文章编号：2095-2163(2020)07-0170-06 中图分类号：TP315 文献标志码：A

面向政府采购数据的工程化采集方案设计

王 宏，夏 禹，常静静
（西安石油大学 计算机学院，西安 710065）

摘 要：政府采购过程中产生的大量招投标数据，基本都以 Web 文本的形式向公众呈现，难以获取结构化数据，严重制约着公众对政府采购过程的知情、分析和监督。本文提出一种基于 Web 挖掘的政府采购数据的工程化采集方案，构建了一套面向政府采购公开数据的结构化数据形成体系。首先，通过对招投标信息来源和结构的分析，设计基于 Scrapy 爬虫框架的工程化数据抓取平台；其次，结合基于规则和基于统计两种抽取方式，设计专用信息抽取器；最后，根据领域特点建立阶段性数据清洗中心，分层过滤数据，最终输出可用于分析和挖掘的结构化数据。系统实验结果证明了该方案的可行性和优越性，为政府采购信息公开发挥监督和引导职能提供了有力的技术支撑。

关键词：政府采购；Web 挖掘；Scrapy 爬虫；信息抽取；数据清洗

Design of engineering collection scheme for government procurement data

WANG Hong, XIA Yu, CHANG Jingjing
(Collage of Computer Science, Xi′Anshiyou University, Xi′an 710065, China)

【Abstract】 A large amount of bidding information is generated in the process of government procurement, which is presented to the public in the form of Web text. It is difficult for people to obtain structured data behind it, which seriously restricts the ability of realization, analysis and supervision of the public for the process of government procurement. This paper presents an engineering data collection scheme based on Web mining for government procurement data, and constructs a system for the structured data in public government procurement field. At first, an engineering data crawling platform based on Scrapy crawler framework is designed by analyzing the source and structure of bidding information. Secondly, a special information extractor is designed by combining rule-based and statistics-based information extraction methods. Finally, a stage data cleaning center is established according to the characteristics of the field where the data is filtered hierarchically, and the final output can be used for analysis and mining. The system experimental results prove the feasibility and superiority of the scheme, and provide strong technical support for the supervision and guidance function through the public information of government procurement.

【Key words】 Government procurement; Web mining; Scrapy crawler; Information extractor; Data cleaning

0 引 言

政府采购一直是中国经济社会发展的重要组成部分，随着信息技术的发展，政府采购的招投标方式也由线下转为线上，积累了大量相关数据，并依法对公众公开[1]。现阶段，公开信息时不提供可直接应用的数据集，而是以政府采购网站上的 Web 文本为主[2]。这对政府采购领域的大数据分析产生了巨大阻碍，一定程度上限制了公众对政采工作的监督力度。同时，参与政府采购的企事业单位广泛，业务层级划分严格，采购类型不一，涉及的商品与服务繁杂[3]，要对其进行精准的分析和研判，就必须通过一个工程化的系统来实现对政府采购数据的采集。

鉴于此，本文提出了一种基于 Web 挖掘技术的政府采购数据采集方案，方案从工程化角度出发，通过建立灵活、便捷的前端数据抓取平台，实现模块化可扩展的数据清洗中心和信息抽取器，结合自然语言处理的性能优化等过程，构建了一个完整的政府采购数据网络获取与结构化数据形成体系，更高效和便捷地分析政府采购数据。

1 研究背景

政府采购领域中数据采集的相关研究较少，且大多集中在数据的分析上，对数据如何获取和处理基本未做出较为清晰的说明。例如：直接使用已有的结构化数据对包含政府采购的公共资源交易数据进行分析[4]；文献采用人工收集和整理的数据对政府采购中建设项目做社团分析[5]。使用爬虫、数据提取和数据清洗技术来提取结构化数据，但未形成体系，且采集数量不足[6-7]；利用爬虫和网络文本抽

基金项目：教育部产学合作协同育人项目(201802224022)。

作者简介：王 宏(1968-)，男，硕士，副教授，硕士生导师，主要研究方向：云计算应用、大数据技术、信息系统集成；夏 禹(1992-)，男，硕士研究生，主要研究方向：智能计算与可视化；常静静(1995-)，女，硕士研究生，主要研究方向：软件工程、物联网技术。

