仅费时、费力，而且不易更新，因此领域术语自动识别已经成为汉语自然语言处理方面重要的研究课题。

术语识别是基础性的研究工作，有助于领域词典的更新、领域本体的构建以及句法分析的研究。术语识别研究通常分为候选术语提取和识别两个步骤。候选术语提取可以视为术语边界识别的问题。对于汉语，字符之间没有明确的切分边界，识别起来非常困难，Ji 等[1]提出利用特定词汇作为边界取

国家自然科学基金(61671070)和国家语言文字工作委员会重点项目(ZDI135-53)资助
收稿日期: 2017–06–05; 修回日期: 2017–09–02; 网络出版日期: 2017–11–07

代寻找候选术语与其上下文之间的特征关系。Yang 等[2]通过观察候选术语的上下文边界信息，利用条件随机场(conditional random field algorithm, CRF)对候选术语进行识别。有些学者将术语识别的过程看成候选术语在不同领域语料中分布的过程，利用词频–逆文档频率(term frequency-inverse document frequency, TF-IDF)[3]和信息熵[4]等方法进行研究。还有些学者倾向于使用直接统计的方法，根据候选术语的频次、语义和上下文信息[1,5–6]进行术语提取。这些方法通常需要一个大而精确的领域术语语料库作为基础知识进行学习，但是，目前尚无公开的大规模领域术语语料库。

本文放弃使用领域语料库资源，利用互联网中已标注的领域信息，提出一种新的利用 URL-Key 的领域术语提取方法。

作为 Internet 上用来描述信息资源的字符串，统一资源定位符 URL (uniform resource locator)近年来引起研究者关注，被应用在 Web 网页主题分类[7]、查询分类[8–10]以及广告关键词提取[11]等研究中。URL 主要由协议、域名以及路径 3 个部分组成。域名通常由域名主体和域名后缀两部分构成。例如，对域名 match.sports.sina.com.cn 来说，match.sports.sina 是域名的主体词，com.cn 是域名的后缀，而域名的主体词通常与信息资源的类别有关。路径则是指信息资源存放的具体位置。人们在构建路径时，为了方便信息资源归类，通常将不同的资源放在不同的路径下，且为路径起的名称与资源的主题相关，如 football 的路径下应该放与足球相关的信息资源。

通过上述对 URL 的分析发现，人们在申请域名或者创建路径时，通常会思考与信息资源的相关性。虽然各网站的域名不同，但是相同主题下的域名主体词及路径引导词的使用可能相同，通常会集中使用一些领域性强的词语，如体育类中的 sport、军事类中的 mil、汽车类中的 auto 等。也就是说，互联网中的域名主体词及路径引导词被赋予人们智慧的结晶。本文将 URL 中含有领域信息的域名主体词及路径引导词统称为 URL-Key。

本文利用 Yang 等[2]提出的方法提取候选领域术语，将候选术语作为关键词串放入搜索引擎中进行检索，得到相关的反馈信息。与以往的研究不同，本文放弃了所有文本反馈信息，只使用反馈的 URL 作为桥梁，利用 URL-Key 的领域性来判断 URL 的领域性，从而判定候选术语的领域性。例如，候选领域术语为“凯美瑞”，搜索引擎返回 URLs= {“http://www.52car.net/”, “http://www.yicars. com/”, …}，由于 car 是汽车类 URL-Key 的关键词，那么 http://www.52car.net/和 http://www.yicars.com/则为汽车领域的 URL，所以很容易看出“凯美瑞”为汽车领域的领域术语。本文根据 URL-Key 在不同领域中使用的差异性，对张宇等[12]收集的领域 URL 进行领域 URL-Key 提取，构建领域 URL-Key 词表。再根据候选领域术语反馈的 URL 结果集，构建候选领域术语的 URL-Key 向量，最终利用支持向量机(Support Vector Machine, SVM)提取领域术语。

1　领域术语识别

一个句子通常由领域术语和非领域术语组成。领域术语一般为实词，非领域术语一般为虚词、停用词或通用词等。非领域术语通常分布在领域术语的左右，构成领域术语切分标记。利用切分标记，很容易将领域术语从句子中分离出来。

