文章编号：1003-0077(2016)04-0037-07

专利中基于语义角色的术语相似度计算方法

姜利雪，季 铎，蔡东风

（沈阳航空航天大学 知识工程研究中心，辽宁 沈阳 110136）

摘 要：术语是由一个到多个单词按照某种语义角色组合而成的，传统的基于统计的相似度计算方法，将术语看作一个基本单元来进行计算，忽略了术语内部的语义角色，且对于上下文信息不丰富的术语，无法利用统计的方法取得理想的效果；基于语义资源的相似度计算方法，所涵盖的词语有限，因此不包含在语义资源中的术语便无法计算相似度。针对这些问题，该文针对专利提出了基于语义角色的术语相似度计算方法，该方法弥补了传统方法的不足。该文对术语内部的单词进行语义角色标注，通过共享最近邻方法计算单词的相似度，然后根据不同的语义角色，利用单词相似度来计算术语相似度。实验表明，该方法与传统方法相比，取得了较好的效果。

关键词：术语；内部语义角色；共享最近邻；术语相似度；专利文本

中图分类号：TP391 **文献标识码**：A

Measuring Term Similarity Based on Internal Semantic Role in Patent Text

JIANG Lixue, JI Duo, CAI Dongfeng

(Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang, Liaoning 110136, China)

Abstract: The Chinese term is composed of one or multiple words with certain semantic roles. The traditional similarity calculation methods based on statistics, which regard the term as a basic unit for similarity computation, ignore the semantic roles inside a term. This paper presented a method for computing similarity of Chinese terms based on the internal semantic roles, i. e. calculating term similarity according to the different semantic roles assigned to them automatically. Experiments show that the proposed similarity calculation method achieves better results than traditional methods.

Key words: term; internal semantic roles; shared nearest neighbor; term similarity; patent text

1 引言

随着科学技术的不断发展，各种专利不断涌现，术语作为科技发明成果的载体也随之增多。如何能够在海量的术语中，正确的计算术语之间的相似度，准确的找到相似的术语，这在术语的理解，同义词扩展查询，术语翻译等领域起着重要作用。从语言学的观点看术语，单词是术语的构成材料，术语则是由这些构成材料形成的产品，因此可以说，一切术语都是由单词构成[1]。那么，术语的相似度可由单词的相似度得到。单词相似度的计算方法有两种：基于统计的词语相似度计算方法和基于语义资源的词语相似度计算方法。

基于统计的词语相似度计算方法，一般是利用词语的上下文环境作为词语的特征来计算相似度。基于统计的词语相似度计算方法历史悠久，早在 1993 年 Dagan I[2] 就提出在大规模语料库中，使用词的上下文特征可以较好的计算词语之间的相似度。后人的很多研究都是针对这种相似度计算方法进行研究和改进，2003 年 E Terra[3] 对不同长度的上下文进行了实验，实验表明，上下文的尺寸对相似度结果具有很大的影响，上下文窗口越长相似度计算的准确率越低。2009 年于水等[4] 引入模糊数学中隶属函数的概念计算词语上下文信息的模糊重要度，提出了一种基于语境的词语相似度计算方法。

基于统计的相似度计算方法，是一种无监督的机器学习方法，完全根据所选择的语料进行相似度

收稿日期：2014-08-20 定稿日期：2015-05-08

基金项目：国家“十二五”科技支撑计划项目(2012BAH14F00)

计算，计算结果具有客观性，且无需大量人力编写规则等，其通用性强。但是，该方法却受限于所使用的语料库，这样难以避免由于数据稀疏问题所导致的词语相似度计算不准确的问题[4]，而且该方法所使用的语料大都是通用领域的文本，所使用的单词专业性不强，所以在大规模语料中，其上下文可以很好的表达单词，相似度计算结果也较为理想。但是对于专利文本，其用词较为单一，而且对于新产生的术语，其上下文信息不丰富，所以利用基于统计的方法来计算术语相似度，并不能得到很好的效果。

