doi:10.3772/j.issn.2095-915x.2019.05.010

基于 schema 的信息安全标准资源解析研究

开放科学（资源服务）标识码（OSID）

邢玉艳[1] 刘耀[1] 刘茹[2]

1. 中国科学技术信息研究所 北京 100038；
2. 北京大学 北京 100091

摘要：精准医疗伦理的信息安全标准体系构建过程中，会使用大量的资源类型，其中最为重要的就是标准资源，标准资源具有“非结构化”、“缺少语义信息”的特点，这对资源中重要文本信息和结构信息的提取造成了困难。本文提出针对信息安全标准数据特点的资源解析方法，基于“非结构化数据 - 半结构化数据”的转换思想和 XML Schema 技术，设计并实现了针对非结构化 PDF 文件的资源自动解析工具，将源数据中的非结构化数据转化为半结构化的 XML 格式数据，并以“整体式存储”方式存入 MongoDB 数据库中，实现了在 Solr 服务器中的检索功能和完成信息安全标准体系构建过程中信息提取、资源解析的工作。通过对比实验证明了基于 XML 数据的资源解析工具比 JSON 数据作为数据中间件，对信息安全标准的解析效果更佳。

关键词：非结构化数据；资源解析；XML Schema；数据存储；信息安全标准

中图分类号：G350

Application Research on Parsing and Storage of Standards Resource based on Schema Technology

XING Yuyan[1] LIU Yao[1] LIU Ru[2]

1. Center of Information Technical Support, Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Peking University, Beijing 100091, China

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”中课题 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）。

作者简介：邢玉艳（1991-），硕士研究生，研究方向：知识工程与知识发现；刘耀（1972-），博士，研究员，研究方向：知识工程与知识发现，E-mail：liuy@istic.ac.cn；刘茹（1998-），硕士研究生，研究方向：计算机辅助翻译，自然语言处理。

Abstract: In the process of constructing an information security standard system for precision medical ethics, a large number of resource types are used, the most important of which is standard resources. The organization and structure of resource data as well as the parsing process of semantic information is the foundation of text mining. This paper puts forward the "unstructured data to semi-structured data" transformation scheme, through the data exchange technology based on XML Schema, transforming effectively the unstructured source data into semi-structured XML data, with another essential idea of "integral storage" in MongoDB, successfully reached the final purpose of data retrieval in a Solr server,and completed the work of information extraction , resource analysis in the process of building an information security standard system. Through comparative experiments, it is proved that the resource parsing tool based on XML data is better than the JSON data as the data middleware. The parsing effect on the information security standard is more better.

Keywords: Unstructured data; XML Schema; resource parsing; data storage; standards of information security

引言

在精准医疗革命性的创新为人民带来福祉的同时，关乎医学伦理道德的安全问题也曝露在公众视野中。信息安全标准，作为精准医疗在伦理道德上受到的法律明文约束，是制定针对精准医疗伦理道德标准规范的重要数据资源。因此，要解决精准医疗伦理安全问题，就构建精准医疗伦理的信息安全标准体系，构建标准体系的过程中需要对信息安全标准资源进行充分的解析和预处理工作，这包括对资源进行语义化、结构化处理的过程。

资源解析是根据数据的结构化程度、类型、载体等层面，对获取的数据资源进行针对性的处理和解析，以获取资源的元数据信息、资源本身的结构信息及文本信息，形成统一的数据格式的过程[1,2]。而国家发布的信息安全标准文件均为 PDF 格式文本，具有“非结构化”、“缺少语义信息”的数据特点，这对标准文本内容的深度揭示、对标准体系构建工程的资源解析工作造成了挑战。

因此，针对信息安全标准资源的数据特点，本文提出了一种针对性的解析方法：基于“非结构化数据 - 半结构化数据”的间接转换思想，XML Schema 校验技术，及“整体式存储”的非结构化数据的数据库存储方法，构建了资源解析工具，实现了信息安全标准资源中的信息提取、快速检索。

1 信息安全标准资源解析概述

本文从信息安全标准的数据特点出发，回顾了资源解析的思想以及具体方法，主要从 PDF 格式的信息安全标准文件特点、针对资源的数据特点可采取的资源解析方法等方面进行了研究分析。

