doi:10.3772/j.issn.2095-915x.2017.01.005

基于知识图谱的精细化工辅助研发平台

1. 武汉科技大学计算机科学与技术学院　武汉　430065；
2. 中怡精细化工集团有限公司　厦门　361000；
3. 荷兰阿姆斯特丹自由大学计算机系，荷兰　阿姆斯特丹　1081hv
彭彬[1]　杨晨[1]　蓝锦煌[2]　徐媚[2]　桂其迹[1]　黄智生[3]　顾进广[1]

摘 要　为每一个化学分子找到合理的归宿是化学工作者的使命，以 CAS 号为中心构建化学物质的知识图谱有助于帮助科学家为化学分子寻找新的用途。我们在精细化工领域进了有益的尝试。首先用知识图谱描棕精细化工产品的生产工艺，以此为基础集成工艺链上化学品基本知识库、反应知识库、专利及学术文献、红外图谱、质谱等。研发平台集成了本体编辑、可视化及基于机器学习知识图谱构建，支持化学分子表达式的知识库查询及可视化展示等功能。

关键词：知识图谱，化学工艺，化学知识库
中图分类号：TP182, TQ2

Knowledge Graph Aided Research and Development Platform for Fine Chemical Industry

1. College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
2. Ecogreen Fine Chemical Industry Group Limited, Xiamen 36100, China;
3. Department of Computer Science, Vrije University Amsterdam, 1081hv, Netherlands
PENG Bin[1]　YANG Chen[1]　LAN JinHuang[2]　XU Mei[2]　GUI QiJi[2]　HUANG ZhiSheng[3]　GU JinGuang[1]

Abstract　It is the mission of the chemical engineers to find a reasonable destination for each chemical molecule. The construction of the knowledge graph with CAS is helpful for the scientists to find new uses of chemical molecules. We have made a useful attempt in the field of fine chemicals. Firstly, this paper described

基金项目：本文受国家自然科学基金项目(61673304，61272110)；国家社会科学基金重大计划项目(11&ZD189)；武汉市科技攻关计划(2016060101010047)；软件工程国家重点实验室(武汉大学)开放基金（SKLSE2012-09-07）的资助。
作者简介：彭彬（1990-），硕士研究生，研究方向：语义WEB，email：827772196@qq.com；黄智生(1957-)，教授，博导，研究方向：语义WEB；顾进广(1974-)；教授、博导，研究方向：语义WEB与新型网络计算，email：simon@wust.edu.cn。

doi:10.3772/j.issn.2095-915x.2017.01.005

the production process of fine chemical products with the knowledge graph, which is based on the basic knowledge base, the reaction knowledge base, the patent and the academic literature, the infrared spectrum, the mass spectrum and so on. Then, the research and development platform integrated the functions of ontology editing, visualization and machine learning based knowledge graph construction, and the query of knowledge base and visualization for chemical molecular expressions.

Keywords: Knowledge graph, chemical process, chemical knowledge base

1 前言

大数据[1]时代提供了全新的信息环境，促使着我们的社会及其许多行业积极应对这个大数据时代提供的重大机遇。化工行业也面临着这个大数据时代的转型挑战，把信息化的手段引入到化工行业的工作流程[2]之中，使得我们能够更有效地利用大数据所提供的各种信息资源，更全面地把握行业状态，去挖掘新的信息。大数据处理的特征之一是多维数据、非结构化数据，例如化工行业的相关文献、专利、原料上下游关系，供应商上下游关系等。处理这些数据的一个基础是要建立一个可集成不同非结构化数据源的基础性元数据及化工相关的知识图谱。

Google 在 2012 年时提出了知识图谱[3][4]的概念，并用于 Google 搜索引擎上，提高了 Google 查询结果的质量。谷歌高级副总裁艾米特·辛格（Amit Singhal）博士认为“构成这个世界的是实体，而非字符串。知识图谱能够理解真实世界中的实体以及实体间关系”。自谷歌知识图谱之后，美国的微软必应，中国的搜狗、百度等公司也在短短的一年内宣布了各自的“知识图谱”产品。这些公司都希望利用知识图谱技术返回给用户真正需要的信息。知识图谱通常是以语义技术标准语言 RDF/RDFS 或者是对 RDF/RDFS 语言在逻辑表示上的扩展如网络本体语言 OWL 来表示的，其语义模型可以被看成是一个以 < 主语，谓语，宾语 > 形式表达的三元组集合。

表1 知识图谱规模

知识图谱	大小		
	实体	关系类型	三元组
Freebase	40M	35,000	637M
Wikidata	18M	1,632	66M
DBpedia(en)	4.6M	1,367	538M
YAGO2	9.8M	114	447M
Google Knowledge Graph	570M	35,000	18,000M

如表 1 所示，列举当前较大规模的开放知识库，这些数据通常以图形式的数据结构存储。通常具有很高覆盖面，可用于通用知识图谱的构建，也可用于行业知识库的构建，只是在行业知识库图谱的构建过程中，需要对数据进行一次筛选和映射。

