电力领域语义搜索系统的构建方法①

姬 源, 谢 冬, 周思明, 王向东

(贵州电网公司 电力调度控制中心, 贵州 230039)

摘 要: 电力领域不断累积大量的数据资源, 包含相关标准规范、技术文档、管理文档、故障解决记录等等, 如何对这些文档进行快速查询和智能搜索, 对于电网调度与故障恢复具有重要价值. 传统的电力领域搜索系统都是基于关键词来实现, 存在查准率和召回率低的问题, 无法理解业务语言, 无法支持语义推理. 本文设计实现一种基于知识图谱的电力领域语义搜索系统的构建方法, 通过智能领域分词技术对非结构化数据进行语义知识提取, 组织并存储为知识图谱, 基于知识图谱来实现支持推理的语义搜索. 介绍了领域语义搜索系统构建流程, 并进行平台实现, 实验表明该方法查准率和召回率均有较大提升.

关键词: 电力; 电网调度; 故障恢复; 知识图谱; 语义搜索

Construction Method of Semantic Search System in Power Domain

JI Yuan, XIE Dong, ZHOU Si-Ming, WANG Xiang-Dong

(Guizhou Electric Power Grid Dispatching and Control Center, Guiyang 550002, China)

Abstract: Large amounts of data resources including relevant standards, products and technical documents, document management, fault recover records, etc. in the power domain continue to accumulate. How to fast query and search of these documents has important value for grid scheduling and fault recovery. The traditional search system is based on the key words matching, which cannot find accurate answers for query business terms. This paper designs a semantic search system for power domain. We research on word segmentation technology, knowledge graph and inference engine. The design architecture and key modules of the system are introduced, and the effectiveness of the method is evaluated by experiments.

Key words: power; grid schedule; fault recovery; knowledge graph; semantic search

电力领域不断累积大量的数据资源, 包含相关标准规范、产品和技术文档、管理文档、故障解决记录等等, 如何对这些文档进行快速查询和智能搜索, 对于电力设备运行维护和故障恢复具有重要价值. 传统的搜索系统都是基于关键词来实现, 无法支持根据业务语言来查找准确的答案[1].

知识图谱的概念首先由 Google 进行实践并倡导, 是下一代搜索引擎技术的核心. 传统的网页搜索引擎对网页直接建立索引, 提供网页的关键词检索. 知识图谱则将所有网页中的知识提取出来, 构成一个图结构, 图中节点代表实体, 边代表关系. 基于知识图谱可以支持语义搜索, 即支持通过关系来进行搜索. 比如搜索: “中国的首都”, 系统直接返回结果“北京”, 而不是返回包含“中国的首都”几个关键字的网页. 这样的搜索能准确理解用户的搜索意图, 返回精确的答案, 在电网调度中可以发挥重要作用[2].

面向电力领域非结构化信息的搜索技术目前还比较落后, 尚不能满足语义搜索的需要. 主要体现在以下三个方面: 其一是自然语言处理技术, 包括分词、词性标注和实体识别, 目前这些技术已在众多搜索引擎产品中广泛使用, 然而已有的处理技术主要面向全领域, 面向电力领域文本的自然语言处理技术还比较匮乏[7], 主要是缺少相关词库和针对电力领域的算法优化; 其二是本体和知识库构建技术, 这是语义搜索的核心技术, 由于本体构建需要大量的时间和精力, 目前采用自动化构建本体的技术创建的本体质量不高,

① 收稿时间:2015-07-27;收到修改稿时间:2015-10-19

中文本体更是少之又少; 其三是查询语句的语义化理解, 目前国内的研究主要集中于通用搜索引擎的研究, 并且仍停留在初步研究阶段.

因此本文设计实现一种基于知识图谱的电力领域语义搜索系统, 通过智能领域分词技术对非结构化数据进行分析索引, 采用知识图谱存储管理领域知识, 基于推理引擎实现语义搜索. 文章介绍了系统的设计架构和关键模块.

1　系统概述

语义搜索系统的核心是知识图谱的构建. 和传统知识库采用逻辑理论来进行知识组织不同, 知识图谱系统将知识组成一个图结构, 图的边表示实体的语义关系.

知识图谱的构建是语义搜索的基础, 计算机只有像人一样具备一定的知识才能更好的理解用户的意图. 首先, 进行知识图谱的建模, 再从遗留下来的关系型数据库中得到数据, 将它转换成 XML 文件. 然后将 XML 文件映射成 RDF 文件, 这样能够将它们导入到 RDF 数据存储系统中. 通过领域知识库维护, 语义扩展和语义解析来维护知识库. 知识图谱的技术实现方案的总体设计如图 1 所示.

