·信息技术与系统·

基于自然语言处理技术的定题监测功能实现研究*

刘 巍 王思丽 祝忠明 吴志强

(1.中国科学院兰州文献情报中心 甘肃兰州 730030)

摘 要:文章主要描述了在自动监测功能研发过程中,如何引入自然语言处理相关技术,从而提高开放知识资源自动监测采集过程的准确性、通用性、可配置性及松耦合性。研究发现,通过将自然语言处理技术应用在自动监测功能中,可以实现对监测资源中重要概念和实体的自动抽取,并与经过用户配置的语料库进行相似度匹配,最终基于匹配的结果实现自动化定题监测的目标。实践应用证明,文章提出的基于自然语言处理技术的定题监测方法目前已应用在相关项目的建设中且实测效果较好,证明其在一定程度上改进了传统的定源定向监测采集方法,提高了监测结果的准确性,优化和简化了监测参数的配置流程,有效提升了功能的通用性和松耦合性。

关键词:自然语言处理;实体抽取;相似度计算;定题监测;信息采集

中图分类号:TP312 **文献标识码**:A DOI:10.11968/tsyqb.1003-6938.2018057

Design and Implementation of Automatic Monitoring Function Based on Natural Language Processing Technology

Abstract This paper describes how to apply natural language processing technology in the development of automatic monitoring functions, improving the accuracy, versatility, configurability and loose coupling of the process of automatic monitoring and acquisition of open knowledge resources. The application of the natural language processing technology can extract important keywords and entities and similarity match with configuration item which configured by users. Finally, based on the matching results, system can determine whether the target is focused. so as to achieve the goal of automated monitoring. This method has been applied in the development of IIBD platform and has a positive effect. This study has improved the traditional fixed-source monitoring method. The accuracy of monitoring results was improved, and configuration of monitoring parameters were optimized and simplified, and versatility and loose coupling of functions were increased.

Key words natural language processing; entity extraction; similarity calculation; fixed-subject monitoring; information acquisition

大数据环境下，可开放获取的信息资源数量大幅提升,更新速度也不断加快,特别是面向产业的政策、市场、科研、数据、决策等多种类型的信息资源,由于其时效性强,覆盖范围广,且一定程度上具有较大的可信度(尤其是政府、权威机构等发布的),已成为政府、企业、科研机构及其情报研究人员关注的重点。因此,及时发现、分析、管理和利用这些开放信息资源,对于获得最新的情报信息,制定合理的科技战略决策,进行相关情报研究变得十分必要。

本研究主要针对产业情报大数据平台(Industrial Intelligence BigData,IIBD)[1] 建设和应用过程中的双向需求,实现对网络中产业相关政策、动态、数据、文献等开放信息资源进行多来源、自动化、定题监测和采集管理。对开放信息资源的监测,目前在应用方面使用较多的方法包括通过互操作协议或接口进行监测的方法。该类型监测方法由于具备信息发布平台提供的互操作接口，因此在监测精度方面有很好的可扩展性和可操作性，采集到的信息格式较完整且质量较高，但是目前直接提供公开接口的平台并不多,尤其是一些重要的竞争情报类站点、企业网站等

* 本文系中国科学院西部之光青年项目“基于学术大数据的专题化信息自动采集与组织技术研究”(项目编号:Y6AX021001)研究成果之一。
收稿日期:2018-06-11 责任编辑:魏志鹏 通讯作者:刘巍(liuw@llas.ac.cn)

并不提供相应的接口和协议，因而该方式并不具有普适性。此外,基于搜索引擎技术的定向定源监测是目前普遍使用的监测方法，但是该监测方法主要通过网络爬虫技术或工具对待采集的网页进行分析,然后进行采集。这种方法灵活性较好,不受目标站点的技术架构限制。一般来讲,只要能够浏览到的信息都可以监测和采集到。但从操作角度来看,方法需要经过相对复杂的配置才能具有较好的监测效果,且当信息来源站点结构发生变化时，需要即时发现并调整采集规则，这在一定程度上增加了操作难度和工作量,特别是当监测源数量较多时,需要有专人或专门的团队进行相应的跟踪和维护。

