计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(5):190−195 [doi: 10.15888/j.cnki.csa.006922]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于语义相关度主题爬虫的语料采集方法①

周　昆[1,2], 王　钊[3], 于碧辉[1,2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](沈阳市国家税务局 信息中心, 辽宁 沈阳 110013)
通讯作者: 周　昆, E-mail: 982892502@qq.com

摘　要: 针对特定领域语料采集任务, 设计了基于语义相关度主题爬虫的语料采集方法. 根据选定的主题词, 利用页面描述信息, 基于维基百科中文语料训练出的词分布式表示综合 HowNet 计算页面信息相关度, 结合 URL 的结构信息预测未访问 URL 链指的页面内容与特定领域的相关程度. 实验表明, 系统能够有效的采集互联网中的党建领域页面内容作为党建领域生语料, 在党建领域网站上的平均准确率达到 94.87%, 在门户网站上的平均准确率达到 64.20%.

关键词: 生语料采集; 语义相关度主题爬虫; 页面信息相关度; URL 结构信息

引用格式: 周昆,王钊,于碧辉.基于语义相关度主题爬虫的语料采集方法.计算机系统应用,2019,28(5):190–195. http://www.c-s-a.org.cn/1003-3254/6922.html

Corpus Collection Based on Semantic Relevancy Focused Crawler

ZHOU Kun[1,2], WANG Zhao[3], YU Bi-Hui[1,2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Center for Information Technology, Shenyang State Tax Bureau, Shenyang 110013, China)

Abstract: To address the corpus collection, the corpus collection system based on semantic relevancy focused crawler is implemented. Word vector trained by Wikipedia and HowNet are used for calculating page information semantic relevancy with descriptive information according to topical keywords, and the URL structural information is used for calculating the topical relevancy. Experimental results show that this system has better effect on party-construction corpus collection with high precision of average accurate rate 94.87%, while the average accurate rate for web pages is 64.20%.

Key words: corpus collection; semantic relevancy focused crawler; page information semantic relevancy; URL structural information

1　引言

在专业领域进行基于自然语言处理技术的信息处理应用研究时, 无论采用有监督方法、半监督方法, 都需要收集大量领域语料进行模型训练, 所以高效准确的采集专业领域信息, 并构建生语料库是进行相关工作的基础. 主题爬虫能收集与主题相关的网页数据信息, 提高数据相关性, 降低后续处理的复杂程度. 不同专业领域的网站结构以及其所含的领域内容一般不具有良好相似性, 所以主题爬虫很难直接有效的在不同领域间迁移. 将自然语言处理技术应用于特定领域时, 由于通用语料无法满足任务需要, 因此需要采集特定领域语料. 特定领域信息通常具有时效性强的特点, 经常出现新词汇, 仅基于词典匹配的方式采集生语料会导致采集系统的泛化能力不强, 忽略语义上相关却未

① 收稿时间: 2018-11-26; 修改时间: 2018-12-18, 2018-12-28; 采用时间: 2019-01-07; csa 在线出版时间: 2019-05-01

出现在词典中的内容. 此外, 主流网站通常将内容相关的页面存放于相同路径下, 同一网站下链指特定领域页面的 URL 结构上具有相似性. URL 的结构信息可以启发特定领域语料采集过程.

本文针对特定领域语料采集任务, 设计了基于语义相关度主题爬虫的语料采集方法, 根据选定的主题词, 利用页面描述信息, 基于维基百科中文语料训练出的词分布式表示综合 HowNet 计算页面信息相关度, 结合 URL 的结构信息预测未访问 URL 链指的页面与特定领域的相关程度, 可以采集特定领域生语料.