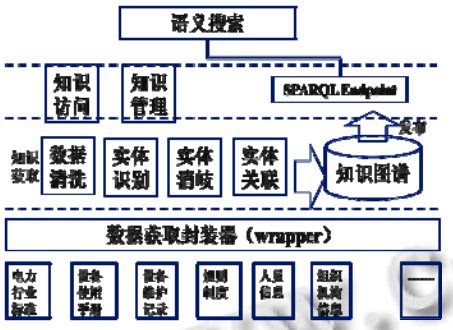

图 1　系统总体架构

知识库的建模是将领域知识映射到一个抽象的模型当中去, 以课题中的知识建模为例, 需要将电力设备维护记录中所用到的知识抽象出来, 映射到知识模型中. 在知识采集抽取阶段, 首先利用从数据库和文档中抽取到的词库, 对抽取对象的非结构化信息进行分词, 识别命名实体. 利用知识模型和实体的上下文关系, 识别实体的类型. 实体抽取出来以后会存储到实体库中去. 整个知识抽取的过程是一个基于模板的过程, 针对不同的格式, 和不同的数据源需要不同的知识抽取模板. 在语义标签知识的生成阶段, 用户可以定制编辑搜索实体的各个属性以及各个属性之间的逻辑规则关系. 然后在后台执行 SPARQL, 自动的生成语义标签, 方便用户搜索一些热门标签. 知识维护包括, 允许对现有的类, 实体和属性进行编辑. 知识管理维护工具用于向用户提供领域知识维护的接口. 电力领域专家可以通过知识管理工具添加领域知识, 维护管理知识库.

2　知识图谱建模

知识图谱是一种技术理念, 并没有统一的表示形式. 不过目前主流的技术是采用本体框架来进行知识的组织管理, 采用图数据库进行数据的存储. 为了构建知识图谱, 首要的任务是对该领域的核心概念进行建模, 形成一个本体的基础框架, 为知识的获取如导入做好准备.

本体是共享概念模型的明确的形式化规范说明, 提供相关领域的知识、概念定义和概念之间的关系, 在本文中为对搜索引擎中领域信息的规范说明. 根据现有的本体构建方法, 结合实际的领域应用, 采用 Protégé 本体构建工具的构建领域本体的过程如下[4,5]:

(1) 领域概念和关系以及相关的领域知识

对于某个特定的领域, 需要明确该领域的概念和关系. 比如概念“变压器”, 按作用可以包含子概念“升压变压器”和“降压变压器”概念之间的基本关系主要有继承关系、部分整体、实例、属性等关系.

(2) 类的定义

首先定义各个基本类, 通过父类和子类来定义类层次; 然后将所有的细化类进行合并. 一个类片段的定义如下:

```
<owl: Class rdf: ID=“升压变压器”>
    <rdfs: subClassOf>
    <owl: Class rdf: ID=”变压器”>
    </rdfs: subClassOf>
</owl: Class>
<owl: Class rdf: about=”#变压器”>
    <rdfs subClassOf>
    <owl: Class rdf: ID=”变电站设备”>
    <rdfs subClassOf>
</owl: Class>
```

其中“升压变压器“是“变压器”的一个子类，同时“变压器 “也是“变电站设备”的一个子类.

(3) 属性的定义和约束

属性的定义包括对象属性(ObjectProperty)和数据类型属性(DatatypeProperty)，对象属性把对象之间进行连接，数据类型属性将数据与对象类型值关联. 下面两个具体例子:

```
<owl: ObjectProperty rdf: ID=”#属于”>
    <rdfs:domain rdf:resource=”#变压器”/>
    <rdfs:range rdf:resource=”#变电站”/>
< / owl: ObjectProperty>
<owl: DatatypeProperty rdf: ID=“地址”>
    <rdfs: domain rdf: resource=”#变电站” / >
    <rdfs: range rdf: resource=“xsd:string”/>
    < / owl: DatatypeProperty>
```

(4) 实例的创建

以变压器的创建为例，其实例片段的 OWL 描述如下:

```
<owl: Class rdf: ID=”变压器 NO10001”>
    <rdfs: comment> An example OWL ontology < / rdfs: comment>
<owl: intersectionOf rdf: parseType=”Collection”>
<owl: Class rdf: about=”#升压变压器” / >
    <owl: Restriction>
    <owl: onProperty rdf: resource=”#属于” / >
    <owl: hasValue rdf: datatype=”&xsd; string”>***NO12 变电站< / owl: hasValue>
    < / owl: Restriction>
    < / owl: intersectionOf>
< / owl: Class>
```

如图 2 所示为创建的电力领域知识图谱的一个片段. 将电力故障恢复记录，以知识图的结构表示. 这样就可以通过关系来进行语义搜索.

