面向精准医疗的信息安全标准资源获取研究

赵 婧 邢玉艳 刘 耀
（中国科学技术信息研究所，北京 100038）

摘要：面向精准医疗的信息安全标准资源来源广泛，具有语义丰富、结构复杂、噪声数据多等特点。为了稳定、有效地对这些资源进行持续获取，并在数据采集时即过滤掉部分噪声数据，本文构建了一体化爬虫，利用自动生成的轻型知识结构库，将开放获取的领域专家知识融入资源获取部件，简化语义爬虫的工程复杂度，实现资源判别、获取、存储的一体化与语义结构生成与进化的一体化。实验结果表明，这种方法对信息安全标准体系资源的自动获取是稳定、高效的。

关键词：信息安全标准资源；一体化语义爬虫；资源自动获取；精准医疗信息；语义结构自动构建
中图分类号：G35 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2020.01.012

Research on the Information Security Standard Resources Acquisition for Precise Medical

ZHAO Jing, XING Yuyan, LIU Yao
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Information security standards for precision medicine come from a wide range of sources, with rich semantics, complex structures, and many noise data. In order to continuously and effectively acquire these resources and filter out some noise data during data collection, this article builds an integrated crawler, and integrates open-access domain expert knowledge into resource acquisition through an automatically generated light-weight knowledge structure library, simplifys the engineering complexity of semantic crawlers, and realizes the integration of resource discrimination, acquisition, storage and integration of semantic structure generation and evolution. The experimental results show that this method is stable and efficient for the automatic acquisition of information security standard system resources.

Keywords: information security standardization resources, integrated semantic crawler, automatic resource acquisition, precise health care information, semantic structure construction

0 引言

精准医疗是个体化的精准治疗模式。精准医疗的应用有助于患者获益最大化和医疗资源高效益[1]。信息标准体系是建设信息安全保障体系的重要条件，是信息系统健壮发展的保障[2-3]。而

作者简介：赵婧（1987—），女，中国科学技术信息研究所助理研究员，主要研究方向：信息资源管理；**邢玉艳**（1991—），中国科学技术信息研究所硕士研究生，研究方向：知识工程与知识发现；**刘耀**（1972—），男，中国科学技术信息研究所研究员，研究方向：知识工程与知识发现（通信作者）。

基金项目：国家重点研发项目“精准医疗伦理、政策法规框架研究”之课题 1“构建安全、可靠的面向生物医学大数据的、跨系统样本和数据共享的保障体系”（2017YFC0910101）。
收稿时间：2019 年 9 月 23 日。

面向精准医疗的信息安全标准是医疗信息资源系统在设计、研发、生产、建设、使用、测评中确保资源的一致性、可靠性、可控性、先进性和符合性的技术规范和技术依据[4]。在自动化与信息化的背景下，在构建精准医疗信息安全标准时，如何对标准资源数据进行自动获取、自动解析并入库，已引起业界学者的广泛关注。为了提高资源获取的效率，解决在多资源融合中存在的资源语义复杂、噪声较多的问题，本文将重点研究探讨一体化爬虫技术，通过自动语义结构生成，对网页内容进行过滤，为自动构建特定领域的信息安全标准提供参考与借鉴。

1　面向精准医疗的信息安全标准资源特点

研究相关资源可以总结得到如下几点精准医疗领域与信息安全领域文本资源的特点。

一是信息安全标准资源类型繁多。针对医疗领域，获取信息安全标准资源不只局限于信息安全标准，而应涉及与该领域相关的文本资源，如与其相关的论文、专利等。通过广泛收集多领域的信息安全标准资源进行自动化知识挖掘，为精准医疗领域提供参考。

二是信息安全标准资源多源化。主要体现为信息安全标准资源来源的渠道较多。除了已经颁布的信息安全标准外，还包括待审批的信息安全标准、学术论文数据库中对信息安全标准的相关研究论文、全国信息安全标准化技术委员会以及领域专家对标准制定的讨论等。