通讯作者：夏 禹 Email：416130253@qq.com

收稿日期：2020-04-18

取的方式进行了数据获取[8]，但这对网络文本规范化要求较高，无法形成工程化抽取方案。总体来说，该领域研究多以算法实验方式进行的尝试性采集和处理，目前还未形成一套规范化、工程化的采集处理方案。

2　方案设计

本文在研究相关领域知识的基础上，充分分析政府采购领域数据结构，引入规则匹配和命名实体识别进行结构化信息抽取，以工程化思想设计整体架构，形成了一套面向政府采购数据的工程化采集方案。方案融合爬虫、信息提取、自然语言处理等多项技术，下面对整体架构、数据爬取、信息抽取和数据清洗四个部分依次说明。

2.1　整体架构

本文将采集过程分为数据爬取、信息抽取和数据清洗三个部分，整体架构如图 1 所示。

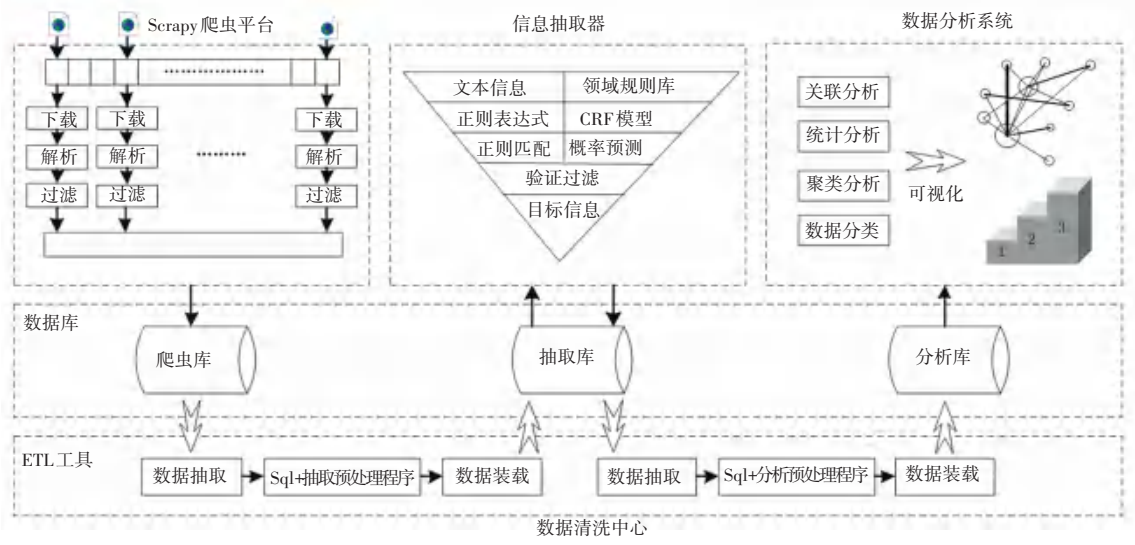

图 1　整体架构图

Fig. 1　Architecture of system

采用 Scarpy 爬虫框架以异步采集、异步存储的方式实现高效爬取；通过 ETL 数据抽取工具完成数据的清洗和转移；使用专用的信息抽取器完成目标信息的抽取；再次通过数据清洗中心送入数据分析系统。整个过程中数据的跨库传递都由 ETL 工具主动发起，大大增加了数据的可控性。同时，分工明确的系统设计降低了各个功能块之间的耦合性和开发维护的成本，也保证了从爬虫到分析数据的可追溯性。

2.2　数据爬取

通过对政府采购中招投标信息的分析，本文基于 Scrapy 爬虫框架设计了符合政采数据特点的工程化爬虫抓取器，内置定时增量爬取、异步加载、动态代理、异步存储机制，保证数据抓取工作高效稳定进行。具体爬取过程，如图 2 所示。