例句: 埃博拉病毒是一种引发埃博拉出血热的烈性传染病。

我们可以利用“是”、“一种”、“引发”、“的”等领域切分标记，将上述句子切分，得到埃博拉、埃博拉病毒、出血热、埃博拉出血热、传染病、烈性传染病等候选领域术语。然后，借助搜索引擎，得到候选术语反馈的 URL 结果集，根据 URL 中含有 URL-Key 的分数，构建候选领域术语的 URL-Key 特征向量。最终，利用已知 URL-Key 的领域性，对未知的候选领域词进行领域性判断，实现从候选领域术语的领域未知性到已知性的转变。

1.1　基于切分标记的候选领域词提取

利用 Yang 等[2]提出的 TCE_DI (term candidate extraction_delimiter identification)方法，使用领域语料库中的切分标记库 DList 分割句子，将一个长句子切分成若干小片段，再以词为单位进行组合，提取候选领域词。如图 1 所示，给定句子 S= C1C2C3C4C5C6C7C8C9…Cn，其中 Ci 代表一个汉字，C1C2 构成词 Wj。句子中有两个切分标记 D1= C3 和 D2= C7C8, D1∈DList, D2∈ DList。DList 将句子切分成 3 个片段: SC1= C1C2, SC2= C4C5/C6, SC3= C9Cn，拼接后，构成候选领域词集合 TC= {C1C2, C4C5, C6, C4C5C6, C9Cn}。

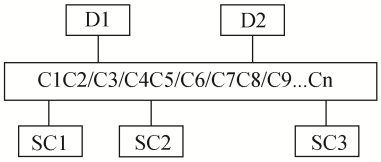

图 1 句子切分样例

Fig. 1 Sample of sentence segment

1.2 基于 URL-Key 的领域词提取

本文利用方差的方法构建领域 URL-Key。利用搜索引擎伪反馈技术，得到相关候选术语的URL反馈信息结果。根据反馈 URL 集合中含有领域URL-Key 的个数，构建候选领域术语的特征向量。最后，利用 SVM 分类器判断是否为领域术语。

1.2.1 领域 URL-Key 构建

在互联网中很多资源都凝聚了人类的群体智慧，如分类网站通常经过人工整理和分类后呈现给用户，具有很高的可信度。张宇等[12]采集了 Yahoo, Google 和 Baidu 的中文分类网页目录，以 Yahoo 网页目录的前 2 层作为分类的标准，并将 Google 和百度网页目录全部映射到 Yahoo 的类别体系中，其 URL 的分布如表 1 所示。

为了得到更多的领域 URL-Key，需要找到更多的领域 URL。由于网络中存在海量的 URL，且每天都不断增长，因此领域 URL 的收集非常困难。通过分析张宇等[12]收集的 URL，发现 URL 分布不均衡，在构建时词语的使用上存在差异，特定词语使用的频次较高，例如: “sport”在体育类 URL 中大量出现，“car”在汽车类 URL 中大量的出现，“mil”在军事类 URL 中大量的出现。因此，可以根据是否含 URL-Key 来判断 URL 的领域性。这样，可以大大地缩减收集领域 URL 的时间，同时将收集领域 URL 的工作转换为提取领域 URL-Key 的工作。

本文利用分割符 T = {: , /,.,-}，对张宇等[12]收集的 URL 进行切分，构成一系列字符串，删除长度为 1 的字串、纯数字串和一些无意义的特殊字符序列(如 http, www, com, cn 等)，最终构建的候选字符串集合为 CandidatekeySet。例如 http://www.tennis.com/player/539/na-li/，经过切分后得到的 Candidatekey 为 tennis, player, na 和 li。根据 Candidatekey 在每个类别中分布的差异性，利用基于方差的方法构建 URL-Key，具体方法参阅文献[8]。

表 1 URL 分布

Table 1 Distribution of URL

领域	URL
体育	1502
汽车	414
军事	255
医疗	2353

1.2.2 基于伪相关反馈的 URL-Key 特征向量构建

伪相关反馈方法假设搜索引擎系统查询反馈的搜索结果排名越靠前，与查询的相关程度越大[12]。本文基于以上假设，构建 TC 的 URL-Key 特征向量，首先利用集合 T 对伪相关反馈 URLs= {url_1, url_2, …, url_l}进行切分，构建一个 bag_i= {Candidatekey_{i1}, Candidatekey_{i2}, …, Candidatekey_{ij}}，目标类别为 C= {c_1, c_2, …, c_n}，其中 i 为搜索引擎返回的前 i 条结果，j 为第 i 个 URL 切分后获取 Candidatekey 的个数，n 为类别个数。为了构建 TC 的领域 URL-Key 向量，需要计算 TC 的伪相关反馈 URL 结果中每一类的 URL-Key 分数 Score(c_n|query)。

$$\text{Count}(\text{URL-Key}_n)=\begin{cases}1, & \text{当bag中含URL-Key}, \\ 0, & \text{当bag中不含URL-Key}。\end{cases} \tag{1}$$