3 知识获取

构建的知识图谱可以涵盖电网调度各个方面的知识，知识的来源主要有两类，即关系数据库和非关系型文本，后者又包含企业内部的文档材料和互联网网页. 下面分别介绍两类数据的知识获取方法.

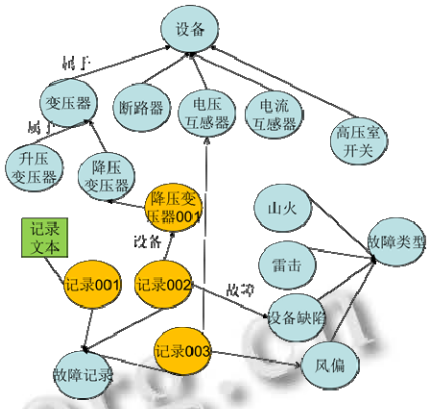

图 2 知识图谱局部图

3.1 关系型数据库到 RDF 数据库

关系型数据库可以直接与 RDF 进行映射，通过自动化的方式转换为本体数据，映射关系如图 3 所示. 数据库中的表对应 RDF 中的类，表中的每一条记录对应着 RDF 中该类下的一个实例，表中的每个字段是该实例的属性[1].

第一步: 首先，从原有的关系型数据库中提取有用信息，例如等等. 再将这些信息转换成 XML 文件. 然后会对 XML 文件中的一些非结构化的信息，例如“故障描述”进行分词.

第二步: 根据上面设计的本体中的类、实体和属性，XML 文件通过递归算法转换成 RDF 文件. 如果 XML 文件中的节点有子节点，该节点就会生成对象属性和实体，然后依次递归的遍历子节点. 否者，只是创建一个数据属性. 这个过程是采用的 Jena API 实现，输出的 RDF 文件.

当将关系型数据库转换成的 RDF 文件，可以将其导入到 RDF 数据库，即本文采用的知识存储系统.

3.2 非结构化知识提取

还有些知识隐含在非结构化数据中，需要对非结构化的数据进行自然语言处理才能提取到知识. 首先需要对非结构化的数据进行分词，然后再提取相应的实体，类和属性. 采用的方法是基于领域知识和模式的分词技术[8].

传统的分词一般是分为基于词典的分词和基于机器学习的分词方法. 基于词典的分词准确率过于依赖词典. 对于一些有歧义的词不能正确的划分，举个简单的例子. 例如，有一个句子是“北京天真好”. 词典里有“天真”，这样句子就被错误的划分成了“北京/天真/

好”. 这是因为通用词典是跨领域的, 没有对词典进行分类, 准确率得不到保证. 基于机器学习的分词方法缺少领域知识, 准确性波动较大. 因此, 本文结合领域知识和模式的方式来进行分词[3].

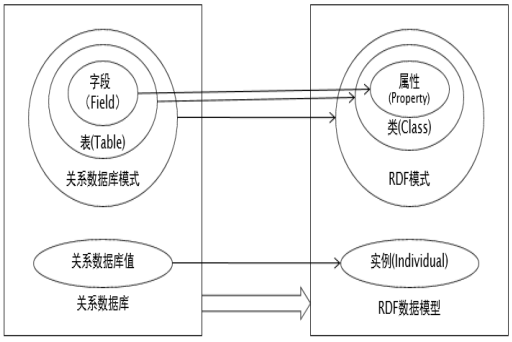

图 3　关系型数据到 RDF 模型转换

首先, 需要由用户指定非结构化短语的模式. 例如, 对于大量技术人员的工作履历进行知识提取, 识别不同人员的专业经验, 履历里存在这一的短语“***某一年在工作单位任**职务”. 需要将其分为三段“省市/公司名/职务”, 这就是提前确定好的模式. 得到模式之后, 可以借助于现有的领域知识. 这个例子的初始情况能得到完整的现有“省市”知识库, 不够完整的“公司名”和“职务名”的知识库. 根据这三个知识库就可以将非结构化的信息分成三个词, 获取的知识也可以反写到这三个知识库.