(1)分析目标页面结构，根据政府采购数据结构设计爬取器(Spiders)，抓取结构并设置初始链接参数；

(2)通过爬虫中间件(Spider Middewares)中的 scrapy-deltafetch 中间件重复过滤，实现增量爬取；

(3)将需要爬取的项目送入调度器(Scheduler)同步管理；

(4)将需要下载的项目通过下载中间件(Downloader Middewares)送入下载器(Downloader)，完成信息的异步加载和动态代理设置；

(5)将项目信息送入项目管道(Item Pipeline)进行异步存储，将正文 URL 通过爬取器重新送入请求队列，直至全部下载为止。

本文形成的爬虫架构，不仅可以通过 SpiderKeeper 等服务挂载多个爬虫定时爬取，而且支持 Scrapy-Redis 框架进行分布式扩展，为后期海量数据爬取提供可能。

2.3　信息抽取

在政府采购招投标信息中，大部分关键性信息都包含在正文中，由于其文本格式的特殊性和多样性，要从中提取出结构化信息并不容易。单独采用基于规则或基于统计的方式抽取文本中的信息都存在很大缺陷。本文结合两种抽取方式构建了信息抽取的基本架构，如图 3 所示。

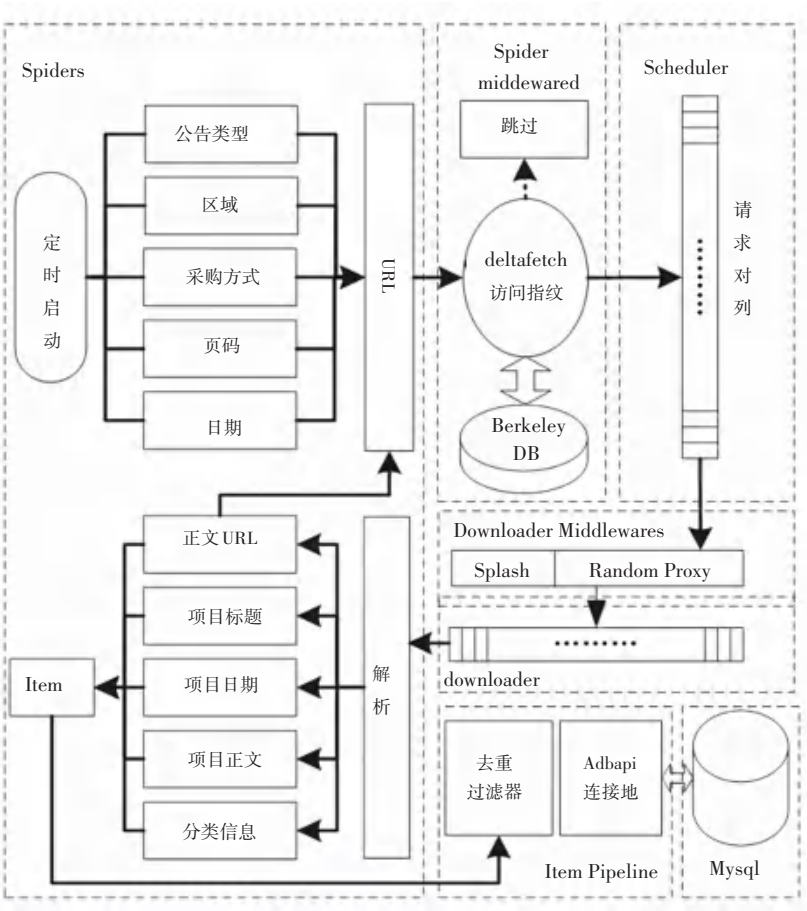

图 2 爬虫功能架构图

Fig. 2 Architecture of crawler function

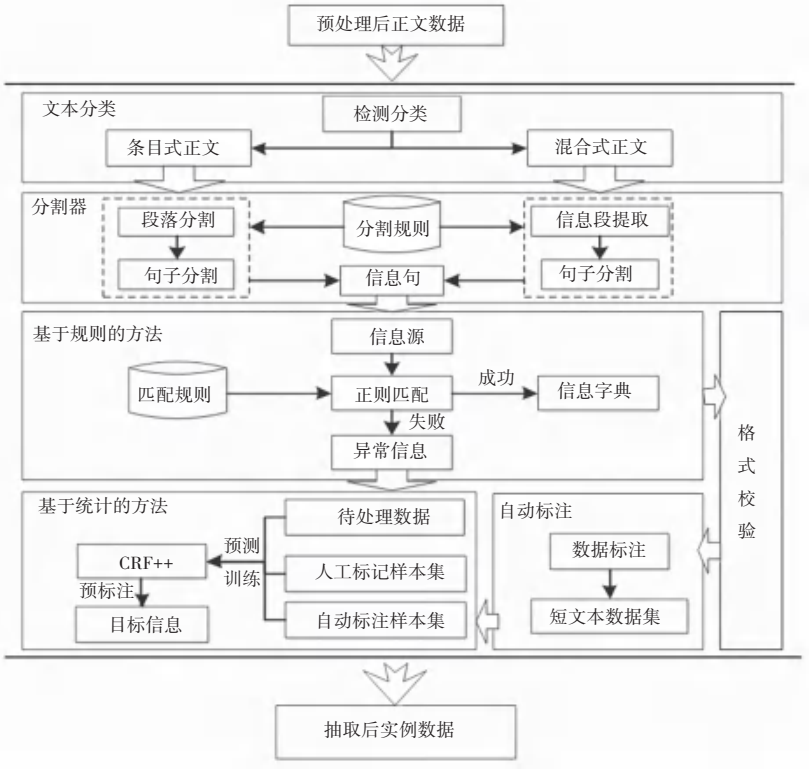

图 3 信息抽取结构图

Fig. 3 Structure of information extractor

(1)文本分类。由于早期政府采购中招投标信息未形成统一的格式规范,文本格式繁杂,但大体上可以分为以数字标号划分信息块的条目式正文和没有明确标号划分的混合式正文。本文通过正则匹配的方式将两类正文分开,方便后期单独处理。