三是数据结构多样化。面向精准医疗的信息安全标准资源的多源化、数据结构的多样化主要体现在资源的种类较多，除了常规的文章外，还有视频、图片、音频中包含的相关文本数据。

四是语义复杂且噪声较多。在海量的文本资源中，讨论的话题十分广泛，比如仅精准医疗与信息安全标准就涉及大量不同的子领域，因此关于精准医疗与信息安全标准的语义十分丰富，同时也存在较多的噪声数据。

五是开放获取资源。面向精准医疗的信息安全标准资源已经发布在互联网上，是开放获取的，可以通过浏览器直接进行访问。基于领域资源开放化的特点，可以依托平台构建相关功能模块，对资源进行自动获取。

2　自动获取资源的通常技术

自动获取资源的关键技术是网络爬虫。网络爬虫可以自动地从互联网中获取指定主题的内容，是互联网资源获取的重要工具。网络爬虫最初是为了实现搜索引擎的功能而发明的。爬虫可以自动获取或更新网站的内容和检索方式，采集其能够请求到的所有页面内容，以供搜索引擎做进一步处理，从而让用户能够对其所需的资源进行快速检索。

（1）普通爬虫。一般地，普通爬虫能够对指定的URL发出请求，抓取其返回的Web资源，根据解析出的URL进一步爬取资源，从而实现对网络资源的获取。如Lee Hsin-Tsang等[5]开发和实现了IRLBOT，处理了遍历网页时产生的海量URL的瓶颈问题以及反垃圾问题，极大地提升了爬虫在单服务器上的性能，具有较强的扩展性；曾锡山等[6]采用多断点控制、并发采集的方法，构建了基于Web文本海量数据挖掘的应用统计系统；于留宝等[7]提出了基于MapReduce的数据采集方案，将爬虫系统部署在hadoop平台上，并采用多个小文件的输入方式，解决了负载均衡的问题。普通爬虫的问题是在待获取的资源规模较大、待抓取的网页较多时，其内容趋于复杂、多样化，爬虫难以高效率地获取与所需资源相关度较高的资源，从而降低了爬虫的效率。

（2）主题爬虫。主题爬虫[8]是通过预先给定一组与主题相关的URL集合，对所需的与主题相关的资源进行爬取；对于从相关资源中解析得到的URL，根据一定的算法，过滤掉与主题相关度不高的URL，将与主题相关的URL加入待爬取队列。如De Bra等[9]提出的fish search算法，该算法通过用户提供的关键词，判断网页遍历过程中提取URL的相关性，在相关性小于阈值时，停止这个方向的遍历；Chakrabari等[10]基于朴素贝叶斯分类模型提出能够判断网页主题相关性URL

访问次序的主题爬虫；汪涛等[11]提出了基于主题相关度分析的主题爬虫，其核心是主题确立模块，结合领域专家提供的词表与自动抽取出的领域相关词，采用向量空间模型进行主题相似度分析，并采用PageRank方法进行URL排序。主题爬虫对爬取内容的过滤方法主要有基于文字内容评价与基于链接评价。这两种主题评价策略在计算主题相关度时均以关键词作为衡量标准，在关键词语义较宽泛时可能会产生大量噪声数据，在关键词不足时可能会漏掉某些关键页面[12]。

（3）语义爬虫。语义爬虫具有一定的语义推理能力，在爬取资源时可以通过对主题进行语义相似度计算等指导爬行，从而获得了一定的决策能力，通过改进语义结构实现爬虫的进化。Ehrig等[13]提出基于复杂本体的文档爬取模型，通过定义本体以及语义相关度计算方法发现相关文档；Su等[14]基于本体计算页面相关度，在爬行过程中对本体结构进行改进，实现智能化爬虫；黄炜等[15]通过本体实现电子商务领域的特征表达，实现语义爬虫，并通过统计与规则方法在爬行中对本体进行改进。总体来讲，基于本体工程与本体学习的语义爬虫在爬行中有一定的语义推断能力，相比传统爬虫有较好的改进。传统的语义爬虫所使用的本体一般由手工构建，存在难度大、成本高、效率低、自适应性差等缺陷，而基于本体工程与本体学习的爬虫，其工程一般较为复杂，本体学习的规则定义较为困难，且爬虫的效率极大依赖于本体的结构以及本体学习的算法，构建本体的复杂度甚至超过普通爬虫加上数据清洗的难度，难以真正融合本体与爬虫。

