基于本体的蒙古语灾害信息检索模型

苏依拉，窦保媛，吉亚图
(内蒙古工业大学信息工程学院，呼和浩特 010080)

摘 要： 由于蒙古语文本数据的匮乏以及语言分析困难等原因，蒙古语信息化处理发展缓慢. 针对该问题，利用语义网本体技术，以自然灾害为本体构建了一个面向蒙古语自然灾害信息的跨语言检索模型，实现了蒙英自然灾害信息间的跨语言检索. 同时，本文提出的跨语言检索模型具有一定的通用性，可为类似应用研究提供参考.

关键词： 灾害本体；蒙古语；WordNet；跨语言
中图分类号： TP 391 **文献标志码：** A **文章编号：** 0254－0037(2016)07－1017－07
doi： 10.11936/bjutxb2015100070

Mongolian Disaster Information Retrieval Model Based on Ontology

SU Yila, DOU Baoyuan, JI Yatu
(College of Information Engineering, Inner Mongolia University of Technology, Hohhot 010080, China)

Abstract: The development of Mongolian language information is slow due to lack of electronic text data and massive difficulty of language analysis. In order to solve the problem mentioned above, a cross language retrieval model oriented to Mongolian natural disaster information was constructed by using Semantic Web ontology technology. Model test results show that this model can achieve a better cross language retrieval effect. At the same time, the cross language retrieval model proposed in this paper has a certain commonality, which can provide reference for similar applications.

Key words: disaster ontology; Mongolian; WordNet; cross-language

由于互联网的广泛使用和自然语言处理技术的发展，本体以使用结构化的、机器和人都能理解的方式来组织和共享知识为契机，成为许多知识密集型应用的重要组成部分[1]. 本体最为著名并被广泛引用的定义是由 Gruber 提出的“本体是概念模型的明确的规范说明”[2]. 而本体学习(ontology learning)的目标是利用机器学习和统计自然语言处理等技术，自动或半自动地从已有的数据资源中获取期望的本体[3-4].

自提出基于本体学习的知识获取方法以来，英语和汉语的知识获取方法都获得迅速的发展. 目前少数民族自然语言处理仍处于知识获取方法的基础技术研究阶段，由于蒙古语编码的特性(如编码不统一)、蒙古语的使用不普遍等原因，相关的应用研究更是缺乏[5]. 内蒙古在中国属于自然灾害多发地区，目前与之相关的研究多是着眼于单一灾害，对于并发或灾害源引发系列灾害的整体状况考虑较少[6].

目前，跨语言信息检索主要有基于双语词典、机器翻译和基于平行语料库这3种方法，中间语言方法作为一种新的检索处理方式也受到了各方的关注，它是伴随检索和翻译过程协同处理思路逐步发展起来的一种方法. 而今，跨语言信息检索技术发展迅猛，Google 页面的“用户偏好”就提供了跨语言检索的功能，支持多种查询语言以及对不同语言构造网页的检索. 2012 年 Google 推出第 1 版知识图

收稿日期： 2015-10-24
基金项目： 国家自然科学基金资助项目(61363052)；内蒙古自然科学基金资助项目(2012MS0904)
作者简介： 苏依拉(1964—)，男，教授，主要从事人工智能、机器翻译方面的研究，E-mail：suyila@tsinghua.org.cn

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

谱,在学术界和工业界掀起了一股热潮[7]. 国内学者对于跨语言检索模型的研究主要集中于中英文信息检索处理,且相对较少,而对于蒙古语、汉语之间的检索研究则更少. 对于蒙汉语间的检索还处在起步阶段,内蒙古大学对蒙汉双语对齐语料库词性标注、词汇对齐和词法分析做了研究,通过对比蒙汉双语简单句子的句子成分、分析短语结构的对应关系,从而总结出蒙古语简单句结构转换为汉语对应句子结构的规律[8].