(2)分割器。数据清洗过滤掉了 Web 文本中的噪音数据,保留了原有文本的段落格式,但无法克服抽取过程中目标数据之间的相互影响。为防止匹配过程中目标信息之间相互影响,本文根据招投标信息特点,在将正文划分为不同的信息块基础上进一步分割为信息句,缩小了匹配范围,并为建立自动标记语料集提供数据支撑。

(3)基于规则抽取。随着招投标信息的逐渐统一和规范,信息的表述方式具有了明显的规律性,通过特定的规则便可以抽取正文中的目标信息。本文利用正则表达式建立了严格的规则体系,可以实现快速准确地提取目标信息。

基于规则的抽取不能对早期不规范的信息文本有效提取,因此本文首先将得到的数据项与其相应句子结合,形成信息字典,再输入自动标注程序形成短文本数据集,为后面的统计抽取提供数据基础。

(4)基于统计抽取。为了弥补规则抽取在不规范文本抽取中的不足,本文利用规则抽取产生并标注的短文本数据集和少量人工标注样本集训练基于 CRF 的命名实体识别模型,并用于提取不规范文本中的目标数据。使用 CRF++工具包作为模型训练工具,输入经 TF-IDF 技术特征化后的样本集作为特征模板,标记规则匹配失败的异常文本信息,最终输出目标信息。

2.4　数据清洗

根据总体架构,数据清洗大致可分为爬虫清洗和抽取清洗两个阶段,分别处于数据爬取与信息抽取、信息抽取与数据分析之间。本文使用开源 ETL 工具 Kettle 完成数据的清洗和迁移工作。

Kettle 不仅提供了可视化的抽取、清洗、转换和装载过程,而且在传统 SQL 脚本的数据清洗之外,还支持基于 Java 的数据清洗方式。政府采购数据中很多数据的预处理工作很难用单纯的 SQL 清洗完成,因此本文在清洗流程中通过嵌入 Java 代码来实现对部分信息的预处理。数据清洗的总体结构,如图 4 所示。