基于语义资源的相似度计算方法，主要依靠语义资源库，由于考虑到了词语的语义信息，相似度计算更具准确性。现在比较常用的语义资源有WordNet、维基百科、《同义词词林》和《知网》等，现在许多语义相似度计算方法都是基于这些语义资源所实现的。1999年，王斌[5]根据《同义词词林-扩展版》将所有的词组织在一棵或几棵树状的层次结构中这一特点，利用两个节点之间的路径长度作为两个概念的语义距离的一种度量。2002年刘群[6]首次利用《知网》来进行语义相似度计算，他提出了把两个词语之间的相似度问题归结到两个概念之间的相似度问题，概念相似度计算归结到义原的相似度计算。2008年，丁林林[7]通过引入义原在义原树中的层次信息和知网中的语义框架等信息对概念间的相似度、相关度计算方法进行了改进。2010年田久乐[8]根据《同义词词林-扩展版》的编码及结构特点，提出了一种基于同义词词林的词语相似度计算方法。2013年Pilehvar[9]把WordNet表示成是一个有向图，利用RankPage算法在“WordNet”中进行随机行走，生成语义签名的方法，以此来计算两个单词之间的相似度。

基于语义资源的相似度计算方法，简单有效，也比较直观，易于理解，但是该方法得到的结果受人的主观意识影响较大，不能反映客观事实[6]，而且这些语义资源库多数是由人为编写，所涵盖的词语具有一定的局限性，因此不包含在语义资源中的词语，便无法计算相似度。相比于普通的词语，术语具有专业性，许多术语并没有涵盖在语义资源中，因此，基于语义资源的相似度计算方法也并不能很好的计算术语的相似度。

考虑到上述问题，针对专利文本，本文提出了一种基于内部语义角色的术语相似度计算方法，在对术语进行分词之后，对每条术语内部的单词进行语义角色标注，然后通过共享最近邻的方法计算各个单词之间的相似度，最后两个术语的相似度即为每类语义角色内的单词的相似度的加权求和。实验表明，利用本文所提出的方法来计算术语相似度，相比于传统的基于统计的和基于语义资源方法，效果有明显的提高。

本文所做的工作有：①提出了一种基于共享最近邻的单词相似度计算方法，用来弥补基于统计的单词相似度计算方法由于数据稀疏所导致的相似度计算不准确的问题；②根据前人对术语的研究和词组型术语的特点，总结了12种语义角色，对词组型术语内部进行语义角色标注；③提出了基于内部语义角色的术语相似度计算方法，即对词组型术语内部单词进行语义角色标注，根据语义角色，利用单词的相似度来计算术语的相似度。

2 研究背景

2.1 术语

术语是科技文献中用来表达专业词语的一种词汇，集中承载着特定领域的核心知识，对于科技信息的传播与交流有着重要的作用。本文对66 905篇已经经过分词的汉语专利文本进行统计，每篇专利文本的摘要和权利要求部分的平均长度为812个单词，平均包含59个术语。在所统计的术语中，各个长度的术语所占的比例如表1所示。

表1 各长度术语所占的比例

长度	1	2	3	4	5	大于5
所占比例/%	20	38	23	10	4	5

由表1可以看出，术语长度多在1—5个单词之间，长度为1的术语仅包含一个单词为单词型术语，长度大于2的为词组型术语，词组型术语占据了整个术语的80%。

2.2 术语语义角色体系

由2.1节可知，词语型术语占据了整个术语系统的80%左右，所以若能对词组型术语进行深入的研究，对我们了解术语有很大的帮助。词组型术语由多个单词构成，各个单词之间存在一定的语义角色，若对术语内部进行语义角色标注，对术语的分析有很好的帮助作用。

《知网》定义了16种义原关系和76种动态角色，清华大学从《知网》中定义的动态角色中归纳总

结了 59 种义原关系并实现了一个语义分析系统，并将该系统应用到语音识别中，取得了很好的效果。陈小芳根据术语的特点定义了 14 种语义关系，并将其运用到机器翻译中，同样取得了很好的效果[10]。

本文在陈小芳等[10]研究的启发下，将语义信息运用到术语相似度计算中。我们根据已有的研究和词组型术语中单词的组成特点，定义了 12 类语义角色，包括否定、外观、类别、方式、技术、性能、作用、使用者、关系、受事、材料和中心词。如表 2 所示。