本文将语义 Web 引入蒙古语信息处理过程中,借助"本体"这一概念,利用汉语作为中间语言,构建了一个面向蒙古语自然灾害信息检索的跨语言检索模型. 通过构建蒙汉词典及中文 WordNet 数据库,将蒙古语灾害信息转换为对应的英文形式,再使用基于 WordNet 的相似度算法将该英文形式与自然灾害本体类名进行匹配,从而完成本体相关信息的查询. 本文提出的跨语言检索模型提供了一种跨语言检索的实现方法,这种具有通用性的设计模式可以为类似应用研究提供参考.

1 基于本体的蒙古语自然灾害信息检索模型

基于本体的蒙古语自然灾害信息检索模型需要对输入的蒙古语灾害信息进行多个层次的处理,具体包括对蒙汉词典的查询、中文 WordNet 汉英信息的转换、英文单词组合同本体中类的匹配和灾害本体信息的查询等.

该设计流程可分为 3 个部分,分述如下.

1) 自然灾害本体的构建

本文使用 Protégé 工具进行自然灾害本体的构建,构建好的本体将被存入数据库,便于后续过程中对本体信息进行检索.

2) 对输入的蒙古语信息的处理、转换

对于用户输入的蒙古语自然灾害信息,首先调用蒙汉数据库词典查找对应的中文信息,如果查找失败则发出错误提示,程序中止;如果查找成功则转到中文 WordNet 中查找中文信息对应的英文信息. 在中文 WordNet 中进行查询的过程中,如果查找到相关的英文信息,则进入下一个处理步骤;如果查找不到对应的英文信息,则使用分词工具对上述中文信息进行分词后再转到中文 WordNet 中进行查询,若此次查找成功则进入下一个处理步骤,否则发出错误提示,程序中止.

3) 基于 WordNet 的词汇匹配并输出查询结果

若用户输入的蒙古语自然灾害信息查找到相关的英文信息,则使用基于 WordNet 的词汇匹配算法计算该英文信息与本体中类名的相似度,选取相似度值最高的类作为匹配结果,然后调用 SparQL 对匹配到的类进行相关查询,并将查询结果返回给用户.

1.1 本体构建

在构建本体时,可以使用斯坦福大学开发的本体图形化设计工具 Protégé,结合 XML、RDF、OWL 等进行蒙古语自然灾害本体的创建.

在创建自然灾害本体时,遵守了由 Gruber 提出的本体构建基本原则[9]. 首先需要搜集各类自然灾害的详细信息,由于这类信息来源较广,此一些专业性强的领域信息获取过程更容易. 获取到自然灾害的详细信息后需要对自然灾害依据一定原则进行分类,将自然灾害按形成过程分为两大类,即突发性自然灾害和缓发性自然灾害. 在此基础上将继续根据自然灾害间的联系性继续进行划分,如突发性自然灾害又可划分为地质灾害、气象灾害等,直至将灾害类型框架建立起来. 具体的本体构建流程如图 1 所示.

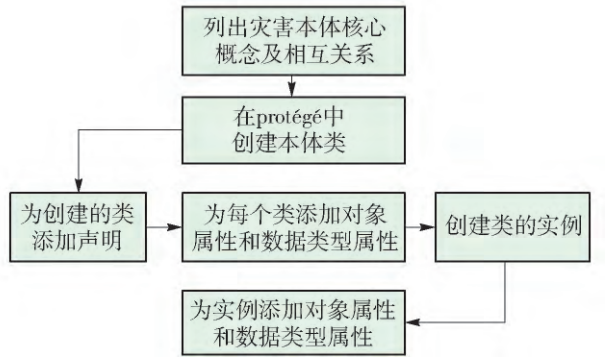

图 1 本体构建流程

Fig. 1 Flow chart of ontology construction

具体步骤如下:

1) 创建特定类名 Nature_disaster.

2) 建立自然灾害类的子类.

首先可以将灾害划分为突发性自然灾害和缓发性自然灾害,接着继续为其构建具有从属性关系的类,如突发性自然灾害又可以分为地质灾害、气象灾害等,如地震(earthquake)、崩塌(collapse)等都属于地质灾害,酸雨、干旱等属于气象灾害,依级在 Protégé 中创建各个类,得到如图 2 所示的树结构.