在化工研发领域中最为关键的信息就是化学品的性能和生产工艺信息。了解各类化学产品的性能和生产工艺，能够为化学工作者在研发新产品的过程提供科学而全面的理论指导，提高产品研发效率。使化学工作者能够开发的

生产工艺中充分利用所有的化学分子，达到到绿色化学的目标。

在研发新产品的过程中，化学工作者为了能够对产品的性能，生产工艺等信息有较为全面的了解，需要借助网络的方式。但是网络中的信息纷繁复杂，人工的筛选工作费时费力。另外，科学合理地整合获得的知识也能更好地为将来科研工作服务。因此，构建精细化工领域的知识图谱将简化化学工作者的工作并协助化学工作者进行新知识的挖掘。

2 研究现状

2.1 知识图谱研究现状

知识图谱的构建方式一般分为三种：自顶向下的方式、自底向上的方式以及这两种方法相结合的方式。自顶向下的方式一般是首先构建顶层关系本体，然后将抽取到的实体匹配更新到所构建的顶层本体中。自底向上的方式则直接从抽取数据中发现到的类别、实体、属性以及关系合并到知识图谱中。不管使用那种方式构建知识图谱，其构建流程分为四个模块：知识获取、知识表示、知识存储以及知识可视化，如图 1 所示。

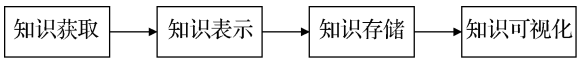

图1 知识图谱的构建流程

知识图谱构建过程中其重要环节之一是寻找合适的数据源并针对数据源进行信息的采集。当前的数据源可分为以下几种：

1. 结构化数据。结构化数据为存储在关系数据库或者面向对象数据库中的数据，在数据库被广泛应用的今天，已经有相当部分的数据都存储于关系数据库当中。但是由于这些都是深网数据，使用通用的爬虫很难获取。

2. 机器可读的开放本体或词典。机器可读的本体或词典通常是人工构建的，因此具有良好的可靠性。用于本体构建最多的当属 wordNet 和 Cyc，不过这两个本体均为英文词典，因此在中文知识图谱构建时难以直接使用。

3. 开放链接数据和开放知识库严格而言属于半结构化的数据。这些数据通常以图形式的数据结构存储。最知名的为 DBPedia、YAGO 和 Freebase，这些数据通常具有很高覆盖面，可用于通用知识图谱的构建，也可用于行业知识库的构建，只是在行业知识库图谱的构建过程中，需要对数据进行一次筛选和映射。

4. 行业知识库和行业垂直网站。它们属于半结构化数据，它们描述的目标是特定的领域，因为专注于特定领域，因此在数据一致性和完整性较完善。知名度比较高的数据集有，DrugBank、豆瓣、百度百科。

关于数据源与数据源采集环节，与英文知识图谱构建相比，中文知识图谱的构建具备以下差异和难点。1. 中文知识图谱构建的开放链接数据和开放知识库相对缺乏。表 1 所列举的知识库包含丰富的英文结构化数据，是做为知识图谱构建的丰富内容。然而这些数据集包含的中文知识非常有限；2. 中文的在线百科没有英文维基百科丰富；3. 没有像 WordNet 一样完整的中文词典库；4. 中文语言的自然特性与英文不同，英文中很多文本抽取与学习的方法不适用于中文。

针对于化工领域，国内还并没有一个相对

doi:10.3772/j.issn.2095-915x.2017.01.005

成熟的化工知识库。首先相对于英文，中文语言本身比较复杂，无法有效的使用领先的技术直接进行分析，而需要进一步的研究改进。其次，中文化工的专业语料库比较匮乏，也没有相关专业人士参与到只是分享。虽然可以利用机器翻译技术将化工相关的英文知识库翻译成对应的中文，而由于语言的差异性，语料知识的不完整性，容易产生语法错误、歧义问题以及专有名词的不一致性，因此无法满足对知识图谱要求高准确性的需求。

2.2 知识图谱辅助研发平台的研究现状

知识图谱通常是采用语义技术标准语言RDF 或者是网络本体语言来表达的。语义技术及其本体工程的相关技术自然被考虑用到精细化工的知识图谱的构建与管理之中。所以，本体工程 [5][6] 属于精细化工辅助研发平台的重要一部分，也是构建精细化工知识图谱的首要条件。研发平台包括本体的建立、修改、重用，实例添加等工作，由于涉及本体自身的建立方法等相关的知识，本体编辑是一个比较庞大的工程，比如就本体建立而言就比较复杂，它首先需要本体的模型，还有大量的实例数据填充，当前国外许多大学和研究机构正在研究和开发的本体编辑工具有很多，下面主要介绍 OILEd、Protégé2000 及 WebOnto 这几个本体编辑工具。

表2 常见本体编辑工具对比

特性	OILEd	Protégé2000	WebOnto
结构	独立	独立	Client/server
输入语言	RDF(S);OIL;DAML+OILXML;	XML;RDF(S);OWL	OCML
输出语言	OIL;RDF(S);DAML+OIL; SHIQ;Dotty;HTML	XML;RDF(S);XMLSchema;Flogic; CLIPS;Java;HTML	OCML;Ontolin-gua;GXL; RDF(S);OIL
存储形式	File	File	File
可扩展性	无	可以	无
实例录入	人工	人工	人工