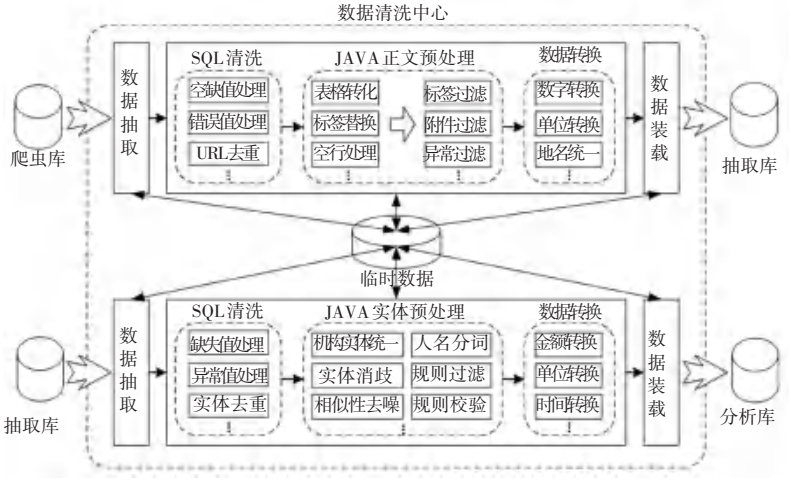

图 4　数据清洗的总体结构图

Fig. 4　Structure of data cleaning

数据清洗过程本身是一个迭代更新的过程[9],在构建清洗规则时需要不断迭代、优化。本文在两个阶段分别构建爬虫清洗迭代层和抽取清洗迭代层,每层都建立了一套完整的清洗流程,如图 5 所示。

在爬虫清洗阶段,对爬取到的项目列表数据进行缺失值处理、错误值处理、去重等一系列基本清洗操作后,还需要对正文数据进行预处理。包括 Web 噪音过滤、标签过滤、空行处理等,在具体实现上,可以先结合政府采购领域知识,过滤掉目标对象中的无关信息,之后通过 jsoup 库与正则表达式配合进行过滤,可有效减少清洗处理的复杂度。

在抽取清洗阶段,抽取后的数据仍会存在抽取不准确,存在噪音等问题。例如,正文本身信息错误所造成的错误值,以及信息项的值缺失等。对此,本文建立了相似性修正模块,对于达到相似性阈值的项目进行填充和修正。利用已有的专家姓名数据构建政府采购领域专家词典,载入 jieba 分词工具,训

练分词模型,对难以分割的专家姓名进行切分,并通过词频统计,对错误专家姓名进行修正,取得了不错的效果。此外,对政府机关单位有包括简称、别名、曾用名,从属部门等多种不同名称的问题,采用实体统一技术将机构实体进行统一,这也是数据分析前的重要一步。