表 2　语义角色类、释义及举例

语义角色	标示符	顺序标号	释义	举例（标示符【】）
否定	FD	1	否定	FD【非/b】易/ad 失/v 性/n
外观	WG	2	表示事物的形状，样式，大小等	WG【I/x 型/k】钮扣/n
类别	LB	3	所属领域，可用于区分	LB【机械/n】锁/v
方式	FS	4	事物的工作方式	FS【气动/b 式/k】马达/nr
技术	JS	5	采用的技术	JS【无/v 菌/n】接种/vn 器/n
性能	XN	6	事物工作能效，特性等	XN【高效/b】减/v 振/vg 垫/v
作用	ZY	7	事物的作用	ZY【减/v 振/vg】垫/v
使用者	SY	8	某事物所属方，可在其后添加“用”	SY【饮品/n（用）】包装物/n
关系	GX	9	主要是方位关系和方向关系	GX【内/f】导管/n
受事者	SS	10	充当“被改变”的实体	SS【种子/n】点播/v 机/ng
材料	CL	11	事件发生或进行所依赖的材料	CL【铜/n】齿轮/n
中心词	ZX	12	术语所表达的对象	弹性/n ZX【凹槽/n】

本文对含有语义角色的 18 744 条词组型术语进行统计，由于每条术语都包含中心词，除中心词以外的其余 11 类语义角色所占的比例如图 1 所示。可以看出，“ZY”、“SY”、“SS”和“FS”这四类语义角色所占的比例较大，其余七类所占的比例很小。

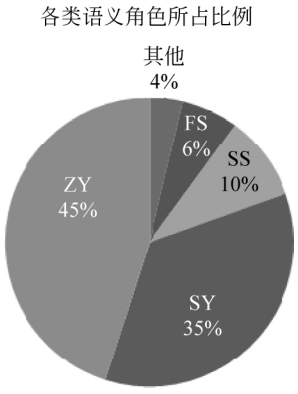

图 1　语义角色比例

本文利用CRF++对词组型术语进行机器自动标注语义角色，12 类语义角色标注的准确率均在 50%以上，其中，“FD”、“SY”、“SS”、“ZY”和“ZX”这五类语义角色的标注准确率在 80%以上。同时，术语内部的全部单词都可以利用这 12 类语义角色进行分类标注。

2.3　基于共享最近邻的词语相似度计算方法

基于统计的词语相似度计算方法，最为常用的计算方法是向量空间模型（VSM：Vector Space Model），把单词的特征处理为向量空间中的向量运算，以空间上的相似度表示词语的语义相似度[10-11]。但是该方法要依赖于所使用的语料库，这样难以避免由于数据稀疏问题所导致的相似度计算不准确的问题。

为了解决相似度计算不准确问题，我们在 VSM 的基础上提出了基于共享最近邻的词语相似度计算方法。共享最近邻（SNN：Shared Nearest Neighbor）是 Levent Ertoz[12-13] 提出的一种聚类算法，它通过使用数据点间共享最近邻的个数作为相似度来处理变密度簇的问题。

根据这种原理，我们提出了基于共享最近邻的单词相似度计算方法，即认为通过 VSM 方法计算得到的词 A 的相似度结果为词 A 的邻居，通过计算词 A 和词 B 的邻居的相似度，来对 A 和 B 的原始相似度进行惩罚，若两者含有较多的共同的邻居，则惩罚相对较小，反之，惩罚较大，然后选择一定的阈值

对相似度结果进行过滤，以此来提高相似度计算的准确性。主要计算步骤如下。

首先对专利语料进行分词、去除停用词等预处理操作之后，利用VSM模型计算单词相似度。

然后利用基于共享最近邻进行相似度计算。通过计算两个词语 A 和 B 的邻居之间的相似度，对 A 与 B 的原始相似度 $sim'(A,B)$ 进行惩罚，计算如式(1)所示。其中，$sim(A,B)$ 为进行惩罚之后的相似度，$sim(NA,NB)$ 为 A 和 B 邻居之间的相似度，计算如式(2)所示。在 $sim(NAi,NBj)$ 中，NAi 为 A 的第 i 个邻居，NBi 为 B 的第 j 个邻居。