从图 2 中可以发现本体编辑工具存在如下问题：1. 大多数工具均支持多种本体语言，但支持 W3C 最新推荐标准 OWL[7] 与工业上应用较多的 JSONLD[8] 的语言不多；2. 大多数工具以文件的形式存储本体内容，只有少数工具支持对本体内容的数据库存储；3. 实例的录入均需要人工手动录入而不能通过创建实例规则，自动化爬取实例数据。

化工辅助研发平台所涉及的知识构建流程主要如下：本体的构建，非结构化数据中提取结构化信息，对文本数据中提取结构化信息，数据的可视化。其中基于本体的构建，常借助上述本体编辑工具，进行可视化的本体设计；将非结构化的数据抽取为结构数据，常用方法是编写爬虫程序从网页中爬取数据，不提供可视化操作。国内常用的本体编辑工具如 Protégé 虽然具有成熟的本体设计与编辑，但是本体实例的数据需要人工录入，且仅支持 OWL 格式。目前还没找到一种合适的化工辅助研发平台，能集成化工本体的构建，非结构化数据中提取结构化信息，对文本数据提取结构化数据，数据的可视化。

3 精细化工知识图谱构建

3.1 精细化工知识图谱分析

构建精细化工领域知识图谱的第一步是确定本体构建领域，在本文中的本体构建领域为精细化工工艺。其次，确定本体构建领域包含的大致内容，化工工艺领域内容主要是化工工艺链上每一个化学品化学品的相关信息，根据化学品的应用分类，可分为基本知识，文献知识，反应知识，技术说明书四大类，根据此分类，本文知识图谱的构建可细分为四个小型知识库的构建。四种知识库可通过原料 CAS 号相互关联，由这四种知识库关联形成精细化工知识图谱。根据工艺包含的内容抽象出信息骨架，将这些信息骨架进行重组形成本体的层次关系及实体属性。

根据本体模型如图 2，化工工艺知识图谱构建的过程可分为三个步骤：构建本体层次关系；构建本体属性；构建本体实例。具体实现过程如下：

1）构建本体层次关系。根据化工工艺所涉及的内容，包括原料端，产品端确定化工工艺本体大致包含化学物的基本信息，文献库，化学反应，技术说明书信息。找到包含信息较全的数据来源，包括化工网，CA 美国化学文摘，国家知识产权局等，通过调研数据挖掘化工工艺信息的层次关系，层次关系如图 3。

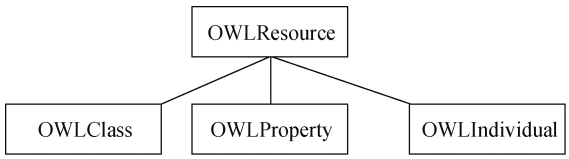

图2 本体模型

2）构建本体属性。根据图 3 的层次关系图，为每一个类添加属性值。化学品基本知识库，文献知识库，反应知识库，技术说明书知识库对应的属性图如图 4。

3）构建本体实例。根据步骤 1，2 定义的本体层次图及属性图，利用数据信息抽取平台，创建规则，抓取相应的本体实例数据，其中信息抽取来源可选取较权威普遍的站点，如化工网，CA 美国化学文摘，维基百科。数据抽取成功后，经过格式转换并上传至知识图谱研发平台。

