基金项目：国家“863”计划基金资助项目“面向农业领域的海量知识资源组织、管理与服务系统研究”(2007AA01Z179)；
作者简介：陈宇（1982-），男，硕士研究生，主要研究方向：软件工程，领域本体；朱建锋，硕士研究生；吴毅坚，讲师，博士；赵文耘，教授，博士生导师；
Email:chen_yu@fudan.edu.cn

一种基于领域本体的新术语扩充方法

陈宇，朱建锋，吴毅坚，赵文耘
（复旦大学计算机科学技术学院,上海 201203）

摘要：本文提出了一种基于领域本体的新术语扩充方法。结合传统基于统计和基于规则的方法，计算出词语在文档中的影响，使用领域本体来体现领域知识，通过在文档中识别出的本体中概念来计算文档及词语的领域相关度，最终获得术语候选项的推荐排序，可以对术语候选项结果进行优化。设计实验及提供结果分析，说明该方法的有效性和可行性。
关键词：领域本体；领域相关度；新术语识别；

A method for new term recognition based on domain ontology

CHEN Yu，ZHU Jian-feng，WU Yi-jian，ZHAO Wen-yun
（School of Computer Science Fudan University，Shanghai 201203）

Abstract：A method for new term recognition base on domain ontology is proposed in the paper. This method combines linguistic rules and statistical methods to get the infection degree from a word to a document. Optimize the term candidate result, using domain ontology to recognize concept in document and to calculate correlation degree between word, document and specific domain. The validity of the method is proved with a practical case.
Keywords：Domain ontology; Domain correlation degree; New term recognition

1. 引言

术语(Term)，是在特定领域中一般概念词语指称[1]，如农业领域中的“氮肥”、“白斑病”等。因为术语本身具有较强的规范性，对术语的认定和收录要经过一个过程，所以术语词典的编纂往往滞后于术语的出现，需每隔一段时间进行扩充。传统术语词典的编撰和扩充需要领域专家手工进行，阅读大量文献，费时费力。如能使用计算机辅助识别出候选项，再通过专家参与确定，会有比较好的效果。计算机辅助术语扩充就是本文所要关注的问题。

术语扩充，也可以说是新术语识别，所谓“新”，是相对已有的领域术语词典而言。领域术语词典，即包含特定领域内部代表本领域特定概念词语的词典。而日常口语和书面语常用词典，往往被称为一般词典或通用词典。

要想进行术语扩充，首先要找到“新词”，主要有以下类别：

a 不在通用词典和领域术语词典里面的词语（也被称为“未登录词”）。

b 在通用词典中，领域内使用频繁，却不在领域词典中的词语。由于初始的领域词典词汇较少，一些通用字典的词，可能还没有及时收录。

c 在通用词典中被识别成多个词，组合在一起才会在特定领域表达完整意思的词语，如“旱黄瓜”（黄瓜的一种），如果在词典中没有，可以能被识别成“旱”和“黄瓜”，应该被合成。

本文针对这几类术语扩充问题进行研究(其中 c 类称为合成词,a 类和 b 类称为基础词)，提出一种扩充方法，结合传统基于统计和基于规则的方法，利用本体技术，使用领域本体来体现领域知识，提供比较丰富的语义信息，进而计算术语的领域相关度，对术语扩充结果进行优化。后续章节将介绍相关工作，该方法的具体步骤，算法、实验设计和结果分析。

2. 相关工作:术语扩充方法和本体介绍

目前对于此问题国内外研究主要有两个方向的思路[2]，一种是基于语言学规则的方法，术语往往具有特殊词缀，词性。利用自然语言处理方法，使用术语构成的语法和词形模式判断词串是否符合术语条件。另外一种是基于统计特征的方法，如逆文档词频 TFIDF[9]、互信息[8]等。统计方法判断术语主要依据特征值构建统计模型，查看词串指定特征值是否符合该模型阈值。未来趋势是将两种方法结合使用。另外要考虑到中文分词困难的特点，目前分词效果较好的是中科院的基于层叠隐马模型的汉语词法分析系统 ICTCLAS[7][10]。以上的研究主要集中在术语单元度（即语言上内部紧密程度）方面，一个完整的语言单位不一定是一个领域术语，所以在满足单元度的基础上，还要从领域相关度的角度进行考察[3]。

本文借助本体作为领域知识的代表,本体的概念源于哲学领域,引入到知识工程领域后，被越来越多关注，被广泛引用的定义是由 Gruber 提出的“本体是概念化的明确的规范说明”，后 Studer 等人进行补充为“本体是共享概念模型的明确的形式化规范说明”。随研究的深入诞生许多本体描述语言，如 RDF、RDF-S、OWL 等[6]。本体的编辑和开发也有一些工具支持，如美国斯坦福大学开发的本体编辑工具 Protégé，惠普公司实验室开源项目 Jena 等。

涉及特定学科领域知识的本体，被称为领域本体，本文将建立一个农业领域病虫害相关的本体作为例子，并以此本体作为领域知识代表，辅助术语扩充。

3. 文档的结构化定义及本体表示法

术语扩充的载体是领域相关的文档，文档本身可以提供丰富的结构信息[4][5]，本文对于文档的定义如下

定义 1：文档空间 DS：包含各类文本文档，记为 DS=$\{d_1, d_2 \cdots d_j \cdots\}$，用 d_j 表示文档空间 DS 中任意一个文档。

定义 2：文档空间的词集合 W(DS)：文档空间 DS 中所有文档中的词语，记为 W(DS)=$\{w_1, w_2 ... w_i ...\}$，用 w_i 表示文档空间的词集合 W(DS)中任意一个词语。

