



滤模块对互信息矩阵进行处理,将存在于互信息矩阵当中而不存在于领域本体当中的词对过滤掉,得到关键字集。这里的关键字集是用于用户最终进行检索时所用的关键字的集合。为了更好的对领域本体进行优化,我们利用语义发现模块对关键字集进行分析,新的语义词对,得以新语义集。在参考本体分类模块的作用下,根据参考本体当中所保存的不同内容,分别采用本体补充模块,本体净化模块或本体修正模块对参考本体当中的内容进行处理,最终达到对领域本体的修改,努力使之保存为更准确的语义关系。

### 1.3 相关描述

为了更好的对上述模型进行描述,我们对模型当中的各组件进行了相应的形式化描述。

#### (1)互信息矩阵的形式化描述

首先定义关键词向量:  $K = (key_0, key_1, key_2, \dots, key_n)$

互信息矩阵可定义为:  $M \cdot M = (M_0, M_1, \dots, M_n)$

其中,  $M_i = (M_{i_0}, M_{i_1}, \dots, M_{i_n})$  表示的是两个关键字之间的互信息量。  $M_{ij}$  计算公式为:

$$M_{ij} = MI(KEY_i, KEY_j) =$$

$$I(key_i, key_j) = \log \left( \frac{p(key_i, key_j)}{p(key_i)p(key_j)} \right) =$$

$$\log \frac{\frac{c(key_i, key_j)}{N}}{\frac{c(key_i)}{N} \times \frac{c(key_j)}{N}} = \log \frac{c(key_i, key_j) \times N}{c(key_i) \times c(key_j)}$$

根据最大似然估计,在语料规模足够大的情况下,可以认为单词出现的概率为其出现的次数。其中  $c(key_i, key_j)$  表示单词  $key_i, key_j$  有序同现的次数,  $c(key_i)$  表示关键词  $key_i$  出现的次数,  $c(key_j)$  表示关键词  $key_j$  出现的次数,  $N$  为文档库中所有单词的个数<sup>[3]</sup>。

#### (2)领域本体的形式化描述

$$DO = (C, R, A, I, P)$$

其中,  $C$  表示的是类 (classes) 或概念 (concepts): 指任何事务,如工作描述、功能等。从语义上讲,它表示的是对象的集合,包括概念的名称,与其它概念之间的关系集合,以及用自然语言对概念的描述。

$R = \{part-of, kind-of, instance-of, attribute-of\}$ , 表示的是词对之间的关系 (relations)。

$A$  表示的是公理 (axioms): 代表永真断言,如概念乙属于概念甲的范围。

$I$  表示的是实例 (instances): 代表元素。从语义上讲实例表示的就是对象。

$P$  表示的是属性 (Property): 代表对概念性或性质的

描述。

#### (3)参考本体的形式化描述

$$RefO = ( \{ NewKey, FomKey, Type, Desc, Num \} )$$

其中,  $NewKey = \{ Key_1, Key_2, \dots, Key_n \}$  表示的是用户反馈的关键字。此类关键字分为两类,一类是在领域本体当中没有的,另一类是在领域本体当中存在错误关系的关键词。

$FomKey = \{ Key_1', Key_2', \dots, Key_n' \}$  表示的是与用户反馈的关键词相关联的且已存在于领域本体当中的关键词。

$Type = \{ 0, 1, 2 \}$  用于表示用户建议对参考本体当中关于该词对的操作,因词对关系缺失建议添加的为 0 因词对关系错误建议修改的为 1 建议删除词对关系的为 2。

$Desc = \{ D | Null \}$  用于表示当参考本体当中存在错误词对时,用户词对关系进行建议性的描述  $D$ , 该字段可为空  $Null$ 。

$Num = \{ D \}$  用于表示用户要求对参考本体当中当前词对进行修改的次数。

## 2 保证领域本体正确性问题

在  $MO\_CRM$  模型中,我们拟解决二个关键问题:

(1)如何从互信息矩阵中发现新的语义信息:互信息矩阵存放的是通过对训练文档集进行互信息计算之后,所得到的关键词对之间的相关联的度量。但是这种度量不能准确判断矩阵当中具有互信息量的两个关键词对是否在语义上具有关联性。这一问题的解决模块,主要是协同过滤算法 (CFA) 与语义发现算法 (SDA) 共同作用下完成的。这两个算法的详细描述,请参阅文献 [4]。