利用式(1)计算 TC 在各个类别中的个数。由于反馈结果中排名越靠前，与查询串越相关，则反馈的 URL 与 TC 的相关性及位置有一定的关系。也就是说，URL 中含有 URL-Key，且其位置越靠前，Score(c_n|TC)越高。本文将反馈结果的前 10 个看成是权重相同的，后面的结果随着排名的增加，权重逐渐降低，具体分布用式(2)计算:

$$\text{Pos}(i)=\begin{cases}1, & 1\leqslant i\leqslant 10, \\ \dfrac{1}{\log(i+1)}, & i>10。\end{cases} \tag{2}$$

综上所述，查询串的 Score(c_n|TC)与反馈结果中的领域 URL-Key 有关，同时也与位置信息有关，计算公式如下:

$$\text{Score}(c_n|\text{TC})=\sum_{i=1}^{1}\text{Count}(\text{URL-Key})\times\text{Pos}(i)。 \tag{3}$$

根据每一类 TC 含有的 URL-Key 的分数值，最终构建 TC 的特征向量{Score(c_1|TC), Score(c_2|TC), Score(c_n|TC)}。

2　实验与结果分析

2.1　实验数据

实验数据为搜狗实验室开放的搜狐新闻数据，该数据收集了搜狐新闻 18 个频道不同领域的数据。本文以体育、军事、汽车和医疗 4 个领域作为实验数据，如表 2 所示。

表 2　实验语料
Table 2　Experimental data

领域	语料数据文档数	实验数据文档数
体育	419768	100
汽车	29116	100
军事	13958	100
医疗	30790	100

为了解决领域词串低频提取问题，本文在不同领域文档中分别抽取前 100 篇文档做实验，且这 100 篇文章涉及的内容各不相同(如篮球、足球、乒乓球、汽车维修、汽车销售、汽车介绍等)，词语分布分散，会出现大量低频词。例如：润滑油出现 1 次，引擎盖出现 1 次，核动力出现 1 次，侦察机出现 1 次，感冒药出现 1 次，感觉神经功能障碍 1 次。因此，可以模拟出一个真实的领域低频词汇语料库。

2.2　评价指标

用人工统计文章中出现的所有领域术语比较困难，因此用人工标注经过 TCE_DI 处理后的候选领域术语。本文使用正确率作为实验的评价指标：

$$\text{Precision}=\frac{N(\text{Correct})}{N(\text{Detected})}\times 100\%, \tag{4}$$

N(Correct)表示正确检测出的领域术语，N(Detected)表示检测到的领域术语总数。

2.3　实验与分析

2.3.1　领域 URL-Key 分析

领域 URL-Key 的提取是解决问题的关键。本文通过对领域 URL 进行切分，利用方差计算公式，计算每一个候选领域 URL-Key 的领域度，再根据领域度的大小进行排序，取排前 100 位的领域度作为候选领域 URL-Key，如表 3 所示。

通过分析，发现领域 URL-Key 主要有以下特点：1) 由英文单词或英文缩写组成，如 sport, auto, car, hospital, nba, vw 等；2) 为拼音或拼音缩写，如 tiyu, che, js, zaojiao 等；3) 为特殊网站域名，如 autohome, zhibo, tiexue 等。

为了得到准确的 URL-Key，根据领域 URL-Key 的上述特点，用人工过滤掉不相干的领域 URL-Key。最终得到 4 个类别、272 个领域 URL-Key，如表 4 所示。

2.3.2　候选领域术语

利用 TCE_DI 算法对 4 个候选领域语料进行切分，共提取候选术语 17123 个，其中体育领域候选术语 4681 个，汽车领域候选术语 3694 个，医疗领域候选术语 4491 个语，军事领域候选术语 3343 个。经过专家的人工标注，最终领域候选术语分布结果如表 5 所示。其中，体育领域候选术语有 1658 个，占总候选术语的 35.4%；非领域术语有 3023 个，占总候选术语的 64.6%。