$$sim(A,B)=\frac{sim'(A,B)+sim(NA,NB)}{2} \quad (1)$$

$$sim(NA,NB)=\sum_{i=1}^{m}\max\, sim(NAi,NBj) \quad (2)$$

$$(j=0,1,\cdots,n)$$

在对相似度进行惩罚之后，选择阈值 β 对词的邻居进行过滤，对相似度大于 β 的邻居予以保留，相似度小于 β 的邻居进行丢弃，以此更新邻居，在本文中 β 取值0.09，计算如式(3)所示。

$$Neib(A)=\{NA \mid sim(A,NA)>=\beta\} \quad (3)$$

3 基于内部语义角色的术语相似度计算方法

本文根据术语的特点提出了一种基于内部语义角色的术语相似度计算方法，该计算方法主要分为以下几步：(1)语义角色标注；(2)术语相似度计算。术语的相似度由组成术语的单词的相似度计算得到，实例如图2所示。

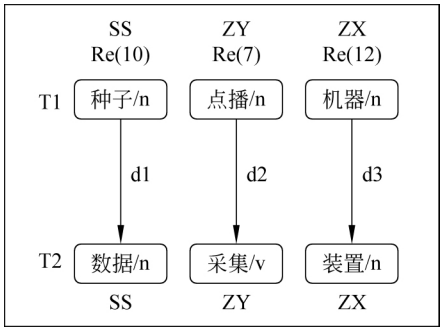

图2 术语相似度计算举例

其中，“SS”、“ZY”和“ZX”分别表示“受事”、“作用”和“中心词”语义角色，Re(10)表示“受事”类语义角色的权重，以此类推，$d1$ 代表单词“种子”和“数据”的相似度，以此类推。则术语“种子点播机器”与“数据采集装置”的相似度即为 $\mathrm{Re}(10)/\tau\times d1+\mathrm{Re}(7)/\tau\times d2+\mathrm{Re}(12)/\tau\times d3$，其中 $\tau=\mathrm{Re}(10)+\mathrm{Re}(7)+\mathrm{Re}(12)$ 为归一化因子。

3.1 术语语义角色标注

通过总结的12类语义角色，我们组织两名具有自然语言处理知识的人员对3 000条词组型术语内部进行语义角色标注(此标注语料可在 http://www.datatang.com/data/45908 处下载)，标注示例如图3所示。并对每一类语义角色赋予不同的权重，记为 $\mathrm{Re}=\{\mathrm{Re}(m) \mid m=1,2\cdots,12\}$，取值范围为0－1，其中 m 的取值为各类语义角色的顺序标号，顺序标号见表2。

```
SS 【 数据/n 】 ZY 【 提供/v 】 ZX 【 装置/n 】
SS 【 电压/n 】 ZY 【 调整/vn 】 ZX 【 单元/n 】
ZY 【 排/v 风/n 】 ZX 【 装置/n 】
SY 【 输送带/n 】 ZX 【 机构/n 】
ZY 【 分类/vn 】 ZX 【 装置/n 】
ZY 【 回/v 油/n 】 ZX 【 过滤器/n 】
SY 【 光盘/n 】 ZX 【 装置/n 】
JS 【 太阳能/n 】 ZX 【 充电器/n 】
SY 【 传感器/n 】 ZX 【 组件/n 】
JS 【 高压/n 】 ZX 【 电池/n 】
FS 【 可/v 动/v 】 ZX 【 电极/n 】
SY 【 快门/n 】 ZX 【 控制器/n 】
```

图3 术语语义标注示例

3.2 术语相似度计算方法

术语是由多个单词组成，要计算术语之间的相似度，首先需要知道各个单词之间的相似度。由于语义资源库涵盖词语有限，且专利中的单词多具有专业性，所以基于语义资源的相似度计算方法并不能得到很好的效果，因此我们利用2.3节中提出的基于共享最近邻的方法来计算单词相似度。

我们把要计算相似度的单词集合记为 Ω，把由共享最近邻方法计算得到的单词相似度集合记为 $Sl=\{Sl(ti,tj) \mid ti,tj\in\Omega\}$。

接下来，我们需要对两个含有语义角色标注的术语 $T1$ 和 $T2$ 进行相似度计算，本节中的单词相似度均从集合 Sl 中获得。

首先计算两个术语中具有相同语义角色的单词的相似度，得到的相似度要根据不同的语义角色乘以归一化后的角色权重。

然后对于 $T1$ 中剩余的没有相同语义角色的单词，则需要与 $T2$ 中的每个单词都计算相似度，取最大相似度结果。得到的最大的相似度结果在乘以相应语义角色的归一化的权重的同时还要乘以一个惩罚因子 δ。