3　一体化语义爬虫的构建

鉴于前述，为克服目前网络爬虫的缺陷，本文提出基于轻型知识库构建一体化语义爬虫，为自动化获取信息安全标准体系资源提供一个稳定而高效的办法。从工程的角度，构建一体化语义爬虫的语义框架时应结合资源判别、资源获取、资源清理与获取能力提升，实现语义框架的自动进化。对于一体化语义爬虫，能够根据用户输入的关键词，自动构建一个基础语义结构，对关键词进行语义推理，并根据下载的网页不断进化结构，以便更好地抓取网页资源，提高爬虫的效率。与传统的语义爬虫相比，本文构建的轻型知识结构所需资源是开放获取的，能够降低工程的复杂度，实现自动化构建。同时，构建的轻型知识库基于相关领域专家的先验知识，对爬虫有较好的指导性，极大地提高了爬虫的效率。该爬虫框架如图1所示。

3.1　语义框架自动生成

一体化语义爬虫通过对用户所需的领域主题进行语义推理，扩充语义结构，并将所爬取的文本赋予语义信息，从而实现爬取过程中的决策与数据过滤。在用户输入领域主题词后，爬虫根据平台资源以及开放获取的Web资源，构建轻型知

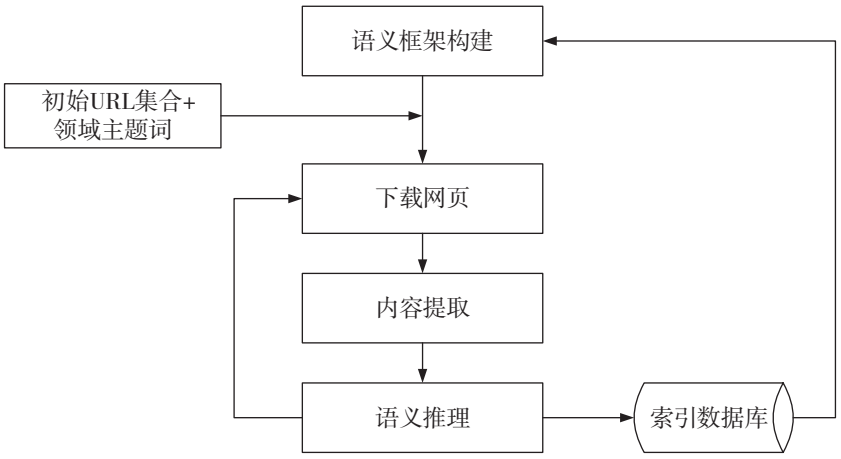

图1　一体化语义爬虫框架

识库。轻型知识库不以构建严格的知识结构为目标，而是利用定义词及属性的关系获取词的语义知识，所定义的结构如图2所示。

在图2中，概念词表达一个具体领域内的主题或概念，是该领域内语义信息的来源，概念词之间具有上下位关系，形成树形结构；概念词通过同义词进行扩展；每个概念词都有其属性，并且通过属性域形成“概念—属性—属性词”的三元组关系。初始语义结构生成所依据的资源主要有3种：结构化资源、半结构化资源以及非结构化资源。表1对这些资源进行了描述。其中，除汉语医学主题词表外，所有的资源均可以在互联网上进行开放获取。

3.1.1　结构化数据

结构化数据包括MeSH医学主题词表与汉语主题词表。汉语主题词表[16]是由领域内的专业人士制定的，规定了术语的层级结构。MeSH医学主题词表[17]由美国国立医学图书馆建立，收录了上万条医学领域概念。通过主题词表提供的领域内专家知识，可以为爬虫构建初始的语义结构。首先用户规定所查询的领域及主题词，选取《哈工大信息检索研究室同义词词林扩展版》[18]对主题词进行扩充，扩充后的词集合在主题词表内进行检索，用匹配到的主题词及其层级结构作为初始的语义结构。以关键词“妊娠性高血压”为例，初始语义结构如图3所示。