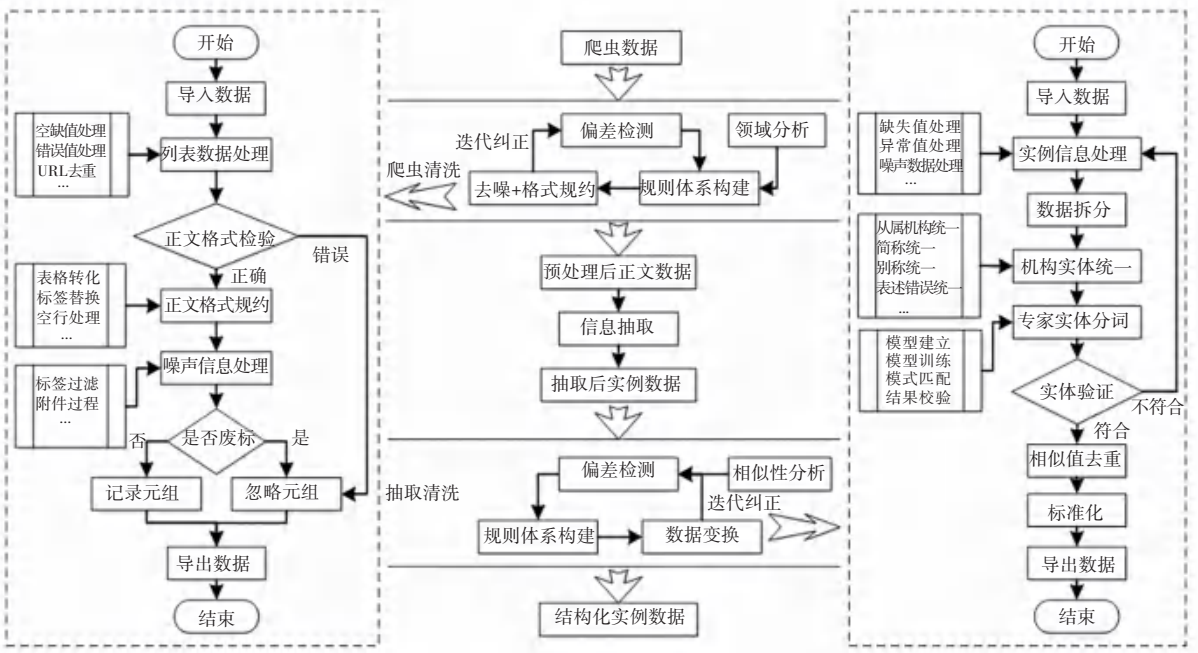

图 5 数据清洗流程图

Fig. 5 Data cleaning process

3 结果展示

本文以陕西省政府采购网中的结果类公告为例进行采集,时间跨度从 2010 年至 2019 年,涵盖了全省各市县区的所有采购方式和采购领域的数据,共计 120105 条。该数据的每一条都表示一条结果类公告,其中包含了该条公告所对应的地区、所属类型、采购方式、发布时间、标题、正文等关键性信息,如图 6 所示。

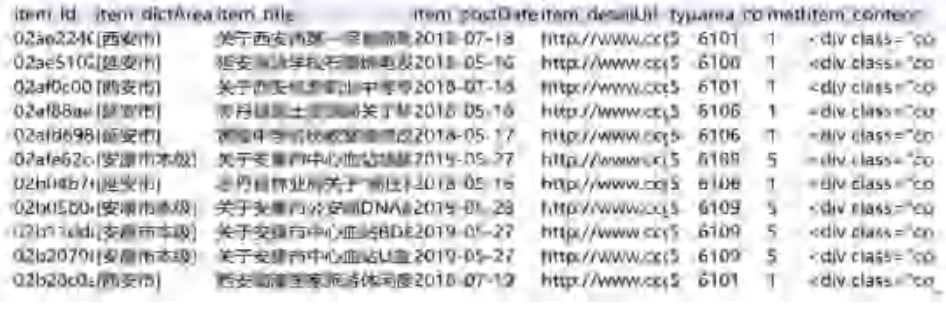

图 6 结构化数据

Fig. 6 Structured data

数据清洗后,通过信息抽取器提取正文数据中的参与方信息,包括采购单位(Purchaser)、代理机构(Agent)、供应商(Supplier)和评审专家(Expert)。由于各参与方之间存在一对多的关系,在对数据进行拆分后共得到158 279 条数据,每一条都表征了各参与方之间的一对一关系。使用基于规则的抽取方式成功抽取87 516 条结果公告,生成了自动标记样本集,结合人工标记的 522 条样本集训练 CRF 模型。最后使用 K 折交叉验证的方法进行模型验证,每次选取70 000条作为训练样本集,12 048 条作为测试样本集,4 次折叠后的模型评估效果见表 1。

表 1 CRF 模型评估结果

Tab. 1 Assessment results of CRF model

Entity	*Precision*/%	*Recall*/%	*F*1/%
Purchaser	71.3	72.5	72.1
Agent	94.5	95.1	94.8
Supplier	87.2	85.9	86.5
Expert	96.1	97.3	96.6

其中，采购单位的识别效果较差，只达到了72.1%，这是由于在结果类公告中采购单位名称复杂多样，包含不同地区、不同时期、不同级别的名称，且没有一个明确的标准，很难取得一个好的实体统一结果。最终在异常文本中成功提取25 413条结果公告，结合基于规则的抽取方式，整体上的抽取成功率达到了 94%，取得了良好的抽取效果。经过抽取清洗得到了148 513条用于数据分析的结构化数据，如图 7 所示。