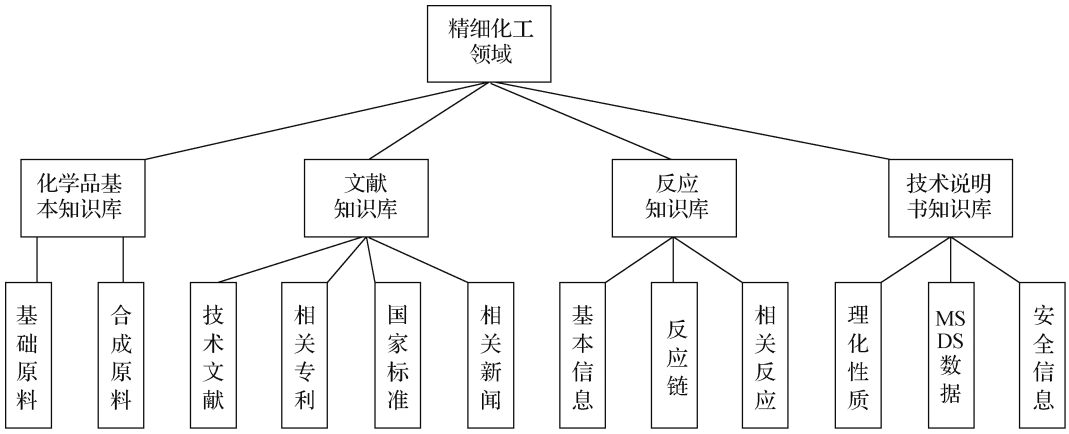

图3 化工工艺本体层次关系图

doi:10.3772/j.issn.2095-915x.2017.01.005

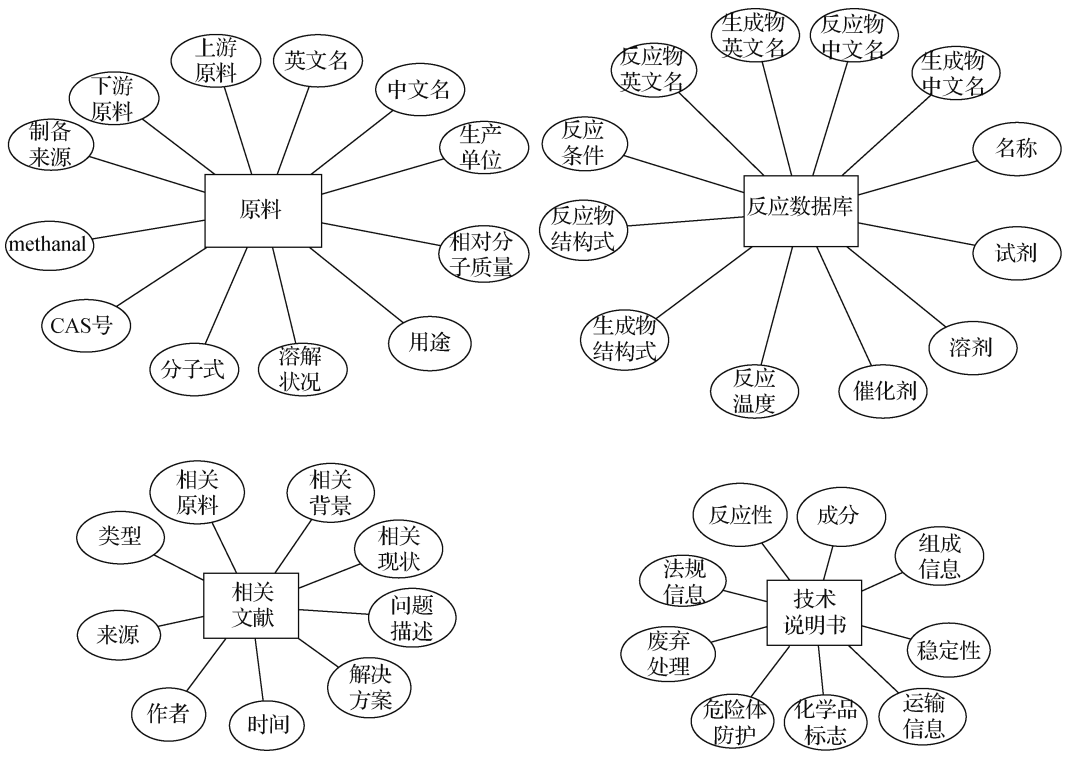

图4 化工工艺本体属性图

3.2 一种半自动化知识图谱构建机制

与传统完全基于人工或机器学习构建知识图谱的方式不同，本文将采用一种半自动化的方式构建精细化工知识图谱。它将分为四个步骤：

1. 如前一节所示，由化学专家和相应知识工程人员设计精细化工产品的核心工艺链上相关化学元素的实体及实体属性。

2. 利用可视化标注工具从专业化学网站或专业文献中抽取可以通过简单标注即可精确获取的实体及实体关系。限于技术的发展，目前仅实现了从专业网站抽取实体及实体关系。

3. 对于无法通过可视化标注即可精确发现的实体及关系，可以利用标注工具抽取相应的文本段，利用机器学习的方法进一步识别相应的实体及关系。

4. 利用机器学习的方法理解技术文献、专利等文档内容，并建立与相应知识图谱实体间的映射关系。

整个知识图谱构建过程可以用图5来表示，图中重点描3述了第2,3两个环节的技术实现细节。

3.3 精细化工知识图谱实例可视化构建

本系统考虑到化工网站中绝大多数实体数据存在于结构化的网页中，因此，本章提出一种面向相似页面的自动化规则抽取方式，其核心思想是模拟人认识网页中结构化知识的过程，利用标注工具及流程语言记录下认识过程，将其转化为图谱数据采集的脚本，然后交给数据抓取进程来执行。本系统不仅可以完成实体与属性关系抽取，还能够兼容常规的单页面级别的关系抽取，可以作为初步化工数据收集和处理平台。用户通过人机交互页面进行实体属性关系标记，保存并生成抽取规则，最后调用抓取脚本结合抽取规则获取精细化工图谱数据。