本研究在对网络开放信息监测方法的相关研究现状进行调研梳理的基础上,结合 IIBD 建设的具体需求，设计开发了基于自然语言处理技术的可配置化互联网开放信息资源的自动监测功能，着重研究和解决了非固定、多源异构情报源采集内容的自动识别和相似匹配的问题,并在 IIBD 平台中进行应用研究,最终实现了对多源异构监测信息的智能识别、长期监测和自动采集发布，并且该功能方法相对于传统自动监测方法来说，在通用性和可配置性方面有所优化和提升。

1　研究综述

竞争情报监测与传统搜索引擎系统所关注的目标和实现方法均有所不同，竞争情报更注重情报获取的精准度和及时性，且一般都在特定领域或主题开展，因此更适合使用定主题的信息采集方法。自 2000 年以来,国内外的信息采集技术逐渐发展成熟,并在相关领域开展了广泛研究和深入应用，所涉及到的相关技术一般包含采集规则/算法/模型的构建、主题内容信息的自动识别和抽取、网页文本的自动聚类与分类技术等。

1.1　监测采集技术研究现状

(1)基于模板匹配的采集技术研究。Bar-Yossef Z 等将在同一网站内多次重复出现的网页头部、导航栏、版权声明、广告等信息块视为噪音信息并定制为匹配模板,并与待处理的网页 DOM 树进行匹配并删除,最后剩下的为主体信息[2]。该类方法属于基于模板匹配的采集技术，应用该方法的前提是同一信息源的内容页面应具有相同或近似的基础展示模板,通过创建和识别模板,然后基于对模板的识别结果进行主体信息内容抽取和采集。

(2)基于 URL 分类的采集技术研究。叶勤勇提出 UFBC 学习算法，基于开源搜索引擎 Nutch 和利用正则表达式进行信息识别和监测采集[3];蒋付彬提出基于决策树的 URL 分类器算法，利用 4 个主要 HTML 标签内容与用户定义主题的相似度构建决策树实现 URL 分类[4];杨镒铭提出基于模式树的 UPCA 分类算法，通过训练提取特定类型的网页链接特征,构建模式树和生成模式规则，形成主题相关的 URL 模式库[5]。该类监测采集方法属于基于 URL 规则的监测采集技术,其应用前提是认为同一来源站点创建的动态网页其内容一般应属于同一个主题且其 URL 格式往往非常相似,基于这一思路,该方法通过各种算法和模型去实现对基础 URL 规律的量化、补充计算,以区分主题无关的 URL 和主题相关的 URL。

(3)基于机器学习的采集技术研究。近年来,对信息监测采集技术的研究方向开始向基于机器学习的方法转换。如 Debnath S 等利用预定义的标签集合对 DOM 树节点进行训练生成分类器[6];王浩提出将采样技术和半监督学习相结合的方法，对传统的 SMOTE 文本分类算法进行改进以实现网络敏感信息的识别[7];Pavlinek M 等提出了基于主题模型表示的半监督式文本分类方法，该方法包括一个基于自训练的半监督文本分类算法和模型，用于识别和确定新文本内容的参数设置[8]。该类方法大多采用需要监督或半监督的机器学习算法，需要基于大量样本积累和训练，或由人工预先标注好一定数量的样本实例，并进行聚类、归纳学习并生成网页分类器(算法和规则)，利用分类器对网页信息进行模式处理。此外,基于内容结构特征和视觉特征,采用相关启发式算法如神经网络算法、贪心算法等构建启发式规则集合，将网页划分为多个可视化块的相关集合以实现内容信息提取等方法的研究也越来越多。如李剑基于 BP 神经网络算法改进 DOM 树结构,按内容相关性将网页划分为多个子模块进行信息内容过滤提取[9];李伟男等基于模拟退火算法训练二阶

隐马尔科夫参数，改进经典的 VIPS 网页分块算法以实现网页主题信息抽取[10]；谢方立提出了基于 DOM 节点类型标注的 NTA 主题信息抽取算法[11]。

1.2 监测采集工具研发研究现状

在实际应用中一般需要根据具体应用需求和不同数据源的结构对上述技术方法进行取舍、改进或整合、综合利用等。同时，在实际开发中，一般会将上述方法与网络搜索引擎和爬虫框架如 Nutch、Heritrix、Scrapy 等进行结合，通过改进监测过程中的某一流程达到提升监测速度或精准度等目标。如谭宗颖等基于网络爬虫技术和文本聚类技术构建了科技发展前沿信息监测与分析平台[12]；刘海波基于 Ajax 和 Web Service 技术实现了网站多栏目多频道的信息监测和实时入库[13]；张智雄等构建了一种支持按需申请、定制服务的科技战略监测服务云平台，通过将网络自由文本转化为结构化的可计算的知识单元，实现对科技领域的态势监测[14]；谢靖等以开源爬虫 Crawler4j 为基本框架，实现了面向网络科技监测的分布式定向资源精确采集[15]；王思丽等也对开放资源及其元数据自动采集策略方法进行了相关实验研究[16-17]。