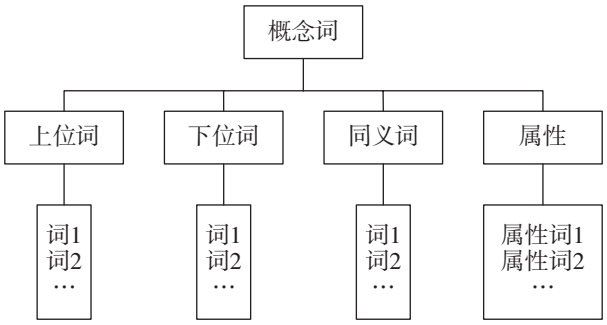

图2　话题轻知识结构

3.1.2　半结构化数据

半结构化数据包括组织良好的文本数据。这种类型的数据包括书籍目录、百科、论文等定义好的层级结构和文本数据。这些数据一般为领域内较为权威的数据，具有明显的结构化特征。标题中一般包含主题词及属性的共现，对应的下级结构则包含属性词（属性值域）。通过分析数据的标题以及文本内容，可以提取概念对应的属性及属性词。

本文所构建的一体化语义爬虫，其概念词抽取模块主要通过条件随机场（Conditional Random Field，CRF）与点间互信息+邻接熵实现。CRF[19]是一类基于马尔可夫假设的无向图模型，通常应用于序列标注中，在数据量不大的情况下可以产生较好的效果。依托平台资源获取医疗领域手工标注的文本数据约6000句，标注方式采用“BIO”标注，作为CRF模型的训练数据集；采用CRF+工具[19]对模型进行训练，得到的模型用于抽取资源中的概念词。作为CRF方法的补充，互信息+邻接熵的方法可以在正文文本中抽取概念词较少的情况下进行补充。这是一种无监督的方法，通过判断互信息（即词间的凝结度）与邻接熵（即词两边的信息丰富度）提取概念词。采用这种方法提取概念词时，首先使用搜狗[20]的医学领域词表，对文本进行分词；使用THULAC[21]对分词后的文本进行词性标注，只保

表1　资源描述

资源类型	资源名称
结构化资源	MeSH医学主题词表；汉语主题词表
半结构化资源	书籍目录，如当当网（www.dangdang.com）；领域相关论文数据（万方数据）；百科数据，如百度、维基百科等
非结构化资源	垂直领域论坛，如丁香园（www.dxy.cn）、信安委官网（www.tc260.org.cn）；博客等

{妊娠性高血压’：{上位词：[[’妊娠并发症’,’高血压’]],’下位词’：[’先兆子痫’,’子痫,’HLLP综合征’],’入口词：[’高血压妊娠性’,’妊娠高血压’,’妊高症]}}

图3　初始语义结构示例

留名词；提取文本中TF−IDF值较高的词，再以这些词为中心，计算互信息与邻接熵进行词串扩展，直至达到阈值，提取出概念词。

3.1.3　非结构化数据

非结构化数据则包括没有明确结构定义的文本数据。这类数据来源于与领域直接相关的垂直类网络资源，例如医疗领域的丁香园、负责信息安全标准制定的信安委官网等。在这类数据中，通过实体概念抽取算法抽取概念词。因为非结构化数据没有结构定义，从中抽取的概念词无法像前述数据一样判断其层级，因此通过词共现及基于向量空间模型的词相似度判断其层级，将其划归到概念词的同义词或属性词中。