3) 为本体中创建的类添加声明.

通过为本体中创建的类添加声明,使概念的定义更加明确,因为某一类具体灾害不可能既属于地质灾害又属于气象灾害. 因此,突发性自然灾害下

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

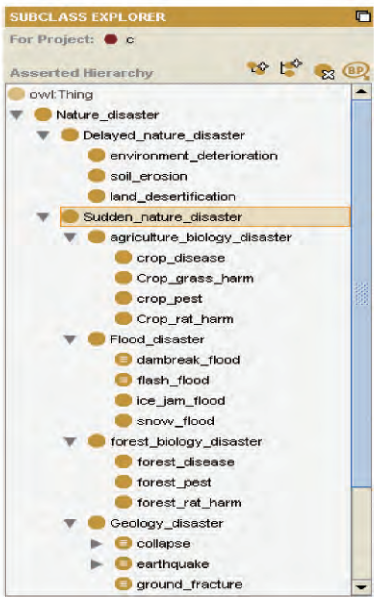

图 2　本体类层次

Fig. 2　Hierarchy figure of classes in the ontology

属一级的类之间就有了互斥关系，例如突发性自然灾害(父类)包括：旱灾、洪涝、台风、风暴潮、冻害等(子类)，若子类属于旱灾的话，则一定不会属于父类中的其他类别，因此旱灾与其他子类之间存在互斥关系. 同理也可为各个层次的兄弟类之间都添加此类声明.

上述过程主要通过建立本体类属性来实现，其中主要用到类的数据类型属性(datatype property)以及对象属性(object property). 本文以地震(earthquake)类属性的构建为例予以说明.

1) 构建 earthquake 类的对象属性

利用 Properties 标签，新建一个对象属性，重命名为 caused_by，代表被引发. 再建立一个对象属性 give_rise_to(引发)，在其右下角 Inverser 框中选择 caused_by 属性，表明它是属性 caused_by 的逆关系(owl: inverseOf).

2) 构建类的数据类型属性

利用 Properties 标签，新建一个数据类型属性 Pre(应对措施)，在 Domain(定义域)中定义该属性的主体的类是 earthquake，Ranger 选取 String 字符串类型. 通过类似的添加可以为所有需要定义的类添加 Pre 属性，对于 Pre 的具体值可以依据不同类别的特征加以描述，如为 earthquake 类的 Pre 属性可以添加如下值:

[illegible]

(就地选择开阔地避震)

[illegible]

(避开高大建筑物或构筑物)

对于上述定义好的类，可为其添加实例，这里暂定其名称为 earthquake_2.

如上所述，可以为与 earthquake 有联系的灾害类创建类似实例，这样就可以为 earthquake_2 这个实例添加谓语“give_rise_to”及宾语“collapse_1”等.

由于本文是对基于本体的蒙古语灾害信息检索模型的研究，其输入的查询内容及输出的结果均需包含蒙文信息. 因而，拟将 earthquake_2 这样的实例名改为蒙古文的形式，具体查询过程为首先找到相应实例，然后对该实例的 comment 属性值进行查询，即可得到包含蒙文的查询结果.

对于已构建好的本体，可通过 Jena 来实现对本体的后续处理，如一致性检查、本体内信息检索等. Jena 是由 Java 语言编写的本体资源框架模型处理工具，能够支持语义网的有关应用. 本体一致性检查以及本体内信息检索均可以通过 Jena 实现; Jena 具有推理功能，为信息处理过程中的推理提供支持; Jena 也支持基于 SparQL (simple protocol and RDF query language) 的本体信息查询，具体过程可以如图 3 所示.

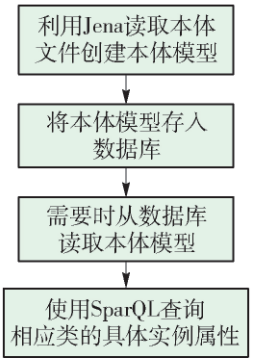

图 3　利用 Jena 处理本体模型

Fig. 3　Deal with ontology model by Jena

1.2　基于 WordNet 的词类匹配

本文研究的一个主要问题就是在蒙文输入转换为英文后如何实现与本体文件中概念的匹配，所以设计过程需要研究英文单词或词组的匹配问题. 由于 WordNet 的相关研究已经包含一些相似度的计算方法，因而拟采用 WordNet 的相似度算法获取输入信息与本体中类名的最佳匹配结果.