2 基于自然语言处理技术的定题监测关键功能设计与实现

本研究所提出的基于自然语言处理技术的定题监测，其方法正是在本团队成员王思丽已提出的采集策略和方法基础上进行了深度改进和优化，通过引入自然语言处理技术，实现对文本中关键概念、实体等的自动抽取并与用户提供的主题、实体等语料进行相似度匹配，从而达到提升监测采集精准度的目标。同时，通过引入成熟的自然语言处理工具和框架，可以在主题、实体概念的抽取过程中自动实现新词发现和语料库的扩展，在一定程度上实现了冷启动的目标，使本方法可以不受半监督学习方法需要标注或准备大量学习样本弊端的影响，提升的采集过程的自动化程度，以及系统整合层面的松耦合性。在配置和操作方面，由于不强制要求提前定义模板或 URL 规则，只需要提供用户关注的主题和实体，也在一定程度上降低了操作复杂度，即时，没有相关技术背景和使用经验的用户也可以配置操作。

2.1 整体功能结构

本研究所提出的基于自然语言处理技术的定题监测整体功能结构主要包含数据准备、采集参数配置、核心概念及实体抽取、相似度匹配及采集发布五个步骤（框架见图 1）。

2.2 数据准备

基于自然语言处理技术的定题监测数据准备步骤主要用于确定基础情报源集合即待采集情报源的基础信息和启动信息。主要包括情报源的名称、网

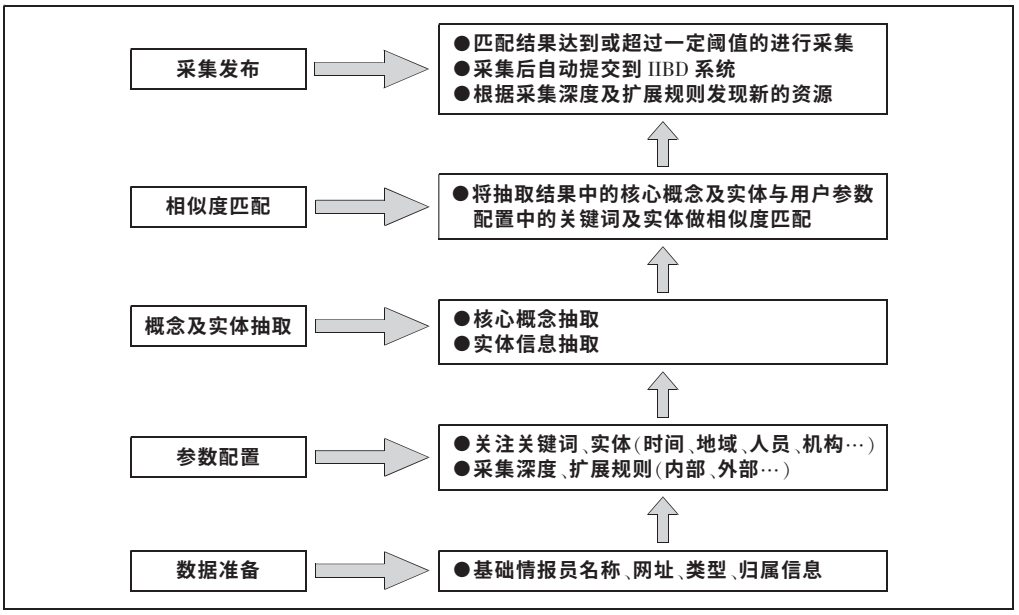

图 1 基于自然语言处理技术的定题监测整体功能框架图

址、类型及归属等基础信息,该步骤一般由具有较丰富相关情报遴选经验的人员或根据用户具体需求进行梳理。所有情报源构成了采集和扩展的基础。

2.3　参数配置

参数配置主要分为两部分，首先是匹配或识别参数配置,主要包括重点关注的关键词、实体(时间、地域、人员、机构、国家等)概念的集合,也可以是相关概念的逻辑组合,如时间 AND (人员 OR 机构 OR 国家) AND 关键词,表示重点关注某一时间范围内,某个人员或机构或国家与某关键词同时出现的信息;另一部分是采集参数配置,主要包括采集深度配置和扩展参数配置，采集深度配置即基于深度优先的原理执行如下操作的次数：①从基础情报源中取出一条信息并对其进行解析；②把解析出的链接和已监测表中的链接进行比较，若已监测表中不存在此链接,表示其未被访问过;③把链接放入监测解析流程中;④处理完毕后,将其放入已监测表中;⑤将当前信息作为基础情报源重复执行①。

扩展参数配置主要控制采集深度的处理策略,如当设置扩展参数为内部时,则新发现的 URL 与基础 URL 相似或处于同一情报源时才进行解析处理否则舍弃。当扩展参数设置为外部时,则无论新发现的 URL 是否与当前情报源处于同一来源均进行解析和分析。