结合结构化数据、半结构化数据、非结构化数据3种资源，则可以为一体化语义爬虫生成初始语义结构，其过程如图4所示。

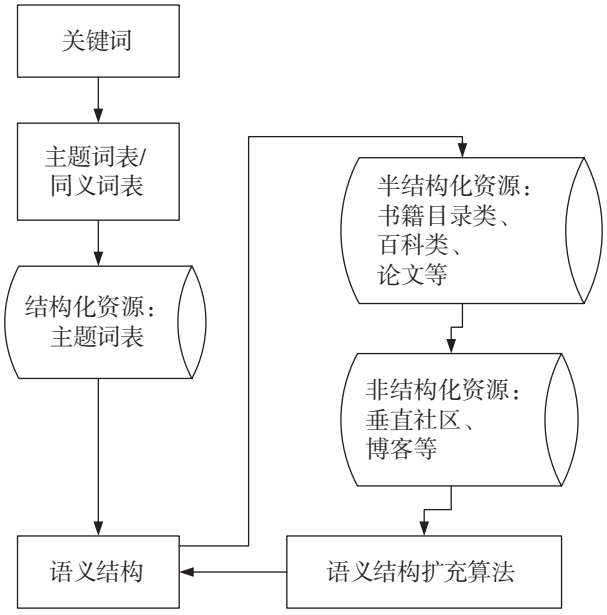

图4　语义框架自动生成

在爬虫爬行过程中，通过语义框架，为非结构的文本赋予语义信息进行语义推理，则可以指导爬虫对资源的获取。此外，所获取的文本在进入索引数据库之后，可以再运行上述语义结构扩充算法，使爬虫的语义结构进一步进化，提高爬虫的运行效率。图5是以“医学化合物”作为关键词时生成的部分结构。

3.2　语义推理过程

通过上述爬虫自动构建的语义框架，一体化资源获取部件对标题文本、正文文本、文本属性词等文本资源进行语义标注。通过语义标注，爬虫可以为待获取的资源赋予结构化的语义信息，从而计算资源与主题的相关度，对所获取资源实现先验判断，有效地利用领域内开放获取的专家知识对资源进行自动化获取。语义标注是基于工程构建的轻型知识库，对待获取的文本资源进行概念与属性的匹配，计算其最终的语义得分。

一体化语义爬虫可以实现语义框架的自动构建。在数据采集过程中，一体化语义爬虫可以对资源进行语义推理，并随着所获取的资源不断增加，可以自动改进语义结构，对所爬取内容构建自适应的语义框架，在爬取过程中不断提高效率。图6是实现的资源获取与自动存储部件的工作流程。

对于获取到的文本资源，将其进行进一步清洗，并且自动化解析入库。对于非结构化的数据，如视频、图片等，将其地址映射到数据库中，采用图像、音频识别等多媒体资源自动化解析方式，从中提取结构化数据，并且解析入库。基于上述流程，针对精准医疗的信息安全标准资源的特点，使用工程化的方法，解决了数据多源

属性同义词：旋光度=[]，公式计算=[过程，计算，求解，达到，表达式，等于，定律，常里，方程，恒里]，腐蚀性=[]，生物碱=[semantciStruc.WordBean@2c767a52，semantciStruc.WordBean@619713e5，semantciStruc.WordBean@708f5957，semantciStruc.W
[糖类]
[性质，物化性质，物理性质，化学性质]
[腐蚀性，脱水性，强氧化性，分解，氧化，氯化，催化，解聚，还原，缩合，聚合，腐蚀，水解，加成，酸性，碱性，显色反应]
[葡萄糖]
[性质，物化性质，物理性质，化学性质]
[腐蚀性，脱水性，强氧化性，分解，氧化，氯催化，解聚，还原，缩合，聚合，腐蚀，水解，加成，酸性，碱性，显色反应]