2 相关工作

针对可用于垂直领域语料采集系统的主题爬虫, 文献[1]提出了 Shark-Search 方法, 利用页面相关信息启发式的判断待下载页面. 文献[2]在 Shark-Search 方法基础上将超链接按区域聚类, 但由于导航栏中包含的导航条目较多, 该方法可能会导致导航栏中的主题相关链接被忽略. 文献[3]利用主题分类树结合网页分块与改进的 HITS 算法, 在主题爬虫的准确率上取得了一定的提升, 但主题分类树对于专业垂直领域存在覆盖度不足的问题. 文献[4]设计了基于分类词关键词词频模型的主题爬虫, 应用于地缘政治这一垂直领域的内容采集, 由于主题相关页面中仍可能包含主题无关链接, 该方法可能导致不必要的页面访问. 文献[5]将文本内容与链接结合用作数据采集的判断, 应用于垂直搜索引擎, 但计算复杂度略高. 文献[6]提出了一种利用语义信息预测待爬取页面相关性的主题爬虫模型, 其利用 WordNet 计算语义相关性, 只针对于英文领域. 文献[7]提出“邻居特征”概念, 利用相同路径下链接内容相似的特点爬取主题相关页面, 但在页面内容的计算上没有充分考虑语义信息.

3 基于语义相关度主题爬虫的语料采集方法

本文提出的方法共包括爬取模块、存储模块、后处理模块三个部分. 爬取模块实现了语义相关度主题爬虫, 对给定网站中特定领域内容进行采集与初步过滤, 并将初步结果提交至存储模块进行持久化, 后处理模块对其进行清洗, 去除页面版权声明等噪声后, 得到特定领域生语料.

3.1 爬取模块

爬取模块首先将种子 URL 添加至任务数据库并由任务生成器提取至任务队列. 对于队列中未访问的 URL, 首先计算其链指页面的描述信息页面信息相关度, 若描述信息不相关, 则认为页面内容属于特定领域的可能性较小, 仅对其进行简单的主题词匹配; 否则, 解析该 URL 链指的页面正文, 并对其中每个未访问 URL, 计算其锚文本的页面信息相关度、URL 结构相关度并综合得到未访问 URL 的优先级并据此决定是否丢弃该 URL. 为解决爬取规模较大时计算 URL 结构信息耗时较长的问题, 对保存于存储模块的历史 URL 采取抽样方式作为近似估计, 以加快计算速度. 爬取模块的工作流程如图 1.

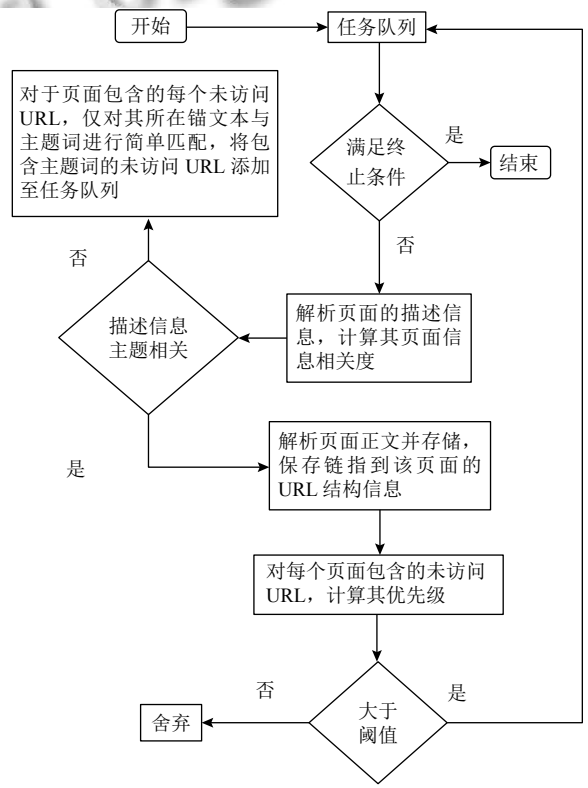

图 1 爬取模块工作流程

3.2 存储模块

存储模块包括内存存储、数据库存储、文件存储三个部分. 爬取过程的中间结果保存于内存. 爬取模块需要加载的部分数据保存于文件. 为解决计算页面信息相关度耗时较长的问题, 存储模块对语义相关度主题爬虫所使用的数据构建倒排索引, 同时将常用数据添加至缓存, 加快爬取模块执行速度. 来自爬取模块的初步结果和来自后处理模块的特定领域生语料, 均由存储模块持久化到数据库.