由于 TCE_DI 算法将语料切分为若干片段，根据每个片段包含词的个数，将候选领域术语分为 1 元候选术语、2 元术语候选术语和多元候选术语，如表 6 所示。可以看出，2 元术语和多元术语中正确领域术语比较多。这是因为有些词语单独出现没有实际的意义，若与领域性强的词组合出现就变成领域词。

表 3　部分领域 URL-Key
Table 3　Examples of URL-Key

领域	URL-Key
体育	sport, nba, espnstar, pingpang, guojizuqiu, basketball, snooker, soccer, tennis, badminton, olympics, olympic, tiyu, tiyv, ty, zhongchao, cbachina, weiqi, guoneizuqiu, sportsbl, chess, saichang, nba, wangqiu, cba, volleyball
汽车	auto, car, xcar, vw, toyota, chinacars, Dongfeng, nissan, vehicles, audi, bitauto, new_cars, chery, newcar, peugeot, webcars, qiche, autohome, che
军事	mil, military, war, army, chinamil, armystar, junshi, js, milnews, tiexue, xinjunshi
医疗	Hospital, beauty, yxy, eat, pharmacy, familydoctor, health, disease, zaojiao, jiankang, pathology, chinamedi

表 4 领域 URL-Key 个数分布

Table 4 Distribution of domain URL-Key

领域	URL-Key
体育	69
汽车	74
军事	55
医疗	74

2.3.3 术语识别

本文将领域术语识别任务看成一个二分类的任务，利用 SVM 对候选术语进行分类。分别随机抽取各领域 80%正确的候选术语和错误的候选术语作为训练语料，20%正确的候选术语和错误的候选术语作为测试语料。如表 7 和 8 所示，体育领域随机抽取 3734 个候选术语作为训练语料，947 个候选术

表 5 术语个数分布

Table 5 Number distribution of Corpus term

领域	候选领域术语/个	领域术语/个	非领域术语/个	占比/%
体育	4681	1658	3023	35.4
汽车	3694	1539	2155	41.6
军事	3344	1196	2148	35.8
医疗	4491	1888	2603	42.0

表 6 术语分布

Table 6 Distribution of domain term

领域	词形	数量	正确数	正确率/%
体育	1 元候选术语	2406	682	28.3
	2 元候选术语	1481	702	47.4
	多元候选术语	794	274	34.5
汽车	1 元候选术语	1509	364	24.1
	2 元候选术语	1680	827	49.2
	多元候选术语	505	348	68.9
军事	1 元候选术语	2292	541	23.6
	2 元候选术语	683	391	57.2
	多元候选术语	369	266	72.0
医疗	1 元候选术语	2899	989	34.1
	2 元候选术语	1037	557	53.7
	多元候选术语	555	343	61.8

表 7 训练数据

Table 7 Train data

领域词结构	体育领域		汽车领域		军事领域		医疗领域	
	训练术语	训练非术语	训练术语	训练非术语	训练术语	训练非术语	训练术语	训练非术语
1 元词	544	1377	287	915	431	1400	789	1526
2 元词	560	621	638	642	211	232	443	383
多元词	217	415	277	123	210	80	273	168
总和	1321	2413	1202	1680	852	1712	1505	2077

表 8　测试数据

Table 8　Test data

领域词结构	体育领域		汽车领域		军事领域		医疗领域	
	测试术语	测试非术语	测试术语	测试非术语	测试术语	测试非术语	测试术语	测试非术语
1 元词	138	347	77	230	110	351	200	384
2 元词	142	158	189	211	180	60	114	97
多元词	57	105	71	34	56	23	70	44
总和	337	610	337	475	346	434	384	525

语作为测试语料；汽车领域随机抽取 2882 个候选术语作为训练语料，812 个候选术语作为测试语料；医疗领域随机抽取 3582 个候选术语作为训练语料，909 个候选术语作为测试语料；军事领域随机抽取 2564 个候选术语作为训练语料，780 个候选术语作为测试语料。

本文利用搜狗搜索引擎，将候选领域术语视为检索词条进行检索。爬取前 100 条 URL 作为数据的实验集合，根据 URL 实验集合中含 URL-Key 的个数，构建候选领域术语的 URL-Key 特征向量。为了比较 URL-Key 方法与 URL 方法，将领域 URL 直接与爬取的 URL 集合进行匹配，构建 URL 特征向量，结果如图 2 所示。可以看出，利用领域 URL-Key 方法的结果好于利用 URL 匹配的方法，这是因为搜索引擎反馈的 URL 不可能与人工搜集的 URL 完全匹配。据统计，每返回 100 条 URL，有大约 20%能进行 URL 匹配，大约 40%能进行领域 URL-Key 匹配。因此，利用 URL-Key 的方法可以解决互联网中每天不断增长的海量 URL 问题。