1.1 信息安全标准数据特点

在信息安全标准体系构建工程中，信息安全标准是工程在资源获取阶段得到的数据资源。通过网络爬虫获取到的中华人民共和国国家标准文件一共是 313 个 PDF 文档，涉及密码技术、个人信息安全、云计算等多个领域。PDF 是一种典型的非结构化数据格式，它的结构性实际

doi:10.3772/j.issn.2095-915x.2019.05.010

仅体现为面向显示的文档组织结构，而非有利于计算机理解的、语义层的结构信息[3]。因此，PDF 这种面向显示的非结构化数据格式对资源的存储和检索造成了困难[4,5]。

由中华人民共和国国家质量监督检验检疫总局和中国国家标准化管理委员会共同发布的 PDF 格式标准文件，是自然语言书写的、非结构化、不携带语义结构的数据资源，但标准文件有统一的要求，包含必要的字段，如中英文标题、分类编号等。

因此，信息安全标准资源数据具有两个重要的特点：一是数据量级小，PDF 文件数目仅为 313 个，单个文件大小在 100KB~10MB 之内，属于较小规格的文件；二是非结构化的文本数据，尽管文件具有相对统一的书写模式，但缺乏语义层面的结构信息和文本信息，即作为非结构化数据，缺乏数据的标引，不利于检索和进一步的文本挖掘。

针对信息安全标准文件“数据量小”的特点，对于标准文件的解析不以速度为第一要求，而是期望实现更高的信息提取准确率。针对“非结构化”的数据特点，对标准资源的解析以提取出元数据信息以及语义结构信息为目的，要求准确、完整地实现对非结构化数据的标引和信息提取，为下一步的文本挖掘和标准规范文件自动生成奠定基础。

1.2 信息安全标准资源解析技术

1.2.1 数据格式间接转换

对类似于信息安全标准文件的非结构化数据解析，目前学者们提出的解析思路可以分为“先转换后管理”和“不转换的一站式管理”两大方向[6]。后者的“一站式管理”思想主要基于全文检索的思路，其应用有如文献 [7] 提出的“基于双层 PDF 和 Lucene 技术的全文检索研究与实现”案例，构建全文检索系统直接检索双层 PDF 文件中的文本信息。全文检索的应用更多的是基于大数据技术的服务，阿里云等平台提供的一站式数据采集、存储、挖掘、和全文检索的大数据服务。可见，针对数据量级较低、非结构化特点显著的信息安全标准资源，全文检索和大数据技术难以应用在该种资源的解析上。

更多的研究和实验证明，基于数据格式间接转换的思想能够实现可持续、更高效的数据信息提取和管理。何卓桁[8]针对异构文本的数据挖掘，提出一种以 XML 文本为中间介质的间接转换方法，将非结构化数据转换为半结构化、最终转换得到结构化数据格式。宋艳娟[9]是“非结构化 - 半结构化格式转换”的代表，通过对 XML 文档的信息抽取，实现对 PDF 源文件的精确查找和管理，完成了 PDF 内容提取系统构建。文龙[10]则完成了“非结构化 - 半结构化 - 结构化”数据格式的转换，以 XML 格式数据作为中间件，完成非结构化数据到半结构化数据的转换，再以关系型数据库中的键值对作为结构化数据，完成 XML 数据到关系向数据库之间的转换。上述研究均表明基于 XML 中间格式的数据转换思想较为理想地实现了资源的解析和入库。

因此，本文采取简介转换的思想，通过将非结构化的 PDF 标准文件转换为半结构化的数据格式，从而实现资源解析和信息抽取。

1.2.2 数据“中间件”技术

文献 [8] 将数据间接转换思想分为两个阶段: 非结构化数据向半结构化数据转换的过程，及半结构化数据的解析过程。在数据转换的第一阶段，即非结构化的源文件向半结构化数据转换的过程中，需要选择半结构化数据的格式，即更适合于待解析数据集的数据“中间件”技术。目前主流的有XML和JSON两种数据交换格式。