3.3 后处理模块

后处理模块主要负责清洗掉噪声数据, 由于部分页面含有大量 URL, 导致正文内容占页面总体内容比例较低, 解析正文时会受到影响, 可能导致得到的正文部分仅包括网站的版权声明等模板信息, 所以需要后处理模块进行清洗. 初步结果经后处理模块清洗后得到特定领域生语料.

3.4 特点

本文提出的基于语义相关度主题爬虫的语料采集方法具有以下特点:

(1) 在基于少量规则的页面分析基础上, 采用语义相关度主题爬虫技术, 有效提高了特定领域相关页面的识别能力, 降低了人工分析网页结构并制定规则的工作量.

(2) 利用网站结构中存在按分类进行页面存储的特点, 基于 URL 结构相关度对候选页面进行处理, 有效提高了页面采集的准确度.

4 关键技术

本章详细介绍词语相似度计算, 以及语义相关度主题爬虫两种关键技术. 词语相似度用以衡量待匹配词与主题词之间的相似性, 语义相关度主题爬虫用以预测未访问 URL 链指的页面是否主题相关. 其中, 待匹配词指描述信息和锚文本中的文本经分词、词性标注、依据词性过滤操作后得到的词语, 包含了文本的语法及语义层面特征, 可以有效表达页面内容.

4.1 词语相似度计算

词语相似度来度量待匹配词与主题词之间的相似程度, 借鉴文献[8]的定义, 词语相似度定义为: 描述词语与词语之间相似程度的一个数值, 取值范围在[0,1]之间. 一个词语与其自身的相似度为 1. 若两个词语在任何语境下都不可互换, 其相似度为 0.

本文利用 HowNet[9]结合词分布式表示计算词语间语义上的相似度. HowNet 是一部人工编纂的语义知识词典, 利用人工覆盖大多数常用词语, 准确性较好, 但由于专业特定领域通常具有时效性强的特点, 词语更新速度快, HowNet 不能完全覆盖, 而词的分布式表示可以由实时语料训练, 时效性较好. 因此, 本系统提出结合词分布式表示缓解 HowNet 覆盖度低的问题. 词语相似度计算步骤如下:

(1) 抽取 HowNet 包含的词语, 根据文献[8]提出的方法, 计算所有 HowNet 包含词与 HowNet 覆盖的主题词之间语义相似性 SimH.

(2) 对中文文本语料进行预处理, 包括语料抽取、繁简转换、分词、去除停用词等.

(3) 利用预处理后的语料训练词分布式表示.

(4) 根据式 (1), 计算 HowNet 包含的词与 HowNet 未覆盖的主题词之间、HowNet 未包含词与所有主题词之间的词分布式表示相似性 *SimD*.

$$SimD(d,t)=\frac{\vec{d}\cdot\vec{t}}{\|d\|\|t\|}=\frac{\sum_{i=1}^{n}w_{di}w_{ti}}{\sqrt{\sum_{i=1}^{n}w_{di}^2}\sqrt{\sum_{i=1}^{n}w_{ti}^2}} \quad (1)$$

其中, $SimD(d,t)$表示待匹配词与主题词之间的分布式表示相似性, d 与 t 分别表示待匹配词向量、主题词向量w_{di}、w_{ti}分别表示待匹配词 d、主题词 t 的分布式表示中维度 i 的大小, n 表示词分布式表示的向量长度.

(5) 根据算法 1, 最终得到所有词与主题词的相似度 Sim.

算法 1. 计算待匹配词与主题词之间语义相似度

输入:
1. 待匹配词 d 与主题词 t 之间的分布式表示相似性$SimD(d,t)$;
2. 待匹配词 d 与主题词 t 之间依据 HowNet 计算出的语义相似性$SimH(d,t)$;
3. 参数α、β, 其中$0\leqslant\alpha\leqslant1$, $0\leqslant\beta\leqslant1$阈值 Th1, Th2.
输出: 待匹配词 d 与主题词 t 之间的语义相似性$Sim(d,t)$.