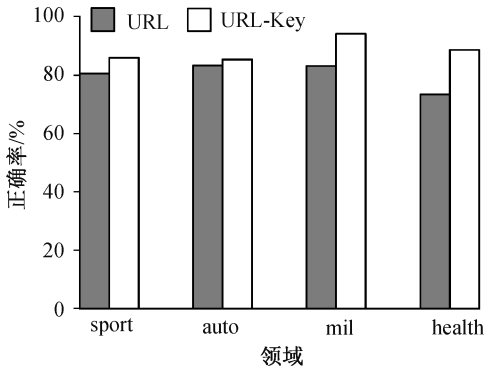

图 2　URL 与 URL-Key 对比实验结果

Fig. 2　Contrast result of URL matching and URL-Key

在进行 URL-Key 匹配的过程中，发现检索到的 URL 中存在着复合领域字串 attatclvolleyhball。Baykan 等[9]以字符为切割单位，利用 all-grams 将 volleyball 提取出来，而本文方法无须对复合的领域词进行处理，只需要通过控制候选术语爬取 URL 的数量，弥补复合词带来的缺陷。是否爬取的 URL 越多，实验结果越准确？我们对爬取不同数量的 URL 进行分析，表明并不是爬取的 URL 数量越多，实验的结果越准确(图 3)。当 URL 取前 70 个时，实验结果最好。这是由于在搜索引擎中，网页越靠后，网页与搜索词之间的关系就越疏远。

利用 URL-Key 作为特征，对于每个候选术语，爬取前 70 个 URL 作为实验结果集。如表 9 所示，体育领域的评价正确率为 86.06%，汽车领域的评价正确率为 87.77%，军事领域的平均正确率为 93.66%，医疗领域的平均正确率为 87.17%。从表 9 可以看出，在各领域中，2 元词的正确率低于 1 元词和多元词。这是因为 2 元词中会有一些单字组成的词语，这些词语通常为领域术语的一部分，本身不能构成领域术语，但在检索返回的 URL 实验集合中却出现很高的 URL-Key 的匹配率。例如：款/福克斯、式/战机、性/皮炎，这些不是领域词，但 URL-Key 匹配率都超过 40%。这是由于搜索引擎会根据用户提供的检索词进行自动扩展检索，以“款福克斯”为检索词，返回的结果中都是关于“新款福克斯”或“福克斯”的信息，因此会影响 2 元词的准确率。

将本文方法与 Yang 等[13]提出的 TV_LinkA 方法进行对比。由于文献评价指标为 Top-*N*，所以本文根据各测试语料中的领域术语个数进行判断，体育领域取 top-337，汽车领域取 top-337，军事领域取 top-384，医疗领域取 top-346 作为评价指标。我