基于 WordNet 的语义相似度算法，主要有基于信息内容(information content, IC)的相似度和基于语义距离的相似度计算方法[9]. 基于语义距离的相似度计算可以参考由 Wu 等[10]提出的经典算法，该

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

算法考虑了 WordNet 中 2 个概念的公共父节点、深度(由于 WordNet 的层次图结构,因而具有深度一说)以及语义距离. 公式为

$$\mathrm{Sim}_{\mathrm{wup}}=\frac{2\times\mathrm{depth}(\mathrm{lso}(c_1,c_2))}{\mathrm{len}(c_1,c_2)+2\times\mathrm{depth}(\mathrm{lso}(c_1,c_2))} \tag{1}$$

式中: $\mathrm{lso}(c_1,c_2)$ 代表 c_1、c_2 两个概念的最近公共父节点; $\mathrm{depth}(\mathrm{lso}(c_1,c_2))$ 代表这个父节点的深度; $\mathrm{len}(c_1,c_2)$ 代表 2 个概念的语义距离. 由此得出: 2 个概念在最近公共父节点一致的情况下,语义距离越短,相似度越大; 概念间语义距离一定的情况下,最近公共父节点深度越大,概念划分越具体,相似度越高.

在基于 IC 的相似度计算中,主要依靠信息量的大小来衡量概念相似度. Lin 算法即为基于 IC 的相似度算法,其计算方式[11]为

$$\mathrm{Sim}_{\mathrm{Lin}}(c_1,c_2)=\frac{2\lg p(\mathrm{lso}(c_1,c_2))}{\lg p(c_1)+\lg p(c_2)}=\frac{2\mathrm{IC}(\mathrm{lso}(c_1,c_2))}{\mathrm{IC}(c_1)+\mathrm{IC}(c_2)} \tag{2}$$

式中 $\mathrm{lso}(c_1,c_2)$ 为 c_1、c_2 的最近公共父节点. 对于 $\mathrm{IC}(x)$ 的计算,本文采用 Nuno[12] 提出的基于 WordNet 本体结构的方法的信息量计算方法,具体为

$$\mathrm{IC}(c)=1-\frac{\lg(\mathrm{hypo}(c)+1)}{\lg(\max_{\mathrm{wn}})} \tag{3}$$

式中: $\mathrm{hypo}(c)$ 为概念 c 的下位词集的个数; $\max_{\mathrm{wn}}$ 为 WordNet 词集中所有名词概念的个数. 需要注意的是此处概念 c 下位词集的个数,不仅仅是概念 c 的直接下位词集的个数,而且是从概念 c 出发的直到叶子节点的所有无重复的下位词集的个数.

为减少每次实验过程中计算信息量的时间,本文通过编程预先计算所有概念的 Nuno 信息量并保存到文件中,在后续计算语义相似度的过程中遇到信息量计算时直接读取信息量文件即可.

通过使用上述信息量的计算方法,将其计算结果带入 Lin 算法的公式中,就可以得到相应的相似度计算结果.

综上所述,这 2 种计算相似度的方法,都考虑了 2 个概念的最近公共父节点,不同之处在于前者同时考虑了概念间的相互距离. 而本文采用的基于信息内容的相似度算法在计算信息量的过程中用到了节点的下位个数. 参考计算语义相似度的相关文献,语义距离、节点的深度及节点下位个数都是与相似度有关的属性,所以拟将这 3 个要素都加以考虑,这可以通过结合 Wu 等及采用 Nuno 计算信息量的 Lin 算法予以实现.

JWS(Java WordNet similarity)中已经实现了 Lin 算法,其中概念的信息量是通过统计在一个大型语料库中概念出现的概率对数的相反数获得的,如

$$\mathrm{IC}(c)=-\lg(p(c)) \tag{4}$$

式中 $p(c)$ 通过统计概念 c 在语料库中出现的概率来获取.