操作步骤:
1) If $SimD(d,t)<0$, then $Sim(d,t)=0$, return $Sim(d,t)$.
2) If $SimD(d,t)<Th1$, then $SimD(d,t)=0$.
3) If $SimH(d,t)<Th2$, then $SimH(d,t)=0$.
4) If $SimH(d,t)=0$, then $Sim(d,t)=\alpha SimD(d,t)$, else $Sim(d,t)=\beta SimH(d,t)+(1-\beta)SimD(d,t)$.
5) If $Sim(d,t)<Th1$, then $Sim(d,t)=0$.
6) Return $Sim(d,t)$.

4.2 语义相关度主题爬虫

特定领域语料采集的基础是语义相关度主题爬虫. 语义相关度主题爬虫使用未访问 URL 优先级 (简称优先级) 来预测未访问 URL 链指的页面与主题的相关程度. 计算优先级时使用页面信息相关度来度量页面描述信息、锚文本与主题的相关程度, 同时使用 URL 结构相关度来度量未访问 URL 与历史 URL 的相关程度, 综合两种相关程度得到未访问 URL 的优先级.

4.3 页面信息相关度

本文使用页面信息相关度来度量页面描述信息、锚文本与主题在语义上的相关程度. 在计算页面信息

相关度时, 使用词语权重来度量每个词的重要程度, 本文定义词语权重如式 (2):

$$w_i = \frac{tf_i * idf_i}{N} = \frac{f_i}{N} * \log\frac{N}{N_i} \tag{2}$$

其中, w_i表示词语 i 的权重, 词语包括待匹配词与主题词. tf_i、idf_i分别表示词语 i 在文档 d 中的词频与逆文档频率, f_i表示词语 i 在文档 d 中出现的频数, N表示文档数目, N_i表示包含词语 i 的文档数目.

主题词集合用于描述特定主题, 待匹配词集用于表示未访问页面的特征, 其定义如下:

定义 1. 选取的主题词集合定义为$TSet$, 包含所有的主题词$word_t$.

定义 2. 待匹配词集定义为$WSet$由待匹配词组成.

由算法 1 与式 (2) 得到词语相似性和词语权重后, 就可以计算页面信息相关度, 计算如式 (3) 所示.

$$Rel(TSet, WSet) = \frac{\sum_{d=1}^{M}\sum_{t=1}^{N} w_d w_t (Sim(word_d, word_t))^2}{M} \tag{3}$$

其中, $Rel(TSet, WSet)$表示主题词集和待匹配词集之间的页面信息相关度, w_d表示待匹配词$word_d$的权重, w_t表示主题词$word_t$的权重, 二者均由式 (2) 得到, $word_d \in WSet$, $word_t \in TSet$, M表示待匹配词集的大小, N表示主题词集的大小, 即$M = |WSet|, N = |TSet|$, $Sim(word_d, word_t)$表示$word_d$与$word_t$之间的词语相似度, 由算法 1 得到.

4.4 URL 结构相关度

本文使用结构相关度度量未访问的 URL 与已访问的历史 URL 结构上的相似程度, 如果未访问的 URL 与历史 URL 结构相似, 则该 URL 很可能也需要爬取. 结构相关度的计算利用式 (4)、算法 2、算法 3 完成.

Jaccard 系数定义为:

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{4}$$

其中, A、B 表示两个集合.