2.4　概念及实体抽取

对采集到的主体内容进行概念和实体抽取需要借助一些第三方自然语言处理工具,在本研究中,对中文的概念及实体抽取我们选用的是 Ansj，对英文内容的概念及实体抽取我们选用的是 Stanford-CoreNLP，以上两个自然语言处理工具包均为开源(具体抽取过程见图 2)。

(1) 预处理。当从情报源监测一个网页信息时,首先利用模板匹配法抽取网页主体信息，并判断信息的语种等，同时去除主体信息中的停用词 (如介词)等,完成对原始信息的预处理。

(2)分词及词性标注。基于主体信息及主体信息语种选择不同的自然语言处理工具进行分词和词性标注。本研究在开发过程中测试了大量开源自然语言处理工具,发现所有工具,特别是中文分词工具的分词粒度都存在过细的问题，普遍会出现分词过度的情况,如“元数据”一词,分词工具通常会基于更通用的语境,将“元”当作一个量词,将“数据”当作一个名词分开抽取和标注,且类似情况很常见。所以,不能直接调用抽取的结果，需要再次封装概念抽取模型和算法。

(3)分词组合。本研究的做法是:首先,根据通用的语境或领域,制订相对宽松的分词组合规则,如连续出现的名词或动词加名词等都可以认为是一个表达完整词义且有意义的概念。利用这一系列组合规则,对信息主体中的所有连续分词进行匹配,抽取出所有符合组合规则的词组。因为本研究制订和遵循的是较宽松的分词组合规则,因此,此时抽取出的词组通常粒度较粗且存在过度组合的情况，同样不能直接使用,需要再次清洗。

(4)关联合并。合并的思路主要依据关联规则挖掘的思想,对每个抽取出的词组再进行细粒度分词,这样就得到若干组候选项集。基于这些候选项集,使用 Apriori 算法或信息熵算法可以挖掘出具有强关联规则的若干组频繁 N(N=1,2,3…)项集,对这些频繁项集进行排列组合，即可得到完整且粒度满足需求的概念集合。

(5)相似归并。至此,要利用这些概念仍存在一个问题，即挖掘出的概念集合中可能存在大量同义概念，如果不进行归并则无法准确判断概念的重要程度(如词频等)。本研究通过词型相似度计算(如 N-Gram 算法等)以及近义词匹配两步完成相似归并

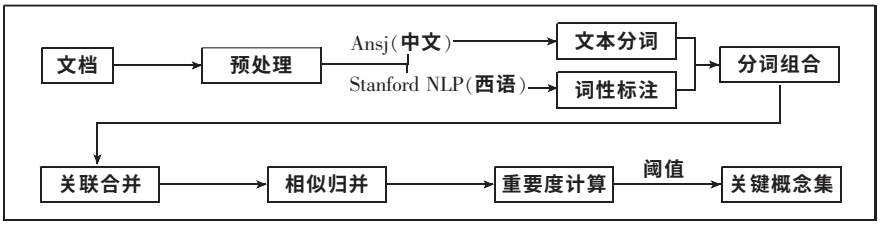

图 2　重要概念抽取流程

的步骤。近义词匹配一般可选择领域相关的叙词表或使用自然语言处理工具中自带的近义词语料库，同时，在此过程中实现相应的机器学习和训练。实测关键概念匹配度超过 75%。