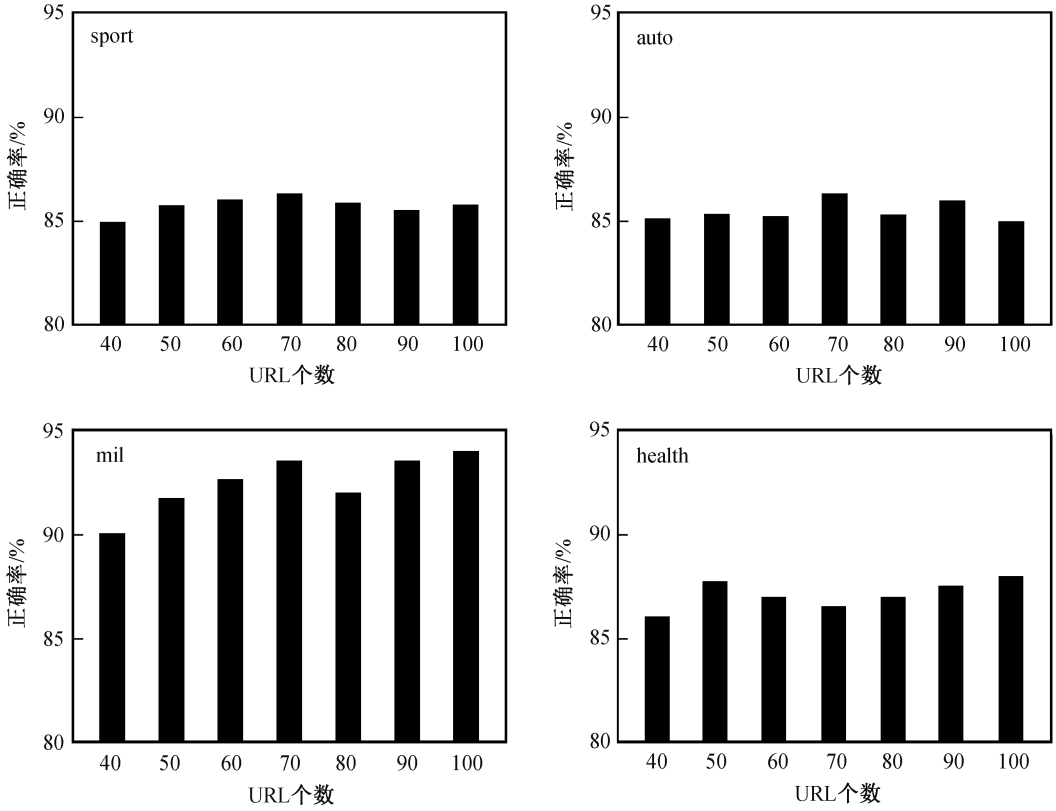

图 3 不同 URL 个数的对比实验结果

Fig. 3 Contrast result of different URL number

表 9 实验结果正确率(%)

Table 9 Experiment results (%)

领域词结构	体育领域	汽车领域	军事领域	医疗领域
1 元词	90.79	86.79	92.39	86.72
2 元词	82.90	82.29	91.26	83.73
多元词	84.48	88.23	97.33	91.07
平均值	86.06	87.77	93.66	87.17

表 10 对比实验结果(%)

Table 10 Contrast results (%)

领域	本文方法	TV_LinkA 方法[13]
体育(top-337)	86.67	66.40
汽车(top-337)	87.77	65.77
军事(top-384)	93.66	70.63
医疗(top-346)	87.17	69.88

们认为 SVM 在测试语料中识别正确术语的正确率与对比实验的正确率可以直接对比。实验结果如表 10 所示。

从表 10 可以看出，TV_LinkA 方法实验结果的准确率没有 Yang 等[13]描述的那样高，主要原因有以下两方面。

1) 领域语料不同，Yang 等[13]用的是 IT 和 Leagl 类语料，本文用的是体育、汽车、军事、医疗类语料，不同质语料可能导致结果的差异。

2) 语料规模不同，Yang 等[13]用的是 IT 领域的数据集，为 6.64 M；本文为了构建真实的低频数据，体育领域数据集为 266 KB，汽车领域数据集为 219 KB，军事领域数据集为 205 KB，医疗领域数据集为 196 KB。Yang 等[13]应用 HIST 算法，根

表 11　各领域低频分布

Table 11　Distribution count of every domain

领域	候选术语类型	数量	低频术语	低频率/%
体育	1 元候选术语	2406	1578	65.58
	2 元候选术语	1481	1102	74.41
	多元候选术语	794	674	84.89
汽车	1 元候选术语	1509	989	74.82
	2 元候选术语	1680	1257	74.82
	多元候选术语	505	428	84.75
军事	1 元候选术语	2292	1368	59.68
	2 元候选术语	683	555	81.25
	多元候选术语	369	316	97.83
医疗	1 元候选术语	2899	1652	56.98
	2 元候选术语	1037	904	87.17
	多元候选术语	555	503	90.63

据候选术语，提取领域句子。由于候选术语出现的频次低，所以提取的领域句子数量少，影响 HIST 算法的结果。因此，语料规模是影响 TV_LinkA 方法的最主要原因。

本文的方法无须考虑实验数据规模，实验结果非常稳定，无论对低频或高频术语都非常适用。

2.3.4　低频术语识别

在候选术语集中只出现 1 次的术语称为低频术语，各领域低频术语分布如表 11 所示。可以看出，本文实验语料的低频率非常高。体育领域语料的低频术语有 3354 个，低频率为 71.65%；汽车领域的低频术语有 2674 个，低频率为 72.39%；军事领域的低频术语有 2239 个，低频率为 66.96%；医疗领域的低频术语有 3059 个，低频率为 68.11%。