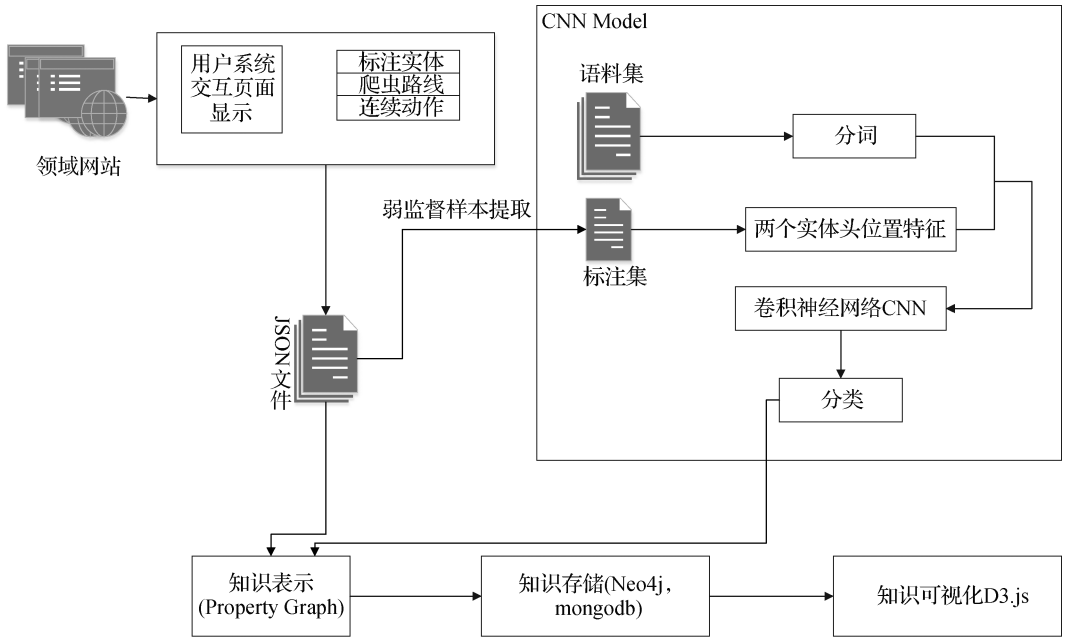

图5 化工精细知识图谱构建系统体系结构

基于化工网站实体数据的抽取系统的体系结构，总共分为三大系统模块：用户系统相互交互的页面展示，用户手动分离实体属性关系标注过程，模板生成以及抓取过程。

本系统是一个面向实体与属性关系分离的自动抓取系统，通过手动创建一个或者多个实体与属性的关系页面，完成对实体与属性分离的操作，系统会自动生成相应的规则进行数据抽取工作。

例如图 5 所示，该图显示的是化工大辞典的一个化工数据展示，此时可将“化工信息”作为所有化工数据的实体，包含了多个属性，如图 6 所示，例如“中文名称”为其属性，而“兔耳草醛”为这个化工数据属性的具体信息。

图6 化工数据样例

doi:10.3772/j.issn.2095-915x.2017.01.005

3.3.1 用户系统相互交互的页面展示

由于需要提供一个人工抽取的操作进而形成抓取规则，该模块提出了用户系统的相互交互的页面展示。该模块如图 7 所示，主要有由两个基本框架组成。一个作为需要抓取页面的展示功能，而右边是主要的用户操作按钮，首先用户需要输入要标注的页面的 URL 链接，系统会通过 HTTP 协议来获取网页信息，并格式化和清理文档的标签，对于 a 标签的 href 属性，会自动生成绝对并且正确的编码 HTML 文档，最后将完整的页面显示在框架中，便于之后的实体属性标注操作。

3.3.2 用户手动分离实体属性关系标注过程

用户交互页面展示成功后，用户可以开始进行实体与属性关系的标注过程。如图所示，用户需要创建抓取规则，并通过 Xpath 标注实体位置，以本例说明，用户创建了“中文名称，中文同义词，英文名称，CAS 号，分子式，分子量，EINECS 号”等，之后用户需要通过手动点击页面定位属性位置，如图 8 所示：

3.3.3 模板生成以及抓取过程

用户将标注工作完成之后，生成抓取规则，即抓取模板的生成，将已经标注的实体以、属性和属性值对应的 Xpath 保存，最后将所有的规则保存成 XML 格式的数据信息，系统会识别 XML 格式信息来进行抓取工作。如图 9 所示，抽取模式的数据：

图7 用户抓取交互界面

图8 用户定位属性界面

```
<xsl:stylesheet version=\"1.0\" xmlns:xsl=\"http://www.w3.org/1999/XSL/Transform\">
    <xsl:template match=\"/\">
        <xslAllType>
            <香料信息>
                <中文名称>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[1]/td[2]/a\"/>
                </中文名称>
                <中文同义词>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[2]/td[2]\"/>
                </中文同义词>
                <英文名称>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[3]/td[2]\"/>
                </英文名称>
                <英文同义词>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[4]/td[2]\"/>
                </英文同义词>
                <CAS号>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[5]/td[2]/b\"/>
                </CAS号>
                <分子式>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[6]/td[2]\"/>
                </分子式>
                <分子量>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[7]/td[2]\"/>
                </分子量>
                <EINECS号>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[8]/td[2]\"/>
                </EINECS号>
                <相关类别>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[9]/td[2]\"/>
                </相关类别>
                <Mol文件>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[10]/td[2]/a\"/>
                </Mol文件>
                <Mol文件url>
                    <xsl:value-of select=\"//div[@id='ContentPlaceHolder1_ProductProperty']/table[2]/tbody/tr[10]/td[2]/a/@href\"/>
                </Mol文件url>
```