算法 2. 计算 URL 之间的结构相似性

输入:
1. URL1, URL2
2. 权重参数α、β
3. 阈值参数$threshold$
输出: 是否结构相似 $isHrefStructSame$

操作步骤:
(1) 将 URL1、URL2 分别分割成主机名部分 Host1、Host2, 路径名部分 Path1、Path2.
(2) 如果 URL1 的主机名与 URL2 的主机名之间存在包含关系, 则返回 $isHrefStructSame$ = true.
(3) 将 Host1、Host2 按'.'分割, 各划分成若干个部分, 对二者对应部分按照式 (4) 计算 Jaccard 系数, 得到各个部分的计算结果 $J_1, J_2, \cdots, J_n$. 如果二者的对应部分均包含"www"(即两个 URL 均以 www 开头), 则此对应部分 Jaccard 系数为计算结果的$1/d$, d为常数.
(4) 主机部分相似性: $hostSim=\sum_{i=1}^{n} w_i J_i$, 其中 n 取 Host1、Host2 分割后的最小数目, w_i 表示第 i 部分的权重, 且$w_1=a'$, $w_{i+1}=w_i-b, i>0$, a,b均为常数, J_i 由上一步得到.
(5) 将 Path1、Path2 按'/'分割, 各划分成若干个部分, 对二者对应部分按照式 (4) 计算 Jaccard 系数, 得到各个部分结果 $Jp_1, Jp_2, \cdots, Jp_m$.
(6) 路径部分相似性$pathSim=\sum_{j=1}^{m} w_j * Jp_j$, 其中 m 取 Path1、Path2 分割后的最小数目, w_j表示第 j 部分的权重, 且$w_1=a'$, $w_{j+1}=w_j/2, j>0$, a'为常数, Jp_j由上一步得到.
(7) 判断: $\alpha hostSim+\beta pathSim$ 是否大于$threshold$, 若是返回 $isHrefStructSame$ = true; 否则返回 $isHrefStructSame$ = false.

算法 3 利用 URL 结构信息判断待爬取链接是否需要爬取

输入:
1. 未访问的 URL
2. 历史 URL
3. 采样数目 n
输出: URL 结构相关度 struct

操作步骤:
1) 从历史 URL 中随机抽取 n 个 URL 作为抽样样本.
2) 计算未访问 URL 与每个抽样样本间的结构相似性, 利用算法 2.
3) 如果未访问 URL 与过半样本结构相似, 返回 struct = 1. 否则返回 struct = 0.

4.5 未访问 URL 优先级

是否爬取一个未访问的 URL 最终由未访问 URL 优先级 (简称优先级) 决定, 若优先级大于阈值, 则认为该 URL 需要爬取. 优先级由未访问 URL 所在页面的描述信息、URL 所在锚文本、URL 结构三个因素共同决定. 优先级计算公式如式 (5).

$$p(url) = \alpha Rel(f_{\text{info}}, TSet) + \beta Rel(a_{url}, TSet) + \lambda struct(url) \tag{5}$$

其中, $p(url)$表示优先级, $Rel(f_{\text{info}}, TSet)$表示主题词集$TSet$与未访问 URL 所在页面的描述信息$f_{\text{info}}$之间的页面信息相关度, $Rel(a_{url}, TSet)$表示未访问 URL 所在锚文本a_{url}与主题词集$TSet$之间的页面信息相关度, 二者由公式 (3) 计算, $struct(url)$表示未访问 URL 的结构相关度, 其计算过程如算法 3, 取值{0,1}, α、β、λ是平衡参数, $0 \leqslant \alpha \leqslant 1$, $0 \leqslant \beta \leqslant 1$, $\lambda > 0$. 考虑页面的实际构成

情况, 借助层次分析法, 确定平衡参数取值.

最终, 语义相关度主题爬虫对未访问 URL 的爬取策略为: 如果未访问 URL 优先级大于阈值, 则爬取该链接链指的页面, 否则不爬取.

5 实验结果与分析

本文选取党建领域作为特定领域, 选用维基百科中文数据库全库数据、某党建网站党建专题下 2017 年 9 月至 2018 年 2 月约 1 万篇党建新闻数据、搜狗文本分类语料库部分数据来训练词分布式表示. 维基百科中文数据库信息量大、内容范围广, 党建专题数据包含有大量主题相关的内容, 内容专业性强, 搜狗文本分类语料库包含多种类别, 对应采集过程中遇到的各种类别. 语料详细信息见表 1.

表 1 训练分布式表示所使用的语料

数据	数目 (条)	大小 (KB)
维基百科中文数据	5 759 538	817 639
党建专题数据	141 583	43 303
搜狗文本分类语料库	118 452	24 741

为确保主题词的专业性与准确性, 大部分主题词选自中共中央党校出版社出版的《党的建设词典》, 为保证时效性, 同时选取了百度百科党务知识类别下的部分词条, 共选出 1552 个词作为候选主题词, 涵盖时政、理论、作风等方面. 为了保证系统的执行速度, 结合党建语料的统计信息, 从候选主题词中精炼出 429 个主题词作为最终的主题词. 部分主题词及其权重如表 2.