根据式(1),需要计算 2 个概念 (c_1,c_2) 的公共父节点 $\mathrm{lso}(c_1,c_2)$ 以及其深度 $\mathrm{depth}(\mathrm{lso}(c_1,c_2))$ 和 2 个概念的语义距离 $\mathrm{len}(c_1,c_2)$.

JWS 中采用

$$\mathrm{Sim}_{\mathrm{wup}}=\frac{2\mathrm{depth}(\mathrm{lso}(c_1,c_2))}{\mathrm{depth}(c_1)+\mathrm{depth}(c_2)} \tag{5}$$

用 2 个概念的深度和来代替式(1)中的分母,简化了运算过程. 2 种方法的结合,综合考虑了路径与信息量对于相似度的影响,理论上具有可行性. $\mathrm{Sim}_{\mathrm{wup}}$ 表示使用 Wu and Palmer 计算所得 2 个词的相似度, $\mathrm{Sim}_{\mathrm{Lin}}$ 为使用 Lin 算法所得相似度,用 $\mathrm{Sim}_{\mathrm{wn}}=k\mathrm{Sim}_{\mathrm{wup}}+(1-k)\mathrm{Sim}_{\mathrm{Lin}}$ 作为最终的相似度结果. 对于 k 的取值,通过相似度数据测试集来确定. 对于一个已有人工取值的数据测试集,可以用选定算法计算相似度,根据人工值和计算值的相关度来判断算法的优劣[13]. 通过阅读相关文献,可以看到这种判断方法为很多研究人员认可并实践. 相关性计算公式为

$$\rho_{xy}=\frac{\mathrm{Cov}(x,y)}{\sqrt{D(x)D(y)}} \tag{6}$$

式中: $\mathrm{Cov}(x,y)=\mathrm{E}((x-\mathrm{E}(x))(y-\mathrm{E}(y)))$; $D(x)$、$D(y)$ 表示方差; ρ_{xy} 的值越大表示相关性越高.

本文选用 WordNet similarity 353 数据集作为测试数据集. WordNet similarity 353 测试集分为 3 列,分别为 word1、word2 和两词之间的相似度的人工取值. 测试集中的相似度范围为 $[0.23,10]$,而 $\mathrm{Sim}_{\mathrm{wn}}$ 的范围为 $[0,1]$,所以在找出最佳的 k 值之前首先要将测试集中相似度的数值映射到 $[0,1]$,方法为

$$\mathrm{Sim}'_{353}=\frac{\mathrm{Sim}_{353}-\mathrm{Min}(\mathrm{Sim}_{353})}{\mathrm{Max}(\mathrm{Sim}_{353})-\mathrm{Min}(\mathrm{Sim}_{353})} \tag{7}$$

式中: Sim_{353} 为测试集中初始的相似度; Sim'_{353} 为映射到 $[0,1]$ 的相似度; $\mathrm{Max}(\mathrm{Sim}_{353})$、$\mathrm{Min}(\mathrm{Sim}_{353})$ 分别取 Sim_{353} 的最大值和最小值.

接着令 k 从 0.01 开始,每次循环都需计算并将 k 递增 0.01,再通过控制循环找到使得与相关系数最大的 k. 通过计算得到相关性最高的系数为

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$k=0.1$，说明该算法中 Lin 的方法起的作用更大.

随后，本文基于由 Rubenstein 和 Goodenough 提供的 RG(Rubenstein & Goodenough) 测试数据集[9]，分别采用 Wu 等、Lin 及两者结合的方法计算 RG 测试数据集相似度，然后同人工取值相比较，根据式(6) 计算相关性，具体如表 1 所示.

表 1　相关度计算结果

Table 1　Calculation result of correlation

方法	相似度
Lin	0.863 333 1
Wu and Palmer	0.797 612 98
Lin_wu	0.864 073

通过上述过程可以得知，使用基于信息量和路径结合的相似度算法可以得到更高的相关度值，该方法具有可行性. 因而可以使用它进行英文检索词与本体类名的相似度计算，实现匹配目标.