(6)重要度计算。将进行相似归并处理后的概念集合，综合利用词频、TF-IDF 以及概念在信息主体中的位置权值（如出现在题名中的概念和出现在正文第一段中的概念可以被赋予更高的权值），综合制订算法，计算出每个概念在信息中的重要度，根据阈值取出满足阈值条件的概念即认为是信息的关键概念集合。同时，对关键概念集合中的关键概念根据重要度进行排序，可实现根据实际需求获取指定数量关键概念的功能。

抽取出的关键概念集合将会有两个用途，首先，是作为当前信息的标签与用户参数配置中设定的关注概念进行匹配，判断是否是用户关注意愿较高的信息；其次，是将关键概念集合中的每个概念作为新词发现的结果加入语料库。如果在其他信息中再次发现相似概念则可直接进行抽取。

(7)实体抽取。时间、领域相关的标号、识别码等信息的抽取可结合词性并采用正则表达式匹配的方式抽取，地域、人员、机构等信息可利用分词工具中的实体抽取方法抽取，同时可根据上下文结构进行筛选判断，最后再与相关词表进行匹配达到准确抽取的效果。

2.5 相似度匹配

当抽取出一系列关键概念和实体对象后，需要与用户在参数配置中设定的识别参数进行匹配。首先进行词型的匹配，然后基于词表进行词义匹配。实体需要结合实体规范库对实体对象进行统一表述，然后进行匹配。最后制订符合实际情况及需求的匹配度计算方法。将词性匹配、词义匹配及实体匹配的结果和数量等信息带入匹配度算法中得到匹配度。最终，通过与匹配度阈值比较，判断当前信息是否是用户关注的目标，并进行采集。

2.6 采集发布

自动采集发布主要包括以下流程：

(1)面向 IIBD 的自动登录验证配置。支持用户在采集发布接口中配置 IIBD 的登录信息（用户名、密码等），接口应用时会自动调用该配置信息和相应登录机制，向 IIBD 发出登录请求并进行验证，最后将登录验证成功与否的标志信息进行返回。登录验证主要是提高系统的安全性，同时也将采集发布功能与 IIBD 主系统实现解耦。当其他系统需要使用本接口时可通过参数配置快速调用。

(2)基于数据包方式的已采集数据与 IIBD 元数据的关联映射配置。支持用户将已采集数据的内容标签与 IIBD 元数据字段进行映射配置，主要包括采集资源类型的映射和元数据结构的映射配置，然后根据配置的信息采用 httpclient 提交 post 数据包的方式，将该信息模拟并构造为表单提交数据的方式，向 IIBD 工作流自动提交与确认发布数据。该步骤同样是实现采集功能与 IIBD 主系统解耦的一部分。

3 案例及应用效果

目前，本研究所述的基于自然语言处理的定题监测功能已经嵌入到产业情报大数据（IIBD）平台中，现已在 10 余家企业、机构的实际应用中完成部署并投入使用。从目前该功能在已部署机构中的使用情况来看，整体反映良好，对近千个监测源进行定题监测和采集，通过基于用户需求的配置，较好地实现了对各种不同类型用户感兴趣的多源异构信息源进行个性化配置并跟踪和采集的应用需求。

在关键概念抽取效果方面本研究随机选取了 500 篇提供关键概念标引的信息，用本文所属方法进行关键概念的自动抽取和对比，发现关键概念的命中率超过 75%。从监测和采集效果方面本研究遴选了 10 个不同类型的网站（综合类、政策类、机构/企业/协会门户类等），配置相应的主题、实体集参数，使用本文所述方法进行机器监测和采集，其结果与人工遴选、采集结果进行对比，机器采集到的信息比人工采集到的信息略多，采集到的信息较人工监测结果覆盖率超过 95%。与基于模板匹配、URL 规则和简单关键词匹配的传统机器采集方法相比，大幅减少了采集量，提升的采集效率和精确度。实现了在不降低查全率的基础上提升查准率和命中率的目标。在实用效果方面，从 IIBD 平台在多家企业投入实际使用的反馈信息来看，可以较好地满足用户在实际