由于标准资源缺乏语义信息的数据特点，对数据格式转换的意义在于更好地实现语义信息的标引。资源解析阶段的语义标注效果不涉及自然语言处理技术，语义标注的作用仅体现在标记非结构化资源文本的元数据信息、以及文本的篇章结构等其他语义信息，目的是为下一阶段的概念抽取、关系抽取、标准规范文本自动生成构建基础 [11]。因此，选择数据格式间接转换的实现方法需考虑资源解析的目的，而非简单的数据性能比较。

XML（Extensible Markup Language，可扩展标记语言）这种半结构化数据格式，以其可以标记数据、并自定义数据类型的特性著称，常作为一种数据交换格式 [16]。XML 允许数据处理工作者通过 XML DTD 或 XML Schema 等校验模式，对文档中元素的属性、次序结构、数据类型、默认值等特性进行声明和定义，从而实现 XML 数据的灵活描述以及语义层面的信息标注 [12,13]。

而 JSON（JavaScriptObject Notation，JS 对象简谱）作为一种轻量级的数据交换格式，以强大的可读性、可扩展性、解析速度等优势应用于 web 服务中的数据交换技术。但是 Baazizi[14] 提出，JSON 数据格式缺少 Schema 模式，因此缺少对 JSON 数据的描述，不利于对 JSON 数据结构的复杂检索。而 Ming Ying [15] 和 Boci Lin[16] 等人对 JSON 的应用探讨则侧重于其作为 web 应用的数据交换格式、在数据传输速率和性能方面优势的体现。随着 JSON Schema 的演化与规范，JSON 作为数据交换格式对于本文的标准资源的解析效果有待探讨。随着 JSON Schema 的演化与规范，JSON 作为数据交换格式对于本文的标准资源的解析效果有待探讨。随着 JSON Schema 的演化与规范，JSON 作为数据交换格式对于本文的标准资源的解析效果有待探讨。

1.2.3 XML Schema校验技术

在数据转换的第二阶段，即对 XML 文档进行校验时，常采用 DTD 或 Schema 模板作为解析或校验标准。胡志刚 [17] 提出在非结构化论文文本数据中，使用 XML 数据及 XML DTD 技术对论文学术信息进行格式化标注，通过匹配 DTD 标签，使得提取出论文中的学术信息变得容易。但是，相较于 DTD 的局限性，XML Schema“具有丰富的数据类型”、“支持用户自定义数据类型”、“语法与 XML 语法一致”等特性则体现出较好的灵活适应性，操作更为简洁，用于定义管理信息等更强大、更丰富的特征 [4,10,18]。

因此，本文采取 XML Schema 为模板进行数据格式转换，并采用 SAX 算法进行 XML 文档的解析，这样的方式能够最准确、最高效地完成对 PDF 向 XML 的转换，以及 XML 文档的解析入库，提供按关键字段数据库检索功能。通过这种方法将标准资源解析入库，为信息安全标准体系的构建奠定基础。

doi:10.3772/j.issn.2095-915x.2019.05.010

1.2.4 非结构化数据索引方法

信息安全标准体系的构建以及标准规范的自动生成建立在对数据格式转换后的 XML 数据的存储和检索管理之上。对于半结构化数据，其高效的存储、检索数据管理方法也随着数据管理技术的更新而更迭。在早期关系型数据库被广泛使用时，大多数学者提出借助 XML 定义数据类型的机制，将非结构化数据转化为结构化数据或半结构化数据，以结构化的键值对形式存储在关系向数据库中，使用结构化查询语言进行检索 [10,19,20]。

而吴胜斌 [21] 认为关系型数据库无法记载 XML 文档中的有序树形结构，不能很好地支持 XML 文档中的层级、顺序、包含等元素间的结构关系，因此难以满足网络环境下 web 应用对检索速率、全文检索等能力的要求。此外，XML 数据模型也会发生改变，因此用关系型数据库来存储 XML 数据不够灵活。

针对这一情形，本文将 XML 文档作为整体进行存储的方法能够保留完整的 XML 数据结构，并依据 Schema 数据类型定义提供数据库检索支持。而 MongoDB 作为一种面向对象的分布式文件存储数据库用于存储非结构化或半结构化数据具有较好的效率。并搭建 Solr 索引服务器。Solr 是基于 Lucene 框架、使用 Java 开发的全文搜索服务器，拥有更灵活的 XML 配置和更优的查询效率 [21,22]。