图5　语义结构示例

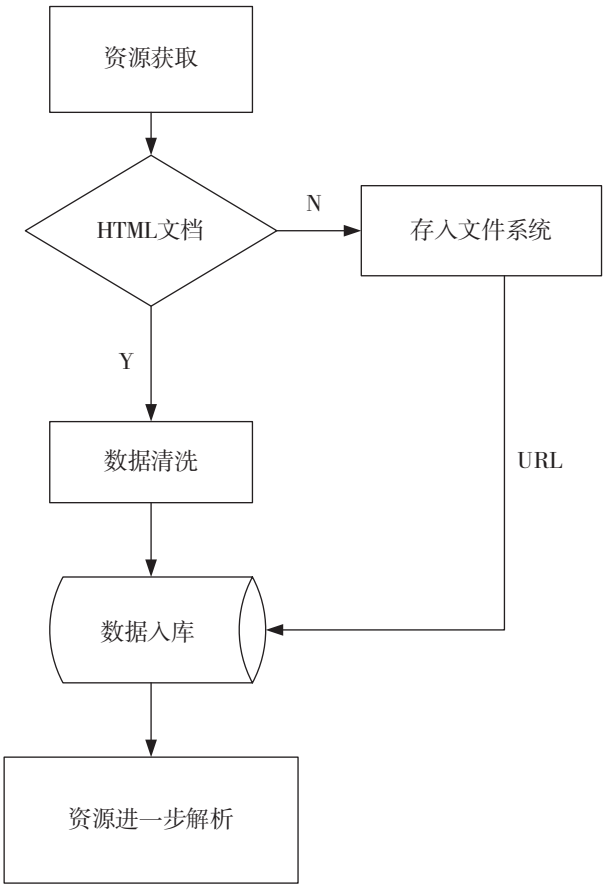

图 6　资源自动获取流程

化、数据结构多样化、噪声多的问题，从而实现了对所需资源的高效获取与分析。

4　实验结果分析

针对多来源收集的精准医疗资源与信息安全标准资源的特点，对面相精准医疗的信息安全标准资源的评估平台所采用的一体化语义爬虫的性能进行实验。这次实验主要是将一体化语义爬虫与基准爬虫（以广度优先策略构建的普通爬虫）进行对比，评估一体化语义爬虫获取数据噪声的过滤效果以及对语义相关度较高资源的获取效率。

选取线程数、爬取时间、机器配置作为控制变量，抓取网页的精确率P（Precision）作为抓取有效性的评价指标，公式采取

$$P=\frac{TP}{TP+FP} \quad (1)$$

在式（1）中，*TP*是True Positive，代表爬取到的文档为主题相关文档；*FP*是False Positive，代表爬取到的文档为主题不相关文档。式（1）的含义为：

$$Efficiency=\frac{\left|\{relevant\ documents\}\cap\{retrieved\ documents\}\right|}{\left|\{retrieved\ document\}\right|} \quad (2)$$

选取标准、政策、论文等与精准医疗相关的开放Web文本资源，设定对多个关键词进行爬取，并对每次所得出的结果进行平均，最终获取资源的结果。对比结果如表2所示。

从表2对比结果中可以看到，一体化语义爬虫资源获取方法是稳定、有效的。与基准爬虫相比，一体化语义爬虫在抓取效率上显著提升，且对有效资源的获取速度也显著提升。分析效率提升的原因主要是：一体化语义爬虫能够自动生成可进化的语义框架，针对资源的语义特点，能够过滤掉大部分的噪声，选择语义相关度较高的内容，同时一体化语义爬虫具有较高的可扩展性，针对多数据结构的特点，将非结构化的数据映射到数据库中，提高了资源获取的广度。

5　结语

多来源的信息安全标准资源为文本挖掘与自动化提供了合适的语料，也为精准医疗领域的信息安全标准提供了重要参考，通过对这些资源

表 2　一体化语义爬虫与基准爬虫的对比结果

实验指标	一体化语义爬虫	基准爬虫
抓取文档数量/个	1241	1469
有效文档数量/个	928	356
抓取速度/（URL/s）	4.1	4.9
有效率/%	74.8	24.2
有效资源获取速度（URL/s）	3.0	1.2