图9 XML规则格式

用户保存规则并进行抓取，对于相似页面的抓取，用户只需导入此规则即可进行抓取，抓取后形成的 XML 格式数据如图 10 所示：

3.4 基于事件 Pattern 的技术文献意图理解

建立基本知识图谱只是精细化工辅助研发平台的第一步，要使研发平台发挥作用，还需要在知识图谱的帮助下，从海量技术文献中发现有价值的研发信息。因此理解技术文献文本并建立与知识图谱中相关实体的关联关系也是辅助研发平台的一个重要过程。然而，由于技术的限制，我们目前很难实现一个化工技术论文或专利的全文理解。但我们认为一个技术文献的意图是研究人员最关注的部分，表现在一个论文的摘要和前言部分，专利的摘要及权利要求书部分，因此在实际实施过程中，我们只对技术文献中的摘要部分进行了具体分析。由于技术文献的写作相对规范，一般表达“研究这个主题的意义，目前存在什么样的不足，本技术文献解决这个问题的思路、技术手段，以及研究或实验的效果”等。我们可以将一个技术文献所表达的意图理解一个或多个事件及其演化的过程，总结技术文献中意图描述部分的事件模式特征 (Event Pattern)，将上述特征模板化 (Template)，然后利用模板来理解技术文献中摘要部分的内容，并建立与基本知识图谱实体间的映射关系。技术实施细节可以通过图 11 表示：

```
<?xml version="1.0" encoding="utf-8"?>
<xslAllType>
 <香料信息>
  <中文名称>兔耳草醛</中文名称>
  <中文同义词>A-甲基-4-(1-甲基乙基)苯丙醛;α-(对异丙基苄基)丙醛;α-甲基-4-(1-甲基乙基)苯丙醛;对异丙基-α-甲基苯丙醛;对异丙醛-α-甲基氢化桂醛;兔耳草醛;仙客来醛;
  <英文名称>3-(4-ISOPROPYLPHENYL)ISOBUTYRALDEHYDE</英文名称>
  <英文同义词>(R,S)-p-Isopropyl-α-methylhydro-cinnamaldehyde;.alpha.-methyl-4-(1-methylethyl)-Benzenepropanal;3-(4-iso-Propylpheny1)-2-methylpropanal;3-(4-Isoprop
  <CAS号>103-95-7</CAS号>
  <分子式>C13H18O</分子式>
  <分子量>190.28</分子量>
  <EINECS号>203-161-7</EINECS号>
  <相关类别>Pharmaceutical Raw Materials;Alphabetical Listings;Flavors and Fragrances;M-N;食品添加剂;食用香料（增香剂）;天然等同香料和人造香料;合成香料;日
  <Mol文件>103-95-7.mol</Mol文件>
  <Mol文件url>http://www.chemicalbook.com/CAS/mol/103-95-7.mol</Mol文件url>
 </香料信息>
 <香料性质>
  <沸点>270 °C(lit.)</沸点>
  <密度>0.95 g/mL at 25 °C(lit.)</密度>
  <折射率>n20/D 1.505(lit.)</折射率>
  <FEMA>2743</FEMA>
  <闪点>228 °F</闪点>
  <CAS数据库>103-95-7(CAS DataBase Reference)</CAS数据库>
 </香料性质>
</xslAllType>
```

图10 抓取结果数据

doi:10.3772/j.issn.2095-915x.2017.01.005

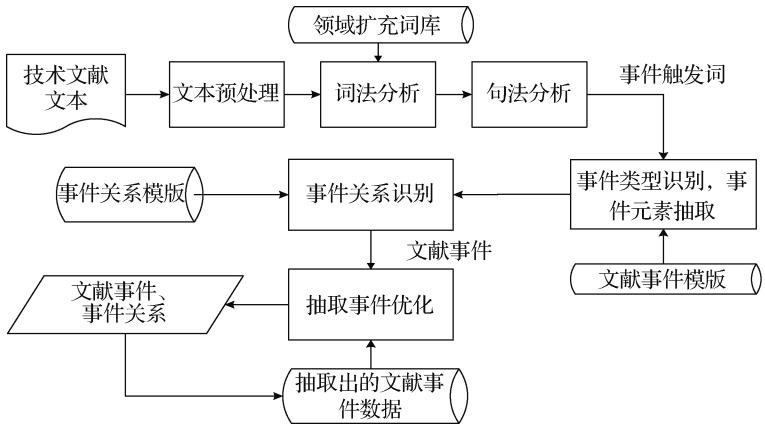

图11 基于事件Pattern的技术文献意图理解

3.5 化学结构检索设计

化工工艺领域，我们是以 CAS 号为中心构建化学物质的知识图谱，由于化学结构式代表着一个化学分子的唯一标号，通过化学结构式去检索想要的化学分子具有重要意义。在化工工艺辅助平台中，化学结构式是以图片的形式呈现给客户，不能满足检索需求，该章节对结构式的查询进行一次探索。化学分子结构式存储常用的存储方法是SMILES编码。SMILES 是一种线性码，即是一种用 ASCII 字符串明确描述分子结构的规范，可用于存储化学结构式信息。在系统中，每一种化学分子的存储的唯一标识是 CAS 编号。如果将 CAS 与 SMILES 建立起关联，即可将结构式查询成为一种可能。故第一步是将 CAS 与 SMILES 之间的关联表存储到知识库中。具体步骤如下：