最后两个术语的相似度即为每类相似度的求

和。在计算时，若 $T1$ 中的当前词 W 在 Ω 中不存在，或者是在 Ω 中，W 与 $T2$ 中的词不具有相似度，那么词 W 与 $T2$ 中的词的相似度赋值为 0.10。示例如图 4 所示。

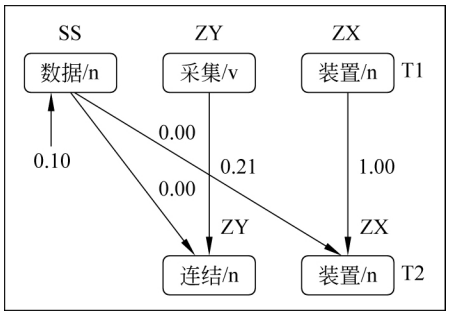

图 4　术语内部相似度计算实例

图 4 中，首先分别计算具有相同语义角色“ZX”和“ZY”的单词的相似度，即“装置”和“装置”的相似度，乘以“ZX”语义类的归一化权重；“采集”和“连结”的相似度，乘以“ZY”语义类的归一化权重。然后，由于“SS”类语义角色，在 $T2$ 中不存在，则分别计算“数据”与“连结”、“装置”的相似度，但是由于在单词相似度集合 Sl 中，“数据”与“连结”和“装置”不具有相似度，则他们的相似度为 0，那就把“数据”的相似度默认赋值为 0.10，在乘以“SS”语义类的归一化权重，还要乘以一惩罚因子 δ。则 $T1$ 与 $T2$ 的相似度为 $\mathrm{Re}(10)/\tau\times 0.10\times\delta+\mathrm{Re}(7)/\tau\times 0.21+\mathrm{Re}(12)/\tau\times 1.0$ 其中，$\tau=\mathrm{Re}(10)+\mathrm{Re}(7)+\mathrm{Re}(12)$。

4　实验

4.1　实验准备

本实验选择 1.8G 大小的中文专利语料，通过术语抽取系统抽取术语，并选择出现频数大于 100 的 18 744 条词组型术语作为实验对象，利用中国科学院分词工具“NLPIR 汉语分词系统”进行分词，然后进行语义角色标注。

4.2　评测方法

由于现在没有可利用的术语同义词词林，因此，我们组织 14 名具有自然语言处理知识的人员对 187 440 条词组型术语进行分类，将具有同义性的术语组织到一类中，构建适用于术语的同义词词林。评测方法为信息检索领域常用的一种系统性能测试指标 F 值。

我们规定，对于术语 Tj，通过相似度计算方法得到的邻居中的任一术语 t，若在术语同义词词林中 t 与 Tj 属于同一类，则认为术语 t 是 Tj 正确的邻居，即为 1，否则即为 0。则术语 Tj 的准确率 $P(Tj)$ 为，Tj 的邻居 $NebTj$ 中正确的邻居的个数除以 $NebTj$ 中邻居的总数 MTj，平均准确率 P 为所有术语的准确率总和的均值，见式(4)；术语 Tj 的召回率 $R(Tj)$ 为 Tj 的邻居 $NebTj$ 中正确的邻居的个数除以术语同义词词林中与 Tj 所属同一类的所有术语的总数 M，平均召回率 R 即为所有术语的召回率总和的均值，见式(5)，其中 N 为术语的总条数；平均 F 值，见式(6)。

$$P(Tj)=\frac{\sum_{i=1}^{MTj}Ki}{MTj}*100\%\quad Ki=\begin{cases}1 & right\\ 0 & wrong\end{cases}$$

$$P=\frac{\sum_{j=1}^{N}P(Tj)}{N}\tag{4}$$

$$R(Tj)=\frac{\sum_{i=1}^{MTj}Ki}{M}*100\%\quad Ki=\begin{cases}1 & right\\ 0 & wrong\end{cases}$$

$$R=\frac{\sum_{j=1}^{N}R(Tj)}{N}\tag{5}$$

$$F=\frac{2\times P\times R}{P+R}\tag{6}$$

4.3　实验介绍

(1) 术语语义角色标注实验

利用 3.1 节中人工 3 000 条术语作为实验对象，将该语料平均分为十组，每组 300 条。利用十折交叉验证，即轮流用九组作为训练语料，剩余一组作为测试语料，利用CRF++0.58 对测试语料进行语义标注。通过计算各个语义角色标注的平均准确率，以证明本文提出的语义角色体系的可用性。