2 信息安全标准资源解析方法

针对信息安全标准非结构化、缺少语义结构的数据特点，为了有效提取资源中的元数据、文本结构等语义层信息，本文基于“非结构化数据 - 半结构化数据”的格式转换思想以及 XML Schema 校验机制，构建了针对信息安全标准资源的解析工具。

该工具由两部分组成：PDF 向 XML 文档格式的转换，以及基于 Schema 模板的入库校验。资源解析工具框架如图 1 所示。

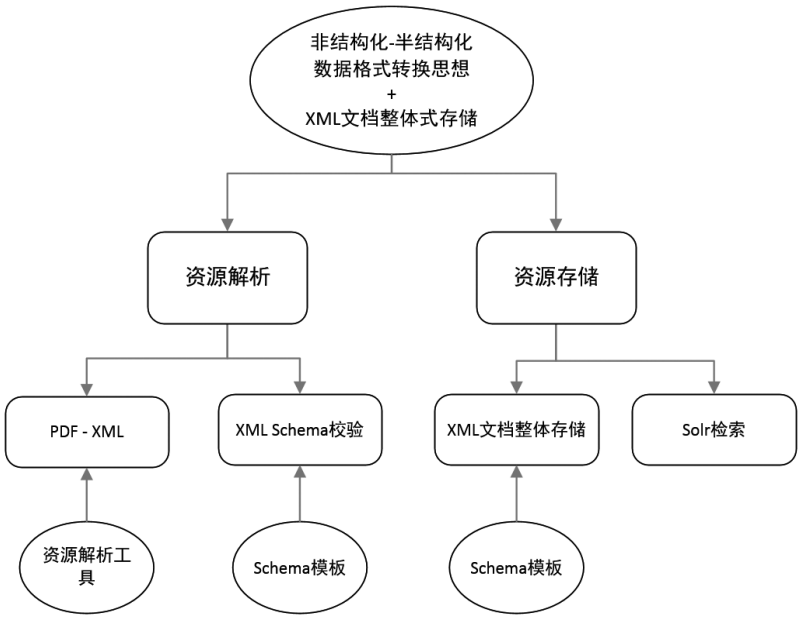

图 1 资源解析工具框架

（1）数据来源

在信息安全标准体系构建工程中，抽取313 个信息安全标准文件进行实验。这些安全标准文件为全国信息安全标准化委员会发布的国标文件，主要格式为 pdf 文档，涉及的领域包括密码技术、个人信息安全、云计算等。

（2）编写 schema 模板

Schema 模板中定义的元素、属性、次序结构、数据类型等关键信息是自动解析工具完成数据格式转换、以及 XML 文档校验的重要依据。

信息安全标准文件中待提取的是文件的元数据、正文及结构信息，因此 Schema 模板定义了 47 个自定义元素。其中，< 元数据 > 元素下的子元素包含了中英文标准名称、标准分类号、起草人等 13 个自定义元素，这些元素的数据类型均为字符串类型；由于标准文件中元数据信息是扁平的，元素之间的结构是平等的，元素之间的次序结构不做强调。而 < 正文 > 部分则由 < 段落 > 和 < 句子 > 组成，并且强调层级结构，因此使用 <sequence> 元素定义次序信息，承接不同等级的元素。