1) 使用第三章讲解到的工具前往化工网中去爬取每一种化学分子 CAS 号对应的 SMILES 结构式；

2) 将爬取结果以 xml 的方式表示。存储格式如下；

<compound id=”7580”>

<cas>101-81-5</cas>

<smile>C1=CC=C(C=C1)CC2=CC=CC=C2</smile>

</compound>

<compound id=”31253”>

<cas>123-35-3</cas>

<smile>CC(=CCCC(=C)C=C)C</smile>

</compound>

3) 将存储结果转化成 RDF[9] 三元组，存储格式如下：

<http://wasp.cs.vu.nl/chem#hasCAS>”101-81-5”.

<http://wasp.cs.vu.nl/chem/id#7580><http://wasp.cs.vu.nl/chem#hasSMILE>”C1=CC=C(C=C1)CC2=CC=CC=C2”.

<http://wasp.cs.vu.nl/chem/id#31253><http://wasp.cs.vu.nl/chem#hasCAS>”123-35-3”. <http://wasp.cs.vu.nl/chem/id#31253><http://wasp.cs.vu.nl/chem#hasSMILE>”CC(=CCCC(=C)C=C)C”.

4) 将三元组导入 Neo4J 知识库中

导入后将 CAS 与 SMILES 关联成功。系

统进行结构式查询是可以实现，首先在可绘制化学结构式的面板控件中画出化学分子的结构式，完成后，改面板控件会反馈给你绘制的化学结构的 SMILES 表达式，将 SMILES 表达式传到 Virtuoso 中，得到该 SMILES 表达式对应的 CAS 号，再通过 CAS 号即可获得对应的化学分子的信息。在实际应用中，往往是需要部分结构式检索，也就是说通过输入部分结构式，能够召回包含该化学结构的所有化学分子。那么就需要在上述步骤中做改进。本章中，改进的思路是通过研究 SMILES 表达式，寻找判断 SMILES 之间是否具有部分匹配关系，本文的实现方式是将获取的 SMILES 表达式切割成一个个不可再分的单位，统计 SMILES 表达式包含的最小单位的种类与个数，然后根据这些比较两个 SMILES 表达式是否具有部分结构匹配关系。具体流程图如图 12：

进行上述步骤后，即可实现化学结构检索功能。总结以上步骤，首先找到 SMILES 与 CAS 之间关联；第二，将关联信息存入 Neo4J；第三，寻找 SMILES 之间的部分匹配关系。本文以苯分子结构为例，化学结构检索结果如图 13、14。

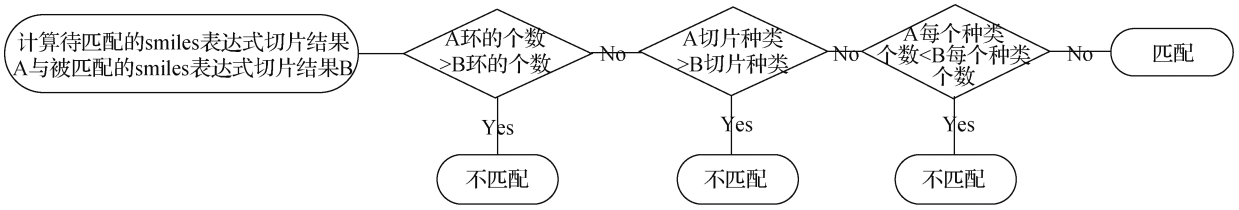

图12 判断SMILES之间结构是否匹配流程图

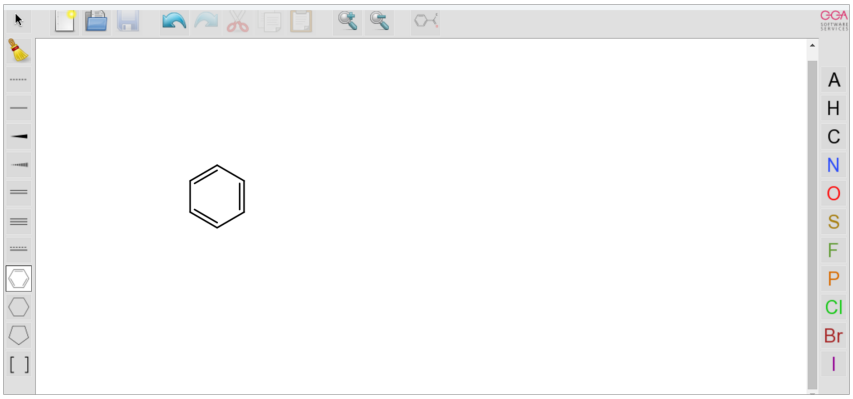

图13 苯结构的子结构查询界面

图14 苯结构的子结构查询结果

doi:10.3772/j.issn.2095-915x.2017.01.005

3.6 辅助研发系统原型

辅助研发平台原型如图 15 所示，辅助平台支持以下功能为：本体编辑，实例管理（抽取、人工录入、查询），技术文献管理（文献导入、文献理解、人工标记），查询及可视化展示。其工作流程可简述为：1. 通过本体编辑功能项，进行本体的层级设计，及属性添加；2. 进入实例录入功能项对本体实例的添加，其中实例的添加借助可视化抽取工具；3. 导入相应的技术文献，通过机器学习的方法建立与图谱实体的映射，允许化学专家手工标记或删除错误的映射关系；4. 借助实例查询，可方便快速查询实例信息，同时提供可视化展示效果。