(2) 通过不同的单词相似度计算方法计算术语相似度实验

分别利用基于《知网》[7]，VSM 和通过共享最近邻方法计算单词相似度，然后利用三种方法得到的单词相似度来计算术语相似度。

(3) 术语相似度计算方法比较实验

分别利用以下三种方法对 18 744 条词组型术语进行相似度计算：①“VSM”方法，即把术语作为一个基本单位，利用其在专利文本中的上下文作为特征来计算术语相似度；②共享最近邻方法，即把术

语作为一个基本单位，在通过“VSM”得到的结果基础之上加入邻居的信息来计算术语相似度；③基于内部语义角色的术语相似度计算方法。对比三种方法得到术语相似度的效果。

根据各个语义角色所占的比例，我们可以知道，“ZY”、“SY”“SS”、“FS”和“ZY”以上五类语义角色对相似度计算所起的作用较大，其余七类所起的作用较小，因此在选择语义权重时，我们仅对以上五类语义角色的权重取平均，其余七类的权重保持初始值不变，初始权重为各个语义角色所占比例。

4.4 实验结果

（1）术语语义角色标注实验

各个语义角色的标注准确率如表3所示。

表3 语义角色机器标注结果

语义角色	标示符	平均标注准确率/%
否定	FD	1
外观	WG	0.63
使用者	SY	0.90
方式	FS	0.70
性能	XN	0.6
材料	CL	0.53
受事	SS	0.84
关系	GX	0.69
作用	ZY	0.94
类别	LB	0.63
技术	JS	0.61
中心词	ZX	0.98

由表3可以看出，在各类语义角色中，“FD”、“SY”、“SS”、“ZY”和“ZX”这五类语义角色的标注准确率在80%以上，其中“CL”这类标注准确率在60%以下，通过观察发现导致这类准确率低的原因主要有以下两点。

首先，“CL”与“SY”这两类语义角色的结构非常相似，“SY”语义类的组成结构多为“名词＋名词”，例如“催化剂/n 载体/n”，其中第一个名词多为“SY”类，第二个名词为“ZX”类。但是对于例如像“铝/n 基线路板/n”这一类术语的组成结构同样为“名词＋名词”，但是“铝/n”却不属于“SS”类，这种结构类似的术语会导致标注错误。其次，由于训练语料不充足，会导致数据稀疏问题，这也会导致标注错误。

（2）通过不同的单词相似度计算方法计算术语相似度实验

实验结果如图5所示。

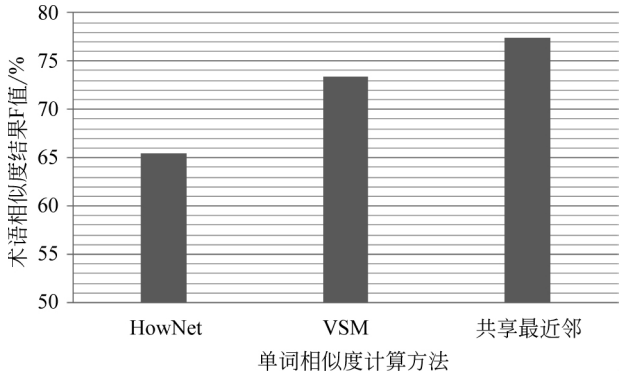

图5 三种单词相似度计算方法效果比较

由图5可以看出，“VSM”方法得到的F值要比“HowNet”得到的F值高，主要原因是，“VSM”方法是利用单词在大量专利文本中的上下文作为特征来计算相似度的，其所得到的结果更符合专利文本中的术语，而《知网》中的单词是面向通用领域的，并不能很好的适用于专利文本。因为“共享最近邻”方法是在“VSM”的基础上加入了邻居的信息来进行改进，所以得到的效果比单纯使用“VSM”方法有所提高。