的获取与进一步加工，能够自动生成新领域的信息安全标准。本文依托信息资源加工平台，对面向精准医疗标准资源的特点与获取方式进行了研究，提出了一体化语义爬虫，降低了语义爬虫的工程复杂度，并对领域资源进行获取。实验验证，构建的一体化语义爬虫对于资源获取部件是具有有效性的，为后续的资源加工与标准生成提供了基础。但是，本文提出的一体化语义爬虫在语义结构生成过程中仍采取单机的数据采集与结构库构建方式。对于开放领域专家知识的获取，一体化语义爬虫仍需对种子资源进行筛选与指定，而对初始资源噪声的辨别能力不强。在现有结构的基础上，需对平台的资源获取部件进行改进，实现分布式的数据爬取与集群式的语义结构进化。同时还需加强爬虫对初始资源的语义理解能力，从而进一步实现资源获取的智能化，提升爬虫的效率。

参考文献

[1] 夏锋, 韦邦福.精准医疗的理念及其技术体系[J].医学与哲学(临床决策论坛版), 2010(11): 5−7, 21.

[2] 杨少华, 李再扬.信息产业技术标准化的理论分析框架及其政策含义[J].情报杂志, 2011(9): 93−99, 105.

[3] DAVID P A, GREENSTEIN S. The economics of compatibility standards: An introduction to recent research[J]. Economics of Innovation and New Technology, 1990(1/2): 3−41.

[4] 周鸣乐, 董火民, 李刚, 等.信息安全标准体系研究与分析[J].信息技术与标准化, 2008(4): 15−20.

[5] LEE H T, LEONARD D, WANG X, et al. IRLbot: Scaling to 6 billion pages and beyond[J]. ACM Transactions on the Web, 2008, 3(3): 8.

[6] 曾锡山, 胡俊荣.WEB文本海量数据挖掘应用中的多点数据采集及处理问题研究[J].情报杂志, 2010(8): 135−139.

[7] 于留宝, 胡长军, 苏林晗.基于MapReduce的微博文本采集平台[J].计算机科学, 2012, 39(z3): 143−145.

[8] CHAKRABARTI S, BERG M V D, DOM B. Focused crawling: A new approach to topic−specific web resource discovery [J]. Computer Networks, 1999, 31(11/16): 1623−1640.

[9] BRA P D, HOUBEN G J, KORNATZKY Y, et al. Information retrieval in distributed hypertexts[C]//Proceedings of the Intelligent Multimedia Information Retrieval Systems and Management. 1994.

[10] CHAKRABARTI S, DOM B, INDYK P.Enhanced hypertext categorization using hyperlinks [C]// Proceedings of the ACM SIGMOD Record, F, 1998.

[11] 汪涛, 樊孝忠.主题爬虫的设计与实现[J].计算机应用, 2004, 24(S1): 270−272.

[12] 刘金红, 陆余良.主题网络爬虫研究综述[J].计算机应用研究, 2007(10): 32−35, 53.

[13] EHRIG M, MAEDCHE A.Ontology−focused crawling of Web documents[C]//Proceedings of the 2003 ACM symposium on Applied computing. ACM, 2003.

[14] SU C, GAO Y, YANG J, et al. An efficient adaptive focused crawler based on ontology learning [C]//Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05). IEEE, 2005.

[15] 黄炜, 张李义.基于语义爬虫的商品信息主题采集研究[J].数据分析与知识发现, 2010, 26(1): 3−8.

[16] 中国科学技术信息研究所情报检索语言研究室.汉语主题词表[M].北京: 科学技术文献出版社, 1996.

[17] MeSH主题词表. MeSH主题词表[M/OL].[2019−07−10].https: //www.nlm.nih.gov/mesh/meshhome.html.

[18] 哈工大信息检索研究中心. 哈工大信息检索研究室同义词词林扩展版 [M/OL]. [2019−07−10]. http: //ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm.

[19] LAFFERTY J, MCCALLUM A, PEREIRA F C N J P O I. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [J]. Proc ICML, 2001, 3(2): 282−289.

[20] 搜狗.搜狗医学领域词表[M/OL]. [2019−07−10].https://pinyin.sogou.com/dict/.

[21] LI Z, SUN M.Punctuation as implicit annotations for Chinese word segmentation[J]. Computational Linguistics, 2009, 35(4): 505−512.