Schema 中定义的部分元素名称如表 1 所示，对应元素在代码编写 schema 如图 2 所示。

表 1 **Schema** 中定义的元素名称

元素名称	中文对照	数据类型
chtitle	中文标准名称	字符串
engTitle	英文标准名称	字符串
CSCNumber	中国标准分类号	字符串
drafter	起草人	字符串
draftingUnit	起草单位	字符串
implementationDate	实施时间	字符串
ISCNumber	国际标准分类号	字符串
references	参考文献	字符串
releaseDate	发布时间	字符串
releaseDepartment	发布部门	字符串
……	……	……

```
<xs:element name="standard">
    <xs:annotation>
        <xs:documentation>标准</xs:documentation>
    </xs:annotation>
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="chtitle"/>
            <xs:element ref="engTitle"/>
            <xs:element ref="info" minOccurs="0"/>
            <xs:element ref="body" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="chtitle">
    <xs:annotation>
        <xs:documentation>中文标准名称</xs:documentation>
    </xs:annotation>
</xs:element>
<xs:element name="engTitle">
    <xs:annotation>
        <xs:documentation>英文标准名称</xs:documentation>
    </xs:annotation>
</xs:element>
<xs:element name="info">
    <xs:annotation>
        <xs:documentation>元数据</xs:documentation>
    </xs:annotation>
```

图 2 **Schema** 模板部分代码

doi:10.3772/j.issn.2095-915x.2019.05.010

（3）自动解析工具编写

自动解析工具包括 2 部分：将 PDF 转化为半结构化数据，以及对半结构化数据的解析。通过调用 PDF2TXT 的解析工具包，将 PDF 文档转化为纯文本的 TXT 文档，再按照 schema 模板中定义的元素、结构、数据类型等对纯文本数据进行解析，提取其中的元数据信息、语义结构信息等，转化为半结构数据文档；针对 XML 文档，采用 XML Schema 进行校验、SAX 算法进行解析，对于通过校验的 XML 文档，提取其中的信息，并在下一步中存入数据库。

其中，提取 TXT 中的字段转化为 XML 文档的过程是通过观察信息安全标准文件具备的特点，总结规律、编写规则，采用正则表达式过滤文本来实现的。通过观察文本特点，思考如何用正则表达式筛选出 Schema 模板中定义的重要字段、并过滤干扰字段，来提取出标准文本中的中英文标题、国际标准分类号、中国标准分类号等字段，以及各字段对应的内容。并且，为了形成结构化数据、统一入库，在实验过程中对冗余的文本进行融合、删除等操作，形成了一套适应精准医疗信息安全标准体系的数据结构。自动解析工具的部分代码如下图 3 所示。

```
public class Pdf2StandardModel {
    public static StandardModel pdf2stanModel(String path){
        StandardModel model = new StandardModel();
      try {

          String pdftxt = PDFToText.getText(new File(path)) ;
        if(StringUtils.isEmpty(pdftxt)){
            return null ;
        }
        String[] senarr = pdftxt.split("\r\n") ;//以换行分隔成数组，得到出图片外的所有信息

          int[] chTitle = {8,0}; //记录中文标题的起始位置
          int[] engTitle = new int[2];//记录英文标题的起始位置
          int[] content = new int[2];//记录本标准  起始行数
          int[] introduction = new int[2];//记录引言  起始行数
          int[] body = new int[2];//记录正文的  起始行数
          int[] references = new int[2];//记录参考文献的  起始行数

          String reg = "^引\\s+言"; //匹配  引      言
          String reg1 = "^附\\s+录.*"; //匹配附录
          String reg3 = "^参\\s+考\\s+文\\s+献";
          String reg4 = "^[\\u4e00-\\u9fa5]{0,}$";//匹配中文标题
          String reg5 = "(.*?)+([A-Za-z]+\\s?)+(.*?)+([A-Za-z]+\\s?)";//匹配英文标题
```

图 3 解析工具部分代码

（4）MongoDB 存储技术及 Solr 查询

最终将得到的 XML 文件采用“整体式存储”方法，将 XML 文档完整地存储在 MongoDB 中，不对 XML 文档做拆分，以实现检索得到完整的 XML 文档的目的。配置数据库到系统平台中，搭建 solr 服务器，即可在 solr 数据库中按“中文标题”、“标准编号”、“一级标题”等字段检索标准资源。

3 实验设计

本文设计了对比实验，采用 XML 数据转换方法与 JSON 数据交换技术两种方式，对 313 份信息安全标准文件进行数据格式转换和资源解析，通过对比两种数据格式针对信息安全标准资源的解析效果，验证本文设计的自动资源解析工具在对缺乏语义信息、数据量小的数据