表 2 部分主题词

主题词	权重	主题词	权重
马克思	0.1	三个代表	0.008 168 279
三严三实	0.036 556 866	三大法宝	0.005 081 246

词语权重基于党建专题数据、依据式 (2) 计算. 对于未在训练语料中出现的词语, 赋予平均值作为其权重. 由于少数权重较高的词语出现在非党建领域页面时会对采集造成影响, 所以对其权重进行人工调整. 例如, “弘扬中国共产党人历史担当精神”(党建新闻) 与“与中国争夺影响力? 俄媒论莫迪与普京私聊弦外之音”(国际新闻) 中均包含“中国”一词, 若“中国”的权重过大可能导致不属于党建语料的国际新闻被采集. 词语权重共覆盖 88 820 个词, 其中人工调整权重的词语共 38 个, 部分被人工调整的词语如表 3 所示, 部分词语权重如表 4 所示.

表 3 部分人工调整的词语

中国	我国	国家	乡镇
项目	知识	公司	学生

表 4 部分词语权重

词	权重	词	权重
巡视	2.903 539 55	甘愿	0.015 283 01
党建	1.989 265 91	无意	0.015 283 01
政治	1.879 143 61	相伴	0.015 283 01

对于算法 1 中阈值选取, 以同义词组间、非同义词组间的语义相似性判断准确率作为评价标准, 选取 420 组词语进行实验, 实验结果如表 5. 最终根据实验结果进行阈值选取.

表 5 算法 1 的阈值选取

准确率	Th1=0.1	Th1=0.2	Th1=0.3
Th2=0.1	0.79	0.72	0.61
Th2=0.2	0.75	0.72	0.61
Th2=0.3	0.70	0.68	0.61

对于算法 2 中阈值选取, 以结构相似性判断准确率作为标准, 选取某网站相同板块以及不同板块下共 100 对 URL 进行实验, 实验结果如表 6, 最终算法 2 的阈值根据实验结果进行选取.

表 6 算法 2 的阈值选取

Threshold	准确率	Threshold	准确率
0.05	0.98	0.25	0.93
0.10	0.98	0.30	0.93
0.15	0.97	0.35	0.93
0.20	0.94	0.40	0.92

对于未访问 URL 链指页面阈值的选取, 即公式 5 阈值选取, 基于表 5、表 6 的实验结果, 利用层次分析法选取$\alpha\beta\gamma$, 并对不同的阈值进行测试, 实验结果如图 2. 阈值越大时, 主题相关限制越严格, 因此采集准确率越高. 当阈值达到 0.2 后, 准确率达到稳定. 此外, 当低阈值时, 由于主题相关限制不严格, 大量主题无关页面被访问, 大量时间耗费在无关页面的访问与判断上, 考虑时间因素下的采集量不高; 随着阈值升高, 主题限制严格, 无关页面被过滤, 无关页面访问耗时较少, 考虑时间因素下的采集量上升; 当阈值进一步升高时, 部分主题相关页面也会被过滤, 导致采集量级反而下降.

最后本文针对党建领域，对多个网站进行实验，实验网站列表如表 7 所示。网站分为两大类，前 4 个

网站为主流的门户网站，包含体育、经济、娱乐、汽车等大量无关领域信息，后 4 个网站属于党建领域网站，其所含内容属于党建领域信息。

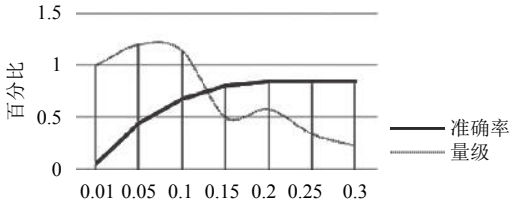

图 2 URL 最终阈值选取

表 7 实验网站

网站	名称	网站	名称
www.163.com	网易网	www.dangjian.cn	党建网
www.sina.com	新浪网	www.djyj.cn	党建研究网
www.sohu.com	搜狐网	www.zgg.org.cn	紫光阁
www.qq.com	腾讯网	www.zzdjw.org.cn	中直党建网