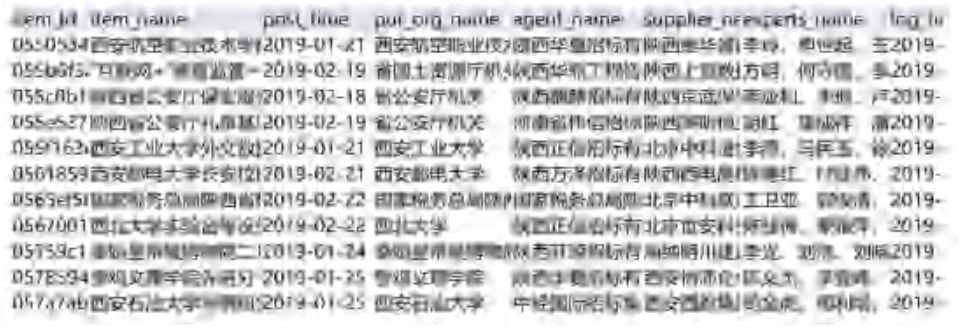

图 7　抽取到的结构化数据

Fig. 7　Structured data of information extractor

利用上述数据本文进行了简单的统计分析和关联分析，结果如图 8、图 9 所示。这表明：通过本文的工程化采集方案获得的数据能满足政府采购数据分析挖掘的需求，其结果可以反应出政府采购领域的一些趋势信息和关联关系，证明了方案的可行性和有效性。

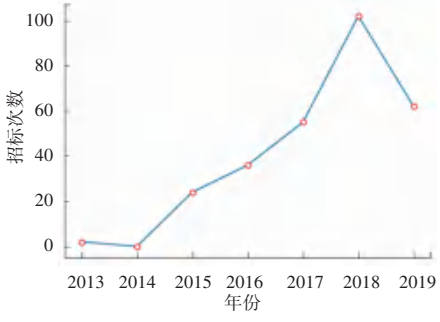

图 8　某大学招标频次趋势图

Fig. 8　Trend map of biding frequency of xx university

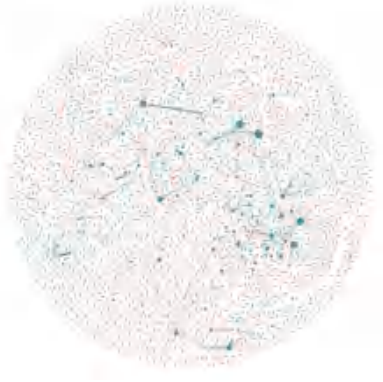

图 9　供应商与采购单位网络关系图

Fig. 9　Network diagram of relationship between Supplier andpurchasing unit

4　结束语

本文提出了一种基于 Web 挖掘的政府采购数据工程化采集方案，将数据爬取、信息抽取和数据清洗功能集成在一个系统框架中。文中提出的方案与其他采集方案不同的是，在设计上，本文方案以数据清洗中心作为系统的运作枢纽，以工程化思想设计各个功能块；在实现上，采用规则和 CRF 相结合的信息抽取方式，更具有普适性和优越性，将数据清洗分为爬取清洗和抽取清洗两步进行，并以陕西省政府采购网中的招投标信息为采集对象，展示了最终形成的结构化数据。结果表明，本文提出的政府采购数据采集方案结构清晰，适应性强，能快速有效地形成用于分析的政府采购数据资源，为数据分析提供有力支持。

参考文献

[1] 丁伟，边漫远，陈超. 浅议 Web 数据挖掘技术在政府采购中的应用[J]. 中国政府采购，2015(4)：70-71.

[2] 万如意. 大数据分析在政府采购领域中的应用：数据、技术与案例[J]. 中国政府采购，2015(12)：52-56.

[3] 刘彦军. 招标投标市场现状及发展[J]. 中国招标，2015(9)：12-14.

[4] 孙涵. 基于公共资源交易领域的知识图谱构建和可视化系统设计[D]. 中北大学，2018.

[5] 王兵. 基于复杂网络的建设项目投标人合谋行为分析[D]. 西安理工大学，2019.

[6] 王宏，门博，雷娜. K 近邻算法在政府采购数据挖掘中的研究与应用[J]. 智能计算机与应用，2019，9(3)：269-272.

[7] 雷娜. 政府采购信息多源聚合与关联分析的研究与实现[D]. 西安石油大学，2019.

[8] 孔维健. 基于图聚类的招投标数据挖掘研究与应用[D]. 中山大学，2015.

[9] 盛怡瑾，黄政，张学福. 面向领域分析的文献数据清洗策略研究[J]. 数字图书馆论坛，2015(12)：2-8.