集上是否能表现出更高的解析性能。

3.1 实验环境

系统环境：64 位 Win 10 操作系统，8GB 内存，Intel（R）Core（TM）i7@2.40GHz

程序语言：JDK1.8，XML，JSON

实验工具： Eclipse Java 4.9.0

数据库：MongoDB 4.2.0

3.2 实验思路

通过编写 XML Schema 和 JSON Schema 两种模板，在自动解析工具中，将 PDF 源文件转换为 XML 和 JSON 两种不同的数据中间件，继而完成对资源的解析和信息提取。实验结果以文本信息提取的准确率、召回率为判断标准，测试两种数据中间件对自动资源解析的影响。

以随机选取的“2015 信息安全技术信息安全保障指标体系及评价方法第 3 部分实施指南 .pdf ”作为解析文件，该文件最长嵌套层级是 7 级。该文件的大小和层级结构均处于数据集的中间水平，具有一定的代表性。

分别编写 XML Schema 和 JSON Schema，其中 XML 文档具有 47 个自定义元素标签，共 285 个节点；JSON 文档也是 47 个自定义字段，234 个节点数。JSON 文档较之 XML，节点数减少的原因在于 JSON 可以以数组的形式“折叠”父节点下平行的节点，例如同样字段名和属性的“句子”节点可以用数组形式表示，位于“段落”父节点下。

4 实验结果与分析

论文设计并实现了针对信息安全标准文件的自动解析工具。本文从自动解析工具自身的解析效果以及对比其他解析方法两方面，验证了本文提出的以 XML 作为数据中间件以及利用 XML Schema 校验的资源解析方法具有较高的鲁棒性。

4.1 信息安全标准解析完成度

本文设计的自动资源解析工具对资源解析和信息提取工作的完成度较高。实验对 313 个 PDF 格式的标准资源进行解析，经过自动解析工具的处理和 Schema 校验，成功解析得到并存入数据库的 XML 文档有 306 个，解析率达到了 97.7%；其次，在 solr 数据库中按“标题”、“标准编号”等 schema 定义的元素名称对标准文件进行检索，成功得到包含检索词的对应字段的内容，以及所有匹配的标准文件。因此，从解析工作的完成度而言，本文设计的资源解析工具高效地完成了资源解析的任务，顺利实现了对 PDF 文档的语义结构提取。

但是，自动解析工具也存在一定的不足。针对信息安全标准文件的解析效果，抽取其中一份转换格式后的 XML 标准文件“信息安全技术 网络安全等级保护基本要求”做分析。将期望解析的结果与实际解析的情况做比较，发现了解析存在抽取了多余信息、未抽取到关键字段的问题。具体情况如表 2 所示。

4.2 对比基于JSON的数据转换方法

另一方面，通过与基于 JSON 数据交换技术的对比实验同样证明了基于 XML 的自动解析工具的鲁棒性。

doi:10.3772/j.issn.2095-915x.2019.05.010

表 2 本文基于 SCHEMA 校验的解析工具解析效果及原因分析

	期望解析结果	实际解析结果	未成功原因
总字段个数	47	43	正则匹配未精准匹配到某些字段；个别字段值为空；个别字段抽取了多余的文字
主题词字段	2	2	标准文件要求的必须字段
元数据字段个数	13	11	<中文标题>、<起草人>两个字段抽取错误 包含了字段以外的其他文字
正文及标题字段个数	32	29	<正文>字段下的<句子>未能够全部解析，真实的句子数目大于解析出的句子数。

统计两种数据格式以及对应 schema 解析出的字段数目、节点数目，并按照如下公式计算平均的准确率和召回率，两种数据交换格式在语义信息标注上的对比结果如表 3 所示：

准确率 = 提取无误的节点数 / 实际提取节点数 （1）

召回率 = 实际提取节点数 / 期望提取节点数 （2）

表 3 XML 与 JSON 解析资源提取信息结果对比

	XML	JSON
校验模板	XML Schema	JSON Schema
嵌套(父子)结构	标签嵌套	对象、数组交替嵌套
期望解析字段数	47	47
实际解析字段数	43	41
提取无误的字段数	43	41
期望提取节点数	285	234
实际提取节点数	270	219
提取无误的节点数	265	211
准确率	98.1%	96.3%
召回率	94.7%	93.5%
解析速率	0.62s	0.57s