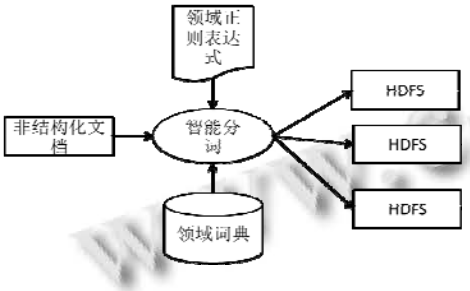

图 4　基于模式和领域知识的分词技术

3.3 知识访问与维护管理

知识库是动态更新的, 所以允许编辑类、实体和属性. 当一个新的类被创建时, 往往需要导入一类实体, 而新导入的实体和原有的实体之间的歧义性需要消除. 维护领域知识库的流程如图 5 所示, 如果新添类中的实体和新类的所有父类的实体相匹配, 这个已经存在的实体就会直接添加到这个新类中. 如果没有匹配上, 就需要为这个新类创建一个实体.

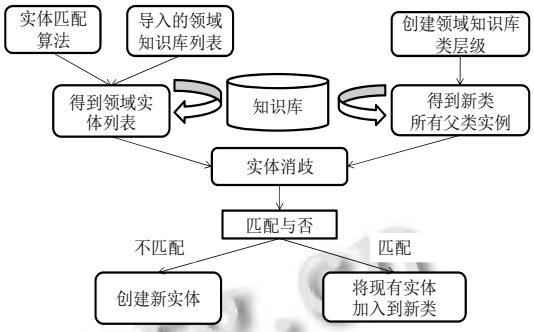

图 5　知识库维护

4　语义搜索

本文介绍的知识图谱系统存储于 RDF 数据库, 因此语义搜索采用 SPARQL 语言来表达.

```
SELECT ?b
WHERE
{ ?b a :变压器
  ?b 属于 故障记录 001
}
```

比如这个简单的语句, 可以搜索所有变压器相关的故障记录, 包含各种类型变压器相关的故障都会展现. SPARQL 执行时自动进行了推理, 即扩展“变压器”的概念到其子概念. 和关键词搜索不同, 语义搜索可以查询所有语义相近的结果.

对返回的多个结果进行排序, 需要语义度量来计算和查询术语语义上最相关的结果. 语义度量是指对不同的概念的语义距离进行计算, 得出不同的概念之间的相似度. 相似度取值范围为(0,1), 相似度取值越大, 相似度越高.

一些基本数据的定义:

$Dist(C_1, C_2)$:概念 C_1 ,C_2 的语义距离;

$depth(C)$: 概念 C 在树的结构层次中的节点深度;

$\text{weight}(C)$: 概念 C 的权值;

$\text{Sim}(C_1, C_2)$: 概念 C_1, C_2 的相似度, $\text{Sim}(C_1, C_2) \subseteq (0,1)$.

概念 C 的权值计算公式表示为: $weight(C) = 1/Wid(C) * a^{Dep(C)}, a \geq 2$

对于两个词语 w1 和 w2, 其语义距离为连接它们的最短路径上 n 条边的权值总和. 即

$Dist(C_1, C_2)=\sum_{i=1}^{n} weight_i$，两个词汇的距离越远，其相关度越小. 相似度计算公式为:

$$Sim(S_1,S_2)=a*\min(depth_{S_1},depth_{S_2})/(Dist(S_1,S_2)+a*\min(depth_{S_1},depth_{S_2}))$$

通过计算语义度量值，可以对查询结果进行排序返回结果.

5 实验评价

为了评估构建的电力领域语义搜索系统的效率，在贵州电网公司内部进行部署测试. 采样测试数据包含电网设备、电网工作人员、电网故障恢复记录、电网站点信息等总记录数 5 万条，相关文档资料 1 千篇. 分别通过关键词搜索和语义搜索来实现一些典型查询，进行案例的比较分析.

通过几个典型的案例来进行分析比较关键词搜索和语义搜索各自适用的场景，并对比二者的搜索的查准率和召回率. 查准率指返回的结果中正确结果的占比. 召回率指返回的正确结果与实际存在的正确结果的占比. 下面给出的 8 个查询条件是采用自然语言描述的，在测试时，语义搜索将转换为 RDF 的 SPARQL 语言来查询，关键词搜索将转换为包含 Like 关键词的 SQL 语句来查询，采用支持全文搜索的数据库可以将 Like 关键词执行全文索引搜索. 虽然仅给出了 8 个测试条件，但是都是有针对性选择的，同类别的查询都可以达到类似的效果.