（3）术语相似度计算方法比较实验

我们利用三种方法计算术语相似度，然后比较结果。实验结果如图6所示。

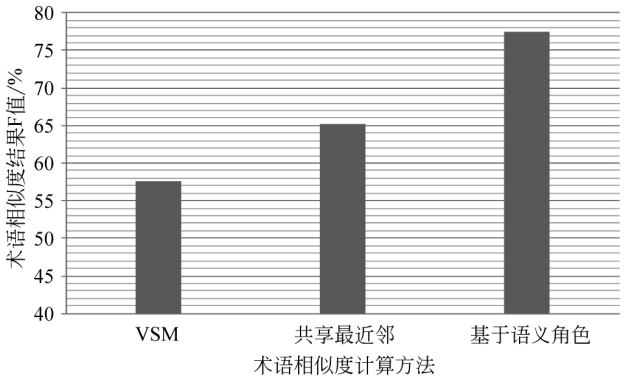

图6 三种术语相似度计算方法效果比较

由图6可以看出，通过“VSM”方法计算得到的F较低，主要原因是，专利文本格式较为统一，术语的上下文信息并不能完整的表达其含义，而且对于一些新产生的术语，其上下文信息并不丰富，所以利用“VSM”方法得到的效果并不理想。共享最近邻方法是在“VSM”方法的基础上进行改进的，能够在

一定程度上缩减由于数据稀疏等问题造成的相似性计算不准确的问题。基于内部语义角色的相似度计算方法，没有利用术语不丰富的上下文，而是加入了语义角色的信息，根据语义角色，通过单词的相似度来计算术语的相似度，效果与其它两种方法相比有很大提升。

5　结论与展望

专利文本较普通文本而言，用词较为单一，所使用的词语多具有专业性，基于统计的和基于语义词典的相似度计算方法不能很好的对术语进行相似度计算，考虑到这一问题，本文结合术语本身的语义结构特征，提出了一种适合术语的相似度计算方法——基于内部语义角色的术语相似度计算方法。我们所提出的方法能够很好的解决传统的基于统计的方法由于数据稀疏所导致的相似度计算不准确和传统的基于语义资源的由于语义资源覆盖不全面而无法计算相似度的问题。

本文所提出的方法，并没有考虑到术语的上下文，仅是利用其内部语义角色来计算相似度，虽然对于术语，其上下文信息并不全面，但是若给予考虑，也可对相似度计算提供一定的价值，所以在将来的工作中，我们考虑可以将“VSM”相似度计算方法和基于内部语义角色的相似度计算方法进行融合，既考虑上下文信息，又考虑术语内部角色，力求能够更好的计算词组型术语的相似度。

参考文献

[1] 冯志伟. 术语形成的经济律——FEL 公式[J]. 中国科技术语，2010，12(2)：9-15.

[2] Dagan I，Marcus S，Markovitch S. Contextual word similarity and estimation from sparse data[C]//Proceedings of the 31st annual meeting on Association for Computational Linguistics. Association for Computational Linguistics，1993：164-171.

[3] Terra E，Clarke C L A. Frequency estimates for statistical word similarity measures[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics，2003：165-172.

[4] 蔡东风，白宇，于水，等. 一种基于语境的词语相似度计算方法[J]. 中文信息学报，2010，24(3)：24-28.

[5] 王斌. 汉英双语语料库自动对齐研究[D]. 中国科学院博士学位论文，1999.

[6] 刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学，2002，7(2)：59-76.

[7] 于林林. 基于知网的汉语词义消歧方法的研究[D]. 沈阳航空工业学院硕士学位论文，2008.

[8] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报：信息科学版，2010 (006)：602-608.

[9] Pilehvar M T，Jurgens D，Navigli R. Align，disambiguate and walk：A unified approach for measuring semantic similarity[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). 2013.

[10] 陈小芳. 汉语术语语义分析技术研究及其应用[D]. 沈阳航空航天大学硕士学位论文，2011.

姜利雪(1988—)，硕士，主要研究领域为自然语言处理。

E-mail：jlxsnow@163.com

蔡东风(1958—)，博士，教授，主要研究领域为人工智能，自然语言处理。

E-mail：caidf@vip.163.com

季铎(1981—)，博士，副教授，主要研究领域为自然语言处理。

E-mail：jiduo_1@163.com