结果表明，本文采取的 XML 数据“中间件”技术在对资源的复杂语义信息标引上具有更准确、更完整、更高效的标引能力。在解析数据量级较小、层级结构较为扁平的信息安全标准资源时，JSON 作为中间数据格式，未能体现出其速率和解析性能上的任何优势。相反，基于 XML 的解析工具提取的字段和节点总数更多，说明成功解析、并能够在数据库中检索的字段和对应信息更多。

出现这样的结果原因有以下几点：虽然 JSON 数据格式具有轻量级、扩展性强等优秀特点，且 JSON 的数据格式是对象与数组的交替嵌套方式，这在描述数据的次序结构上具有很大优势，这可以为 web 应用的实时数据交换提供数据高速、即时传输的保障，但对于本文的信息标注和提取的资源解析任务而言，JSON 数据的快速、即时性并无优势。另外，JSON 采用的数组存储数据的形式对于层级较多的文件相对占优势，但信息安全标准文件的层级结构并不复杂，结构的纵深通常在 10 个节点以内，因此以 JSON 作为数据中间件的优势在层级结构较为扁平的数据集上优势并不突出。反之，XML 作为一种成熟的标记语言，结合 XML Schema 校验模式，提供了对 XML 文档元素、结构、属性等进行灵活的编写和修改的功能，这使得 XML 作为标准文件的语义信息解析和提取的中间数据格式占据了优势。

4.3 实验分析小结

以上实验证明了本文提出的解析方法具有较高的解析完成度、鲁棒性和灵活性。信息安全标准涉及的领域广度较大，PDF 标准文件覆盖密码技术、个人信息安全、云计算等多个领域，而本文根据规律总结编写的解析工具和 Schema 校验模板仍完成了对大多数标准文件的解析，并且提供对比 JSON 数据交换技术，基于 XML 的资源解析工具体现出了较高的鲁棒性。

此外，基于 schema 校验的解析工具中 schema 文件体现出了较好的灵活性，schema 文件可以根据资源的组织特点随时调整，且 schema 校验机制不影响数据的存储空间，schema 文件所需的存储空间极小。

XML 文档存储方面，实验结果中 MongoDB 对 XML 文档的存储支持较好，具有较强的可移植性。本文采用的 XML 文档整体式存储在非关系型的 MongoDB 数据库中的方法相较于文献 [23] 所提出的 XML 文档与关系型数据库的映射机制提升了资源入库的效率。

总体而言，本文编写的自动解析工具及 Schema 校验模板，体现出良好灵活性和简便性，解析性能基本能够满足下一阶段文本处理的要求，并且通过建立 schema 模板确立了统一的数据结构，成功地将非结构化的标准文件转换为结构化数据存储在数据库，并能够灵活检索。但解析工具仍存在抽取到了噪音文字、部分抽取字段空值等不足之处。

5 总结与展望

信息安全标准的资源数据具有数据量级小、非结构化、缺乏语义结构信息的特点，这对于资源的信息提取、文本挖掘、以及和信息检索等工作造成了阻碍。为解决这一难题，本文基于“非结构化数据 - 半结构化数据 - 分布式文件存储”的数据转化思想，设计了自动资源解析工具，构建了 XML Schema 规范，完成了资源数据格式的转化，实现了在数据库中的结构化和整体化检索。并且通过对比其他资源解析方法，证明了本文提出的资源解析工具针对信息安全标注资源解析的有效性。

实验结果显示，XML 数据格式以及 XML Schema 技术对非结构化资源有比较不错的解析效果，但研究仍存在一定的局限：1）实验数据有限。实验进行前，通过网络爬虫等手段获取到符合要求的信息安全标准资源文件 313 个，实验仅完成对获取到的有限资源的解析，因此 Schema 技术对更大数据集的解析效果、以及对应的解析方法有待测试；2）目前的解析效果仍有待提升。Schema 中定义的元素名称和属性与部分标准文件的行文有所差异，导致不能够解析这部分文档，可完善 schema 模板，调整解析工具的设计，结合资源特点进一步提高 Schema 在系统中的利用率。