(2)如何处理领域本体自身关系描述不准确的情况:本体虽然具有准确描述语义关系的特点,但是,由于语言的使用习惯与语言自身都是发展的,因此本体自身的描述在某些时候可能会出现偏差或是过时的情况。而这些偏差及其解决方法可以归纳为三种情况:

(i) 针对本体当中词对之间的描述存在错误的情况,我们将对这种关系描述进行修正。在本系统当中,采用本体修正算法 (ORA) 进行实现。

(ii) 针对本体当中存在错误关系的词对,即词对之间无关的词对,进行关系删除。在本系统当中,采用本体净化算法 (OPA) 进行实现。

(iii) 针对本体当中本应有关系的词对之间缺失相应的关系描述的情况,使用本体补充算法 (OPA) 进行处

理。

为了更好的区别应该采用上述三种手段当中的哪种,是根据参考本体的类型决定的。因此,设置参考本体分类算法(ROCA),对两种情况进行分类处理。

### 2.1 参考本体分类算法

参考本体的构成,主要是来源于用户对检索结果的反馈。当检索结果与用户的意图不相匹配时,用户可以通过“用户反馈模块”构造参考本体,进而实现对领域本体进行更正。通过对参考本体的形式化描述分析发现,需要根据参考本体的内容对后续的不同算法进行调用。相关伪代码为:

```
Public class ReD
{
    Public void ROCA()
    {
        声明并实例化 ReD对象 refd
        If(当参考本体当中的词对为领域本体词对缺失型,即 refd.Type= 0&& 用户有对参考本体通过用户反馈对词对进行描述,即 refd.Des! = null)
        {
            调用 OCA 算法,传入参考本体当中的词对;
        }
        If(当参考本体当中的词对为领域本体词对错误型建议修改词对,即 refd.Type= 1)
        {
            调用 ORA 算法,传入参考本体当中的词对以及对词对关系的描述信息;
        }
        If(当参考本体当中的词对为领域本体词对错误型建议删除词对,即 refd.Type= 2)
        {
            调用 OPA 算法,传入参考本体当中的词对;
        }
    }
}
```

### 2.2 本体净化算法

本体更新缓慢这一特点导致了本体所制定的规则有可能会跟不上语言发展的脚步,从而导致本体自身对语义表达出现偏差。基于这个原因,提出了“本体净化”这一理论,用于对本体自身所存在的错误规则进行删除,使本体的语义表达更为准确。

具体的本体净化模块的算法为:

Step1: 读取 ReD 集中的关键字 词对所对应的信息;

Step2 在 ReD 集中去检索当前读取出的词对是否存在;

Step3 判断当前词对所对应的操作类型 Type 属性是否为 2 如果 Type 值为 2 则跳至 Step4 否则退当前算法;

Step4 读取该词对出现的次数 Num 属性,如果 Num 属性值大于某阈值,则通知专家评审。当专家评审通过,跳至 Step5 否则退出算法;

Step5 删除该词对在本体当中的关系,得到新领域本体 DamO'。

### 2.3 本体补充算法

为了弥补手工创建的领域本体天生的不足,提出了本体补充模块。目的是将在原领域本体当中没有的词对添加到领域本体当中,使本体做到更完善。

为了对领域本体进行补充,首先要对新语义集当中的关键字是否在原领域本体当中存在进行查询。这里,使用本体查询语言 RDQL。这种查询语言可以将输入的查询检索词或是自然语言问句转换为 RDQL query 表达式,然后查询引擎便会执行这个 RDQL query 表达式,运行完毕,返回相应的实例 instance<sup>[6]</sup>。

根据新的语义集的各种情况,设定了 4 个处理规则:

设得到的新语义集中包括有 {A, B} 两个关键词对,利用本体查询语言 (RDQL) 我们规定:

规则 1: 若 (RDQL(A) = TRUE & RDQL(B) = TRUE), 则: !OTA(A, B)。

规则 2 若 ((RDQL(A)! = TRUE & RDQL(B) = TRUE)), 则: OTA(A)。

规则 3 若 ((RDQL(A) = TRUE & RDQL(B)! = TRUE)), 则: OTA(B)。

规则 4 若 ((RDQL(A)! = TRUE & RDQL(B)! = TRUE)), 则: OTA(A, B)。

Input 从 SDA 处得到的 NewSemanticSet 由用户反馈回来形成的 ReD。

Output 领域本体 DamO。

Procedure

Step1: 判断由 ReD 集组成的词对当中的 Flag 标志的值是否为 1。

Step2 根据之前所规定的 4 个规则,判断是否进行补充操作。

## Step3

Case规则 1: 说明 A 与 B 之间不存在任何的关系, 不用进行本体补充操作, 直接舍弃。

Case规则 2 对关键字 B 进行推理, 找出关键字 B 在领域本体当中的位置, 在领域本体当中将 A 置于 B 的父结点的下一级。

Case规则 3 对关键字 A 进行推理, 找出关键字 A 在领域本体当中的位置, 在领域本体当中将 B 置于 A 的父结点的下一级。

Case规则 4 将 A 和 B 两个关键字置为根结点的两个叶子结点。

## 2.4 本体修正算法

领域本体当中保存的是词与词之间的关系, 可以对词与词之间进行推理计算<sup>[7]</sup>。但如果词与词之间的关系描述有偏差甚至是错误的, 那么将会直接导致本体的失效。因此, 当我们一旦发现了词与词之间的这种错误的关系, 根据用户的反馈信息, 在此基础上, 我们可以结合专家意见, 对这些反馈信息进行整理, 然后作用于领域本体当中, 使本体更为准确与受用。

具体的本体修正算法描述为:

(1) 读取 ReO 集中的关键字 词对所对应的信息;

(2) 专家对 ReO 集中 Desc 字段的描述是否赞同, 如果赞同, 则转到 4 否则转到 3

(3) 写入专家根据用户反馈的意见所进行的领域相关的意见;

(4) 利用 Protege 工具对领域本体当中对应词对进行修正, 得到新领域本体 DomO'。

## 3 实验

为了便于与传统的基于关键字匹配的信息检索和基于概念的信息检索进行对比, 本文采用信息检索系统的一般评价标准, 利用查准率和召回率来定量分析 QSEBMI 较传统的问句扩展系统的优劣, 并用 F 量度 (F-measure, Van Rijsbergen, 1979) 来综合精确率与召回率指标。

关于查准率、召回率以及 F 量度的计算公式为:

$$precision = \frac{|A \cap R|}{|R|}, recall = \frac{|A \cap R|}{|A|}$$

$$F = 2 * (recall * precision) / (recall + precision)$$

A 表示信息检索系统获取的数据记录的集合; R 表示数据全集中所有与用户查询相关的数据记录的集合。

本文的实验主要针对特定的语料库进行检索测试。我们用 Spider 从网上获取了计算机类文档集作为实验

数据, 共计 200 余篇, 文章中语言为中文, 且长度不等。我们手动准备了 20 余条真实的查询问句, 针对不同的计算机领域进行查询。在权值的计算中参数取值如下:  $\alpha = 0.1, \beta = 0.1$ 。分别将本文所提出的 MO\_CRM 模型与 QSEBMI 模型及不进行查询词扩展的方法进行比较, 将其得出的查全率和召回率以及其 F 量度值进行比较, 如图 2 所示:

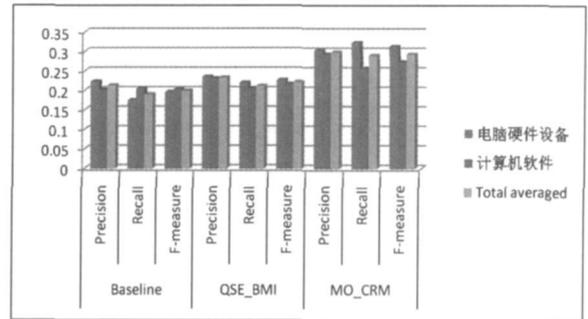


图 2 实验结果比较图

由图 2 的实验结果可知, MO\_CRM 模型通过引入“互信息”与“本体”并使用了改进算法后, 较传统的问句扩展系统和在 QSE\_BMI 系统在查准率、召回率和 F 量度上都得到了一定的改善。