总体来看语义搜索可以满足更多用户的搜索需求，且达到更高的查准率，可以结合领域来进行复杂查询条件的定制分析. 举例说明. 问题 1，关键词返回 5 个结果，其中 3 个为错误，材料中出现“雷击”和“故障”两次，但相互没有关系. 语义搜索返回 6 个结果均为正确结果，还查询出材料中未出现“设备”两字，但包含“电压互感器”. 问题 2，两个搜索结果一样，因为所有变压器名字都包含“变压器”关键词，所以可以找到 5 个. 如果存在设备父类、子类名字无重复词语的情况，语义搜索仍然可以正常找到所有实例. 问题 7 和 8，这种类型的搜索关键词无法找到结果，因为可以对所有人员信息提前进行语义分析，添加语义标签，因此可以支持此类型语义搜索. 这种类型关键词搜索无返回结果，因此查找率为 100%.

表 1 测试查询条件

编号	查询条件
1	雷击出现故障的设备
2	变压器数量超过 20 的站点
3	贵阳变电站数量
4	贵阳金关变电站包含哪些设备
5	出现故障的设备都位于哪些地点
6	20-30 岁之间的工程师有哪些
7	李某某的同级同事有哪些
8	具有丰富故障恢复经验的工程师

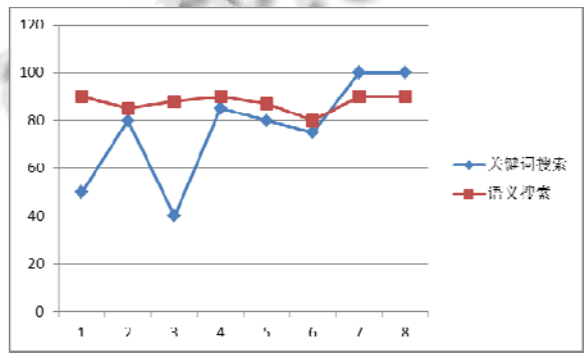

图 6 查准率

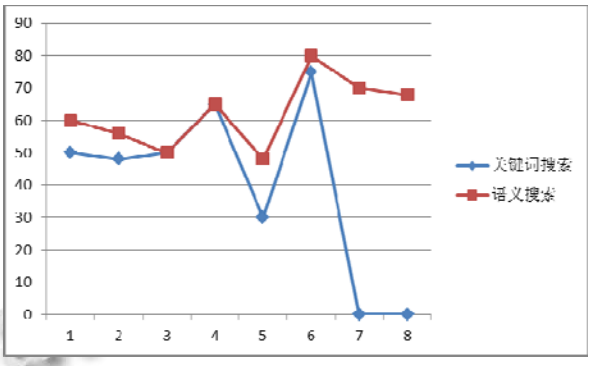

图 7 召回率

6 结语

本文针对现有搜索技术在电力领域存在的准确性和召回率问题，提出一种基于知识图谱的领域语义搜索系统框架，给出电力领域数据资源采集、知识提取、知识图谱构建，到支持领域语言的语义搜索的整个流程. 通过搭建实验平台，并采用真实数据进行评估，该框架在搜索的查全率查准率都有较大提高.

构建领域知识图谱是一项复杂工程，在各环节还存在很多技术挑战，本文只是给出了一些初步的思路和方法. 如果要实现知识高效准确的自动提取，还需要结合自然语言理解、深度学习等相关技术，也是本

文未来的研究方向.

参考文献

1 王珊,张俊,彭朝晖,等.基于本体的关系数据库语义检索.计算机科学与探索,2007,(1):59–78.

2 苏明明,宋文,基于本体的语义搜索引擎解决方案研究新进展.现代图书情报技术,2008,(11):24–28.

3 ICTCLAS 汉语分词系统.http://ictclas.org/.[2010-07-10].

4 RDF model and syntax specification. 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

5 SPARQL, http://www.w3.org/TR/rdf-sparql-query/.

6 陶涛,王栋.电网调度运行管理中存在的问题及解决措施分析.电子技术与软件工程,2014,21:179.

7 吴克河,何霞,李廷顺.基于 Lucene 构建电力企业搜索引擎分析器.电力行业信息化年会,2008.

8 车海燕,冯铁,张家晨,陈伟,李大利.面向中文自然语言文档的自动知识抽取方法.计算机研究与发展,2013,4:834–842.