2　模型检验

在实验模型的检验过程中，本文采用自上而下的设计步骤，包含蒙古语到英语的转化流程，本体模型的存储及调用过程，基于 WordNet 的匹配算法用以实现检索关键词与本体信息的匹配，使用 SparQL 实现本体中信息的检索以及最终的处理结果展示. 检索模型的实现采用 MySQL 及 Java 作为数据库及编程语言.

2.1　实证分析

1) 文本转化

对于输入的蒙古文查询词，需要通过处理转化为英文查询词，进而与本体中的类名进行匹配. 设计过程中使用蒙科立整词输入法作为蒙古文的输入环境. 本文使用该输入法、结合 MySQL 及数据库可视化软件 Navicat 构建的蒙汉词典为如图 4 所示.

对于蒙古文查询请求，在数据库中进行蒙汉匹配，得到对应的中文词汇集合. 调用蒙汉词典数据库，以实现该处理过程.

在蒙古文转换为英文的过程中，用到了汉语作为中转语言，在蒙英翻译工具偏少的情况下，通过这种处理方式间接实现中文输入的跨语言检索.

对于中文词汇的处理结果，可以通过在中文 WordNet 数据库的相关表中找到它们对应的中文词汇信息，再根据这些信息找到对应的英文单词信息，从而将中文词汇集合转换为英文词汇集合. 根据对中文 WordNet 中表内容的分析，对带有蒙汉对照的表进行联合查询，可以得到处理后的英文单词组合.

[illegible]	洪水灾害
[illegible]	海岸带灾害
[illegible]	地面沉降
[illegible]	森林虫害
[illegible]	构造地震
[illegible]	冻害
[illegible]	海洋灾害
[illegible]	农业生物灾害
[illegible]	森林病害
⋮	⋮
[illegible]	森林鼠害
[illegible]	泥石流
[illegible]	地震
[illegible]	农作物病害

图 4　蒙汉数据库词典

Fig. 4　Mongolian-Chinese database dictionary

本文使用 wn_synset 表中的 word 字段 + ss_type 字段 + sense _ number 字段(这 3 个字段与英文 WordNet 中是对应的) 来定位英文 WordNet 中的一个同义词集. 本体中的类名由名词和形容词组成，因为在英文 WordNet 中没有定义形容词的上下义及其他关系，无法为形容词计算相似度，所以在处理过程中将形容词过滤，ss_type 字段均为 n(名词) ，即可以使用中文 WordNet 中 wn _ synset 表的 word 字段 + sense_number字段来定位英文 WordNet 中一个名词的同义词集.

如果转化后的中文词汇在中文 WordNet 中找不到相应的记录，则调用中文分词工具对该中文词汇进行分词，接着在中文 WordNet 中对分词后的结果进行查询，如果可以找到相应记录则进行后续匹配处理，否则返回空值并退出.

2) 信息匹配

对于经中文 WordNet 查询得到的英文单词集合，需要与本体中的类名进行匹配，这也是本文实现的关键性问题. 实现这一过程需要用到英文单词与英文本体类名的匹配实现，本文采用基于 WordNet 的词匹配方法，综合考虑词汇的语义距离、深度以及下义词集个数的算法来实现匹配过程.

对灾害本体中所含类，可以对数据库中取出的本体模型调用 listClasses( ) 方法迭代获取; 对于处理后带编号的英文单词集，将集合中的单词同所有类名一一进行相似度的计算，所得的相似度最大的作为匹配结果，匹配分为单个搜索词匹配与多词匹配过程. 本文以“酸雨”为例说明多词匹配过程:

首先“酸雨”经过分词并在中文 WordNet 中进

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

行多表联合查询后得出如下结果:

{雨 = [selva-1, water-6, Hyla-1, raindrop-1, monsoon-1, mack-1, monsoon-3, monsoon-2, Apus-2, thunderstorm-1, Apus-1, pluviometer-1, swift-3, swift-4, scud-1, Pluviose-1, flashing-1, flashing-2, storm-1, storm-3, storm-2, … , brashness-1, soaker-2, rainmaker-2, brashness-2, rainmaker-1, Huguenot-1, water-2, water-3, water-4, water-5, drizzle-1, soaker-1, water-1, rainforest-1, slicker-1, slicker-2, Burberry-1, slicker-3, Apodidae-1],

酸 = [fox-2, fox-3, lactobacillus-1, Oxydendrum-1, propenoate-1, fox-1, Austerlitz-1, acidimetry-1, borosilicate-1, Austerlitz-2, ribose-1, Garibaldi-1, lysine-1, tannin-1, ribonuclease-1, asphyxiator-1, …, bichromate-1, oxalacetate-1, folacin-1, benzoate-1, Narcan-1, fulminate-1, lypressin-1, clabber-1, nuclease-1, aerator-1, lime-1, lime-3, lime-2, ketoprofen-1, lime-5, lime-4, lime-6, sourness-3, sourness-1, bluestone-1, sourness-2]}

再将上述结果与本体中所有类名计算相似度得出如下结果:

{acid_rain = 0. 695 738 646 197 844 9, torrent_rain = 0. 631 658 119 005 674 3,

thunder_storm_wind = 0. 610 877 292 153 405 5, sand_dust_storm = 0. 598 911 659 125 960 1,

typhoon_storm_tide = 0. 589 161 739 304 375 7, …, beach_erosion = 0. 229 200 143 317 986 63, red_tide = 0. 223 953 777 856 392 26,

drought = 0. 215 051 739 750 191 54, soil_salinization = 0. 208 536 674 373 103 5}

由上述结果可知“酸雨”与 acid_rain 的相似度最高,所以取 acid_rain 作为“酸雨”的匹配值.

在实现过程中,可以使用同样的代码对单词或者多词进行处理,因为单词的分词结果还是单词本身,不会影响后续的相似度的计算.

2.2 结果呈现

依据本文阐述的方法找到搜索词对应的本体类名后,可以通过 SparQL 查询得到灾害的应对属性及其引发的灾害类型.

如匹配后得到的灾害类型为 earthquake,可通过对其实例的相关查询最终可以得 earthquake 的应对信息及其引发的灾害类型,实例的查询通过 OntClass(本体类)的 listInstances( ) 方法实现.

如在找到 earthquake 类的实例 earthquake_2 后,利用:

SELECT ? pre WHERE {base: earthquake_2 base: Pre ? pre}

可以获得灾害的应对信息,然后借助:

SELECT ? caused_by_earthquake_2 WHERE {base: earthquake_2 base: give_rise_to ? caused_By_earthquake_2}

获得 earthquake 可以引发的灾害类型实例.

由于实例中并不含有蒙文信息,故对于得到的灾害实例,需要输出其 comment 属性的值,因为该属性值包含了希望得到的结果(所需的蒙文信息). 所以实际查询为得到地震引发的灾害实例,然后对这些实例进行 comment 属性的查询,以查询地面沉降相关的蒙文信息为例,相应的查询语句为: SELECT ? id WHERE { base: land_subsidence_1 rdfs: comment ? id},查询结果如图 5 所示,从图中可以发现关于地面沉降的应对措施以及其可能诱发的自然灾害,如滑坡、坍塌等.

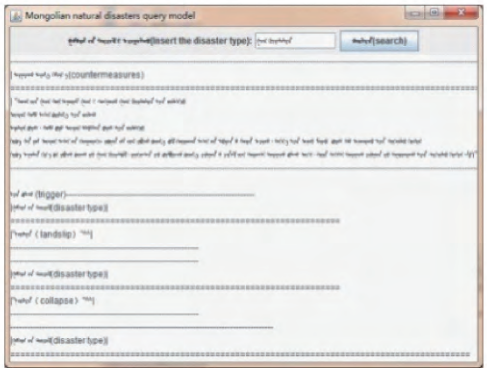

图 5 查询结果展示

Fig. 5 Show of query results

3 结论

1) 将自然灾害间的关系纳入本体构建过程,利用机器学习和统计自然语言处理等技术,使用斯坦福大学开发的本体图形化设计工具 Protégé 构建灾害本体,通过为本体中创建的类添加声明,使概念的定义更加明确.