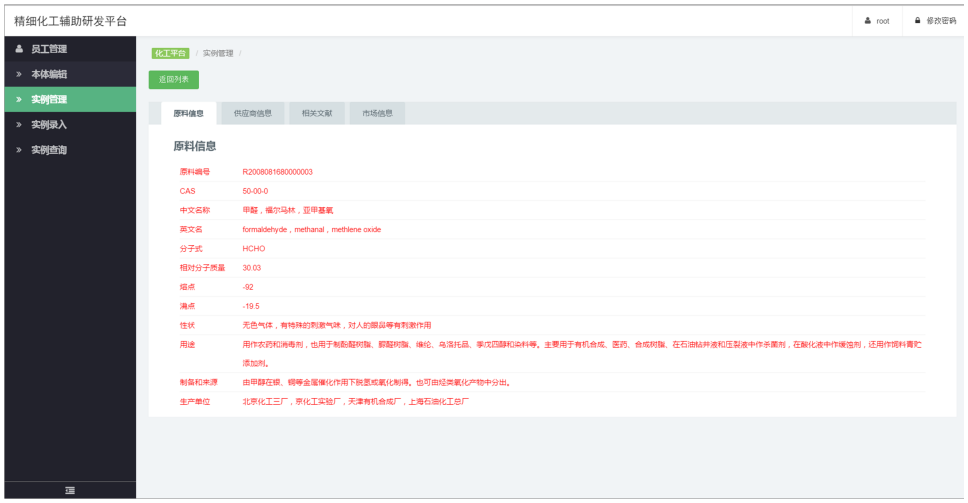

图15 辅助平台效果图

4 结论与展望

本文讨论了一种精细化工的知识图谱构建机制及面向研发的辅助应用，重点讨论了可视化的图谱数据抽取机制、面向事件 Pattern 的技术文献意图理解及支持化学结构式的图谱数据查询等技术实现思路。本文的工作表明利用知识图谱实现行业或领域知识管理和知识发现是可行的，而且具有十分重要的应用价值。

可视化的图谱数据抽取机制现阶段可用于非结构化数据的提取，文中提到的抽取工具现具有通用性，可用于行业垂直网站结构化数据的提取；利用机器学习的方法理解技术文献、专利等文档内容对文本的抽取为知识图谱的构建提供一种新的数据源采集方法；实现一个类 Protégé 的功能，集成知识获取、知识表达、知识存储、知识表示于一体的图谱构建平台，可为快速孵化行业知识图谱提供一种思路；支持化学结构式的图谱数据查询，画化学结构图即可查询图谱实体数据，丰富图谱的查询方式。构建出的图谱亦可快速构成行业词典，为文本的实体及实体关系的识别提供实现基础。

在了解本文工作能给化工知识图谱和知识发现带来的应用价值之外，我们也应该看到，基于知识图谱辅助研发平台的应用效果取决于基本知识库有精细程度和对技术文献理解的深度。因此我们将继续探讨基于知识图谱的技术文献理解及以此为基础的知识发现技术。

参考文献

[1] 王元卓 , 靳小龙 , 程学旗 . 网络大数据 : 现状与展望 [J]. 计算机学报 . 2013, 36(6): 1125-1138.

[2] 中国化工报 . 大数据分析是化工行业趋势，应用能产生效益 . [EB/OL]. [2016-5-12] . http://www.cbdio.com/BigData/2014-11/27/content_1931113.htm.

[3] 陈悦 , 刘则渊 , 陈劲 , 等 . 科学知识图谱的发展历程 [J]. 科学学研究 , 2008, 26(3): 449-4660.

[4] Chen C. Searching for Intellectual Turning Points: Progressive Knowledge Domain Visualization[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(suppl): 5303-53100.

[5] Gomez-Perez A, Corcho-Garcia O, Fernandez-Lopez M. Ontological Engineering[J]. Advanced Information & Knowledge Processing, 2004, 47(2): 69-755.

[6] Perez G, Asunci, et al. Ontological Engineering: with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. (Advanced Information and Knowledge Processing) [M]. Springer, 20077.

[7] Antoniou G, Harmelen F V. Web Ontology Language: OWL[M]// Handbook on Ontologies. Springer Berlin Heidelberg, 2009: 67-92.

[8] JSON for Linking Data[EB/OL]. [2014-01-15].http://json-ld.org/.

[9] Decker S, Melnik S, Harmelen F V, et al. The Semantic Web: the Roles of XML and RDF[J]. IEEE Internet Computing, 2000, 4(5): 63-7373.

[10] Erling O, Mikhailov I. RDF Support in the Virtuoso DBMS[J]. Studies in Computational Intelligence, 2009(221): 59-68.