在以往的研究中低频术语的识别率非常差，本文提出的方法可以很好地解决这个问题。在 4 个领域，3448 个测试候选术语中，共有 2422 个低频候选术语，占总共测试集的 70.25%。正确识别出 2128 个候选术语，其中 638 个术语，1490 个非术语。错误识别 294 个候选术语，其中 70 个术语，224 个非术语。实验结果如表 12 所示。

表 12　低频实验结果

Table 12　Low frequency experimental results

词语类别	正确识别个数	错误识别个数	正确识别率/%
候选术语	2128	294	87.86
术语	638	70	89.03
非术语	1490	224	84.96

对于普通方法，以单篇文章为语料抽取领域术语时，通常舍弃出现频次为 1 的候选术语，这样的做法将导致大量术语不能准确地识别。使用本文方法得到令人兴奋的结果，低频词语识别准确率达到 87.86%，领域术语准确率达到 89.03%，很好地解决了低频术语识别问题。此外，新词在最初的语料中出现的频次一般较低，因此可以推断本文方法对新词识别同样有效。

3　结语

本文利用互联网中凝聚人们智慧的 URL，提取领域 URL-Key，再利用 URL-Key 进行领域词提取，主要有 4 个方面的贡献：1) 充分地借助互联网中的群体智慧，利用互联网中已存在的领域 URL，提取领域 URL-Key 资源；2) 首次提出利用领域 URL-Key 进行领域词提取，解决了领域语料库资源稀缺问题；3) 该方法可以快速、方便地移植到其他领域和其他语言中进行应用，为领域词提取提供了新的思路；4) 该方法可以有效地解决低频领域术语的提取问题。

本文提出的方法仍有需要改进的地方，例如领域 URL-Key 构建的精确性、SVM 选取特征的多样性以及更多领域的扩展性等。这些问题将在未来的工作中进一步研究。

参考文献

[1] Ji L, Sum M, Lu Q, et al. Chinese terminology extraction using window-based contextual information // Alexander G. Computational linguistics and intelligent text processing. Berlin: Springer, 2007: 62−74

[2] Yang Y, Lu Q, Zhao T. Chinese term extraction based on delimiters // International Conference on Language Resources and Evaluation. Marrakech: DBLP, 2008: 247−254

[3] Salton G, McGill M J. Introduction to modern information retrieval. NewYork: McGraw-Hill, 1983

[4] Chang J S. Domain specific word extraction from hierarchical Web documents: a first step toward building lexicon trees from Web corpora // Proceedings of the Fourth SIGHAN Workshop on Chinese Language Learning. Jeju Island, 2005: 64−71

[5] Sornlertlamvanich V, Potipiti T, Charoenporn T. Automatic corpus-based Thai word extraction with the C4.5 learning algorithm // Proceedings of the 18th conference on Computational linguistics-Volume 2. Saarbrücken: Association for Computational Linguistics, 2000: 802−807

[6] 木合亚提·尼亚孜别克, 古力沙吾利·塔里甫. 哈萨克语 IT 领域术语识别研究与实现. 中文信息学报, 2016, 30(3): 68−73

[7] Qi X, Davison B D. Web page classification: features and algorithms. ACM Computing Surveys (CSUR), 2009, 41(2): 12−35

[8] 李雪伟, 吕学强, 董志安, 等. 利用 URL-Key 进行查询分类. 北京大学学报(自然科学版), 2015, 51(2): 220−226

[9] Baykan E, Henzinger M, Marian L. Purely URL-based Topic Classification // Proceedings of International Conference on World Wide Web3.10. Madrid, 2009: 1109−1110

[10] Shen Dou, Sun Jiantao, Yang Qiang, et al. Building bridges for web query classification // Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. Seattle: ACM, 2006: 131−138

[11] Raju S, Udupa R. Extracting advertising keywords from URL strings // Proceedings of the 21st International Conference Companion on World Wide Web. Lyon: ACM, 2012: 587−588

[12] 张宇, 宋巍, 刘挺, 等. 基于 URL 主题的查询分类方法. 计算机研究与发展, 2012, 49(6): 1298−1305

[13] Yang Yuhang, Lu Qin, Zhao Tiejun. Chinese term extraction using minimal resources // Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Manchester: Association for Computational Linguistics, 2008: 1033−1040