本文采用准确率作为评价标准, 准确率即精度, 如式 (6) 所示:

$$precision = relevantResult/AllResult \quad (6)$$

其中, *relevant*Result 表示系统采集的语料中属于党建领域的数量, *Allresult* 表示系统采集到的所有语料数目. 系统准确率结果如表 8 所示. 系统在党建领域网站, 其平均准确率在 94.87%; 在门户网站上, 其平均准确率为 64.20%.

表 8 实验结果

网站	语料 (条)	相关数目 (条)	准确率 (%)
网易网	4656	3344	71.82
新浪网	3789	2649	69.91
搜狐网	3899	2813	72.15
腾讯网	5832	2863	49.09
党建网	28 681	26 948	93.96
党建研究网	1108	1103	99.55
紫光阁	1303	1238	95.01
中直党建网	8074	7866	97.42

在党建领域网站上采集准确率较高, 原因是其内容属于党建领域信息, 噪声较小, 页面内容较为规整, 形式较为统一, 因此准确率较高. 而门户网站的采集准确率均较党建领域网站准确率低, 其主要原因如下:

(1) 门户网站包括大量无关主题, 噪声较大.

(2) 门户网站页面结构较为复杂, 对主题识别影响较大.

(3) 部分时政新闻正文文本较短, 导致正文解析器解析结果较差.

(4) 部分 URL 链接到视频、图片等非文本页面, 经初步过滤后只保留下网站版权等噪声信息, 降低了准确率.

6 结论与展望

本文针对语料采集任务, 设计了基于语义相关度主题爬虫的语料采集方法. 根据选定的主题词, 利用页面描述信息, 基于维基百科中文语料训练出的词分布式表示综合 HowNet 计算页面信息相关度, 结合 URL 的结构信息预测未访问 URL 链指的页面内容与主题的相关程度. 实验表明, 基于语义相关度主题爬虫的语料采集系统能够有效的采集互联网中的党建领域页面内容作为党建领域生语料, 具有较高的准确率. 针对正文提取器提取短文本时质量较差的问题, 下一步工作将对其进行改进, 以进一步提高系统采集门户网站时的准确率.

参考文献

1 Hersovici M, Jacovi M, Maarek YS, *et al*. The shark-search algorithm. An application: Tailored Web site mapping. Proceedings of the 7th International Conference on World Wide Web. Brisbane, Australia. 1998. 317–326.

2 苏祺, 项锟, 孙斌. 基于链接聚类的 Shark-Search 算法. 第四届全国搜索引擎和网上信息挖掘学术研讨会论文集. 济南. 2006.

3 黄仁, 王良伟. 基于主题相关概念和网页分块的主题爬虫研究. 计算机应用研究, 2013, 30(8): 2377–2380, 2409. [doi: 10.3969/j.issn.1001-3695.2013.08.034]

4 魏勇, 胡丹露, 郝晨光, 等. 基于分类关键词词频模型的地缘政治主题爬虫设计. 计算机工程, 2016, 42(2): 45–50. [doi: 10.3969/j.issn.1000-3428.2016.02.008]

5 Almpanidis G, Kotropoulos C, Pitas I. Combining text and link analysis for focused crawling—An application for vertical search engines. Information Systems, 2007, 32(6): 886–908. [doi: 10.1016/j.is.2006.09.004]

6 Du YJ, Liu WJ, Lv XJ, *et al*. An improved focused crawler based on semantic similarity vector space model. Applied Soft Computing, 2015, 36: 392–407. [doi: 10.1016/j.asoc.2015.07.026]

7 Suebchua T, Manaskasemsak B, Rungsawang A, *et al*. Efficient topical focused crawling through neighborhood feature. New Generation Computing, 2018, 36(2): 95–118. [doi: 10.1007/s00354-017-0029-8]

8 刘群, 李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学, 2002, 7(2): 59–76.

9 董振东. HowNet. http://www.keenage.com, 2013.