2) 在基于 WordNet 的词类匹配过程中,使用基于信息量和路径结合的相似度算法可以得到更高的相关度值. 因而,可以使用它进行英文检索词与本体类名的相似度计算,实现匹配目标.

3) 在蒙古文转换为英文的过程中,用到了汉语作为中转语言,在蒙英翻译工具偏少的情况下,通过这种处理方式间接实现了中文输入的跨语言检索.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献：

[1] VALARAKOS A G，PALIOURAS G，KARKALETSIS V，et al. Enhancing ontological knowledge through ontology population and enrichment [M]. Berlin: Springer Engineering Knowledge in the Age of the Semantic Web，2004: 144-156.

[2] BAKHSHADEH M，MORAIS A，CAETANO A，et al. Ontology transformation of enterprise architecture models [M]. Berlin: Springer Technological Innovation for Collective Awareness Systems，2014: 55-62.

[3] 杜小勇，李曼，王珊. 本体学习研究综述 [J]. 软件学报，2006，17(9): 1837-1847.
DU X Y，LI M，WANG S. A survey on ontology learning research [J]. Journal of Software，2006，17(9): 1837-1847. (in Chinese)

[4] SURE Y，SCHNURR H P，STUDER R，et al. Knowledge processes and ontologies [J]. IEEE Intelligent Systems，2001，16(1): 26-34.

[5] 赵小兵，邱莉榕，赵铁军. 多民族语言本体知识库的构建技术 [J]. 中文信息学报，2011，25(4): 71-74.
ZHAO X B，QIU L R，ZHAO T J. Construction technology of ontology knowledge base in multiple minority languages [J]. Journal of Chinese Information Processing，2011，25(4): 71-74. (in Chinese)

[6] 赵兰. 我国政府在自然灾害应急管理中的职能问题研究 [D]. 成都: 电子科技大学，2011.
ZHAO L. Research on the function of the government in the emergency management of natural disasters [D]. Chengdu: University of Electronic Science and Technology，2011. (in Chinese)

[7] 王新才，丁家友. 大数据知识图谱: 概念、特征、应用与影响 [J]. 情报科学，2013，31(9): 10-14.
WANG X C，DING J Y. Mapping knowledge domain of big data: concept，feature，application and impact [J]. Information Science，2013，31(9): 10-14. (in Chinese)

[8] 阿拉塔. 蒙汉双语简单句结构比较研究 [D]. 内蒙古: 内蒙古大学，2013.
A L T. Comparative study of mongolian Chinese bilingual simple sentences structure [D]. Inner Mongolia: Inner Mongolia University，2013. (in Chinese)

[9] 王桐，王磊，吴吉义，等. WordNet 中的综合概念语义相似度计算方法 [J]. 北京邮电大学学报，2013，36(2): 100-106.
WANG T，WANG L，WU J Y，et al. Semantic similarity calculation method of comprehensive concept in WordNet [J]. Journal of Beijing University of Posts and Tele Communications，2013，36(2): 100-106. (in Chinese)

[10] WU Z，PALMER M. Verb semantics and lexical selection [C] // Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. New Mexico: ACL，1994: 133-139.

[11] FELLBAUM C，MILLER G. Combining local context and WordNet similarity for word sense identification [M] // WordNet: An electronic lexical database. Berlin: MIT Press，1998: 265-283.

[12] MAEDCHE A，MOTIK B，Silva N，et al. Mafra-a mapping framework for distributed ontologies [M] // Knowledge engineering and knowledge management: ontologies and the semantic Web. Berlin: Springer，2002: 235-250.

[13] 张凯勇. 基于 WordNet 的词语及短文本语义相似度算法研究 [D]. 长春: 吉林大学，2011.
ZHANG K Y. Research on semantic similarity between words and between short texts based on WordNet [D]. Changchun: Jilin University，2011. (in Chinese)

(责任编辑　吕小红)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net