## 4 结束语

本文提出了一种基于互信息与本体相结合的协同检索模型。该模型将互信息与领域本体相结合, 利用互信息与本体在本质上的互补性, 从而在精确率、召回率和 F 量度上都优于在文献 [3] 当中的 QSE\_BMI 模型, 性能都有所改善。实现用户的反馈信息采集与用户兴趣模型的自动化建立方面将是进行进一步研究的方向。

## 参考文献:

- [1] 袁占亭, 张爱民, 张秋余. 基于概念的 Web 信息检索 [J]. 计算机工程与应用, 2003, 39(36): 173-175.
- [2] 邵兵, 关毅, 王强, 等. 第二届全国学生计算语言学研讨会论文集 [C]. 哈尔滨: 哈尔滨工业大学出版社, 2004.
- [3] 夏磊, 周竹荣. 基于互信息的问句语义扩展研究 [J]. 计算机工程与设计, 2008, 29(1): 163-166.
- [4] 周竹荣, 邱玉辉, 夏磊. 基于互信息和本体的协同检索模型的研究 [J]. 计算机科学, 2008, 35(4): 165-168.
- [5] 许盛中, 蔡乐才. 基于本体的图书智能检索系统的模型研究 [J]. 四川理工学院学报: 自然科学版, 2009, 22(5): 55-57.
- [6] David V. Allet, M. iriam Fernández, Pab b Castells. An Onto-

gy-Based Information Retrieval Model [ C ]. The Semantic Web Research and Applications, Second European Semantic Web Conference, 2005, 3532-455-470.

[ 7 ] 王成敏. 基于规则的汉语名词短语自动识别方法研究 [ J ]. 四川理工学院学报: 自然科学版, 2009, 22( 2 ): 55-57.

## Research on the Improve Model of MO\_CRM

XIA Lei<sup>1</sup>, HE Min<sup>1</sup>

( 1. Department of Computer Science, Chengdu Neusoft Institute of Information Technology, Chengdu 611844, China )

**Abstract** For users on internet used to search information by several simple words, we try to satisfy them by expanding those words. We build the MO\_CRM by combine them mutual information and ontology. But as deep research into the model, we improve the way of revise and complementary of the domain ontology and get the MO\_CRM. As a result, a higher chance of recall and precision by MO\_CRM is obtained.

**Key words** domain ontology, mutual information, reference ontology, collaboration

(上接第 564 页)

### 参考文献:

- [ 1 ] 冯登国. 国内外密码学研究现状与发展趋势 [ J ]. 通信学报, 2002, 23( 5 ): 18-26
- [ 2 ] 于学江. 对称密码体制及其算法研究 [ J ]. 齐齐哈尔大学学报, 2007, 23( 6 ): 38-40
- [ 3 ] 吴昌银, 岳青松. AES 安全性及其影响研究 [ J ]. 信息安全与通信保密, 2006, 32( 11 ): 77-79.
- [ 4 ] 胡向东. 应用密码学 [ M ]. 北京: 电子工业出版社, 2006

- [ 5 ] 谢扬, 王金凤. 浅谈公钥密码体制 [ J ]. 计算机科学, 2008, 35( 4 ): 157-158
- [ 6 ] 张焕国, 冯秀涛. 演化密码与 DES 密码的演化设计 [ J ]. 通信学报, 2002, 23( 5 ): 29-32
- [ 7 ] 王宏杰. RSA 算法、DES 算法的特点分析及结合 [ J ]. 天津科技, 2005, 19( 4 ): 65-69

## Research of Asymmetric Encryption Technology

ZHOU Xiande, ZHAO Fei, ZENG Deming

( Department of Information Engineering, Luzhou Vocational and Technical College, Luzhou 646005, China )

**Abstract** With the rapid development of network technology, data encryption makes information safe. This paper asymmetric encryption algorithm is studied to improve the security of network information, reducing the risk of cyber communications. The paper firstly introduced the basic concepts of cryptography, then discussed and analyzed the characteristics of asymmetric encryption and weaknesses, finally with the introduction of specific cases by non-symmetric encryption data used in encryption and hybrid encryption system showed that this research can be more convincing and realistic significance.

**Key words** data encryption, non-symmetric encryption, Hybrid Encryption