工作中对定题信息监测和采集的需求，总体达到可投入实际使用的标准。

4　结语

本研究在一定程度上实现了通过用户个性化配置，对大量多源异构信息源进行自动化定题监测和采集的功能。在信息内容与用户关注度的匹配方面，通过使用自然语言处理的一些常用方法，有效提升监测的精准度,降低了用户的工作量,实现对传统定向定源监测采集功能的优化和改进。并且在关键概念及实体的抽取过程中，同时支持新词发现和部分机器学习的功能。在架构上通过开发相关接口和提供词典、语料及匹配规则的配置功能实现监测采集功能与主平台的解耦,支持在除 IIBD 以外的其他平台中快速便捷地嵌入。

本研究仍存在很多不足和提升空间，如在关键概念及实体抽取以及相似匹配的部分，目前项目组正在研究将深度学习的一些算法和方法应用进去,用以提高监测采集的智能化,进一步提升精准度,并以此提高监测效率和降低人工成本。此外,单从采集功能角度来看，对基于复杂 ajax 技术构建的情报源以及对微信开放公众号的监测采集的效果仍有待提升。以上问题和不足还需要通过进一步学习和掌握相关技术、工具、方法来予以优化和解决,从而对各类基于大数据概念构建的专题竞争情报平台提供更完善的监测采集支持。

参考文献:

[1] 产业情报大数据平台[DB/OL].[2018-05-08].http://tbea.llas.ac.cn/.

[2] Bar-Yossef Z,Rajagopalan S.Template detection via data mining and its Applications[C].In:Proceedings of the 11th International Conference on World Wide Web,Honolulu,Hawaii,USA. New York,USA:ACM,2002,5(10):580-591.

[3] 叶勤勇.基于 URL 规则的聚焦爬虫及其应用[D].杭州:浙江大学,2007.

[4] 蒋付彬.基于决策树的 URL 分类器算法及主题爬虫平台设计[D].成都:成都理工大学,2016.

[5] 杨镒铭.基于 URL 模式的网页分类算法研究[D].合肥:中国科学技术大学,2016.

[6] Debnath S,Mitra P,Pal N,et al.Automatic identification of informative sections of Web pages[J].IEEE Transactions on Knowledge & Data Engineering,2009,17(9):1233-1246.

[7] 王浩.基于半监督学习的网络敏感信息识别[D].天津:天津大学,2012.

[8] Pavlinek M,Podgorelec V.Text classification method based on self-training and LDA topic models[J].Expert Systems with Applications,2017(80):83-93.

[9] 李剑.基于 DOM 和神经网络的网页净化应用[J].电子科技,2012(1):105-107.

[10] 李伟男,李书琴,景旭,等.基于模拟退火算法和二阶 HMM 的 Web 信息抽取[J].计算机工程与设计,2014,35(4):1264-1268.

[11] 谢方立.基于节点类型标注的网页主题信息提取技术研究[D].北京:中国农业科学院,2016.

[12] 谭宗颖,王强,苍宏宇,等.科技发展前沿信息监测与分析平台的构建[J].科学学研究,2010,28(2):195-201.

[13] 刘海波.动态 Web 信息监测相关技术研究[D].哈尔滨:哈尔滨工业大学,2011.

[14] 张智雄,刘建华,谢靖,等.科技战略情报监测服务云平台的设计与实现[J].现代图书情报技术,2014(6):51-61.

[15] 谢靖,曲云鹏,刘建华.面向网络科技监测的分布式定向资源精确采集研究和应用[J].现代图书情报技术,2011(Z1):26-31.

[16] 王思丽,马建玲,王楠,等.开放知识资源的元数据自动采集策略研究[J].图书馆学研究,2013(12):47-51.

[17] 王思丽,刘巍,祝忠明,等.基于 CSpace 的科技信息可配置化自动监测功能设计与实现[J].数据分析与知识发现,2017(10):85-93.

作者简介 :刘巍(1980-),男,中国科学院兰州文献情报中心副研究馆员 ;王思丽(1985-),女,中国科学院兰州文献情报中心馆员 ;祝忠明(1968-),男,中国科学院兰州文献情报中心研究馆员 ;吴志强(1985-),男,中国科学院兰州文献情报中心馆员。