参考文献

[1] 刘耀 , 穗志方 , 胡永伟 , 等 . 基于内容与形式交互的图书馆资源组织语义化方法研究 [J]. 情报理论与实践 , 2010, 33(10):105-107+112.
[2] 刘耀 , 朱礼军 , 靳玮 . 专利信息资源挖掘与发现关键技术研究 [M]. 北京 : 科学技术文献出版社 , 2018:130-132.
[3] 张秀秀 , 张立峰 . PDF 文件文本内容提取研究 [J]. 科技情报开发与经济 , 2008, 18(36):118-120.

doi:10.3772/j.issn.2095-915x.2019.05.010

[4] 赵刚 , 于悦 , 黄敏桓 , 等 . PDF 阅读器字体解析引基于 schema 的信息安全标准资源解析研究擎的测试方法 [J]. 清华大学学报 (自然科学版), 2018, 58(3):266-271.

[5] 阎玮 . XML 解析技术研究及其实现 [D]. 上海 : 上海交通大学 , 2009.

[6] 杨桥 . 基于 MongoDB 的非结构化数据管理的研究与应用 [D]. 成都 : 电子科技大学 , 2017.

[7] 向禹 , 吴世明 . 基于双层 PDF 和 Lucene 技术的全文检索研究与实现 [J]. 现代情报 , 2014, 34(6):75-78. DOI:10.3969/j.issn.1008-0821.2014.06.015.

[8] 何卓桁 , 刘志勇 , 李璐 , 等 . 异构文本数据转换中 XML 解析方法对比研究 [J/OL]. 计算机工 程 : 1-8[2019-09-21]. https://doi.org/10.19678/j.issn.1000-3428.0054925.

[9] 宋艳娟 . 基于 XML 的 HTML 和 PDF 信息抽取技术的研究 [D]. 福州 : 福州大学 , 2005.

[10] 文龙 . XML 与非结构化数据管理 [J]. 电脑知识与技术 , 2009, 5(6):1306-1308.

[11] 赵东玥 , 杜永萍 , 石崇德 . 基于 BLSTM 的科技文献术语抽取方法 [J]. 情报工程 , 2018, 4(1):67-74

[12] Richard G. An XML schema for enhancing the semantic interoperability of archival description[J]. Archival Science, 2015, 15(3):295-313.

[13] Baqasah A, Pardede E, Rahayu W. XSM – A Tracking System for XML Schema Versions[C]. Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference.

[14] Baazizi M A, Colazzo D, Ghelli G, et al. Parametric schema inference for massive JSON datasets[J]. The VLDB Journal, 2019, 28(4):497-521.

[15] Ying M, James M. Refactoring legacy AJAX applications to improve the efficiency of the data exchange component[J]. The Journal of Systems & Software, 2013, 86(1):72-88.

[16] Lin B, Chen Y, Chen X, et al. Comparison between JSON and XML in Applications Based on AJAX[C]. Computer Science & Service System (CSSS), 2012 International Conference on, 2012.

[17] 胡志刚 , 田文灿 , 孙太安 , 等 . 科技论文中学术信息的提取方法综述 [J]. 数字图书馆论坛 , 2017(10):39-47.

[18] 刘力 . 科技文档信息抽取与格式化技术研究 [D]. 长沙 : 中南大学 , 2010.

[19] 於志勇 , 杨志义 , 於志文 , 等 . XML 数据存储方式的性能评价研究 [J]. 计算机工程与应用 , 2006(17):171-173.

[20] 赵春晖 , 张俊 . 查询专指度与检索多样化的关系研究 [J]. 情报工程 , 2018, 4(4):82-94

[21] 吴胜斌 , 张骏 . 基于数据库的 XML 存储技术设计研究 [J]. 信息技术与信息化 , 2018(10):93-95.

[22] 刘长生 , 周龙 . Oracle 数据库的 XML 存储技术研究 [J]. 电脑知识与技术 , 2018, 14(22):3-4.

[23] 史涛 , 沈艳霞 . XML 文档到关系型数据库的模型映射方法 [J]. 江南大学学报 (自然科学版), 2015, 14(5):590-595.