定义 3：文档词集合 W(d_j)：文档 d_j 中词语集合，是 W(DS)的子集；

与词语有关的表示法：

w_{ij}：表示出现在文档 d_j 中的词语 w_i；

n(S)：表示集合 S 的数量，如 n(DS)表示文档空间 DS 中的文档数量；

f(w_{ij})：表示词语 w_i 在文档 d_j 出现的次数；

g(w_{ij})：如果词语 w_i 在文档 d_j 出现，g(w_{ij})=1，否则，g(w_{ij})=0；

I_{ij}：表示词语 w_i 在文档 d_j 中的影响，也可以说词语 w_i 在文档 d_j 的重要程度。

与领域有关的表示法：

定义 3：三元组集合 TS：本体知识表示，本文使用 RDF（S）的描述形式，把本体同时看成三元组的集合，每一个三元组形如（主体，谓词，客体）。

定义 4：术语推荐项集合 TRS：最终词语作为术语推荐项的结果集合，形如 TR=$\{(w_1, \text{Recom}(w_1)),(w_2, \text{Recom}(w_2)) \cdots (w_i, \text{Recom}(w_i)) \cdots\}$，其中 Recom($w_i$)表示对于 w_i 作为

术语候选的推荐度。

4. 术语扩充指导规则

术语和一般词语相比在出现时间、文中影响度、格式等方面具有的自身特点，定义扩充指导规则如下：

规则 1 时间特性：新术语的出现，往往在一段时间内较频繁，有时效性和新颖性。即可以根据术语所出现的文档的发布时间进行过滤和选择。

规则 2 影响特性：术语在一篇领域相关的文档中往往代表更重要的意思，在文档中重要性较大。

规则 3 位置特性：术语往往在文档的重要位置出现，如文档题目，摘要或者标注为关键字的段落，及首段和末段。

规则 4：格式特性：术语为了便于理解，通常会突出显示或解释，使用黑体，斜体，括号，注释符号等说明。这是术语带有的格式特点。

规则 5：高频特性：术语词在领域相关文献内部，词频较高。由于要在领域内部作为知识交换的重要载体，在领域内使用较频繁。

规则 6：内聚特性：术语词内部各部分结合交紧密，如“阿司匹林”内部结合紧密，而每个字片段所代表的意义不大。

规则 7：领域特性：术语往往出现特定领域的上下文环境，同领域的术语往往共同出现。所在的文档具有很强的领域相关性。

5. 领域本体的建立

不失一般性，本文建立了一个简单的农业方面与农作物生长及病虫害有关的本体作为示例，如**错误！未找到引用源。**所示，实际中可以根据需要确定本体规模和领域方向，开始建立时可以采用最为常见的领域概念知识先建立一个小规模的原型，再根据本文方法部分运行结果再对本体进行细化加工。

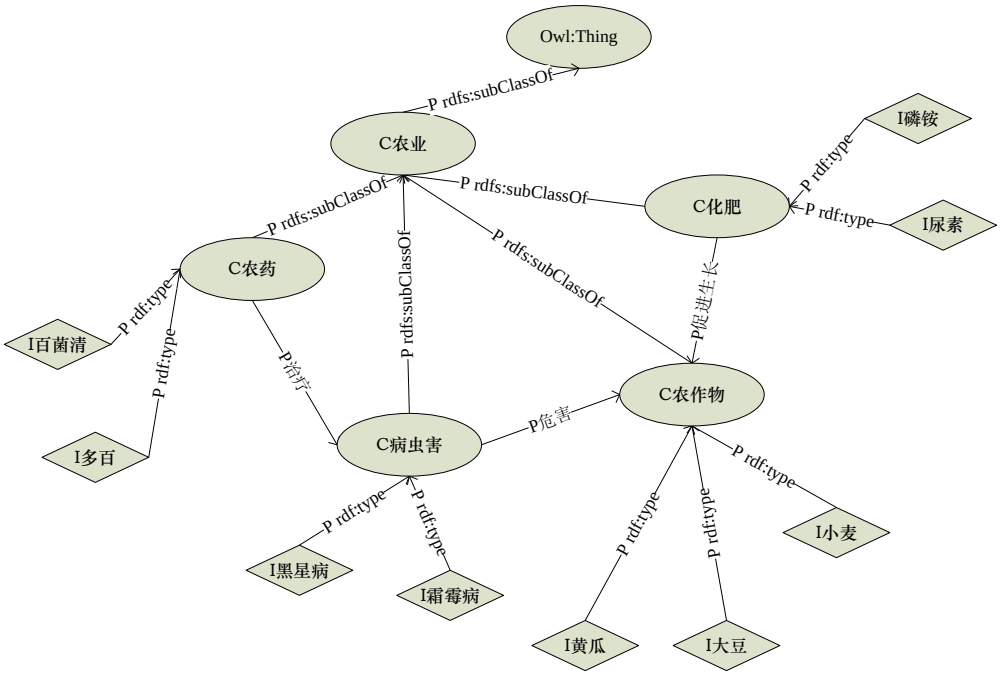

图 1 农业本体例子：与农作物生长及病虫害有关的本体层次片段

本例中主要有C={农业、农作物、化肥、农药、病虫害} 五个大类；R={危害、治疗、促进生长} 三种主要关系。每个类下面建了若干实例，如农作物有黄瓜、大豆、小麦等实例，病虫害有黑星病、霜霉病，农药有百菌清、多百，化肥有尿素、磷铵。图中未能完全展现出

来的还有形如“主体-谓词-客体”的三元组，也是用来描述知识的，如（霜霉病，危害，黄瓜）、（尿素，促进生长，黄瓜）、（百菌清，治疗，霜霉病）提供了概念，及概念之间的关系。本文利用领域本体提供领域的知识来考察文档的领域相关度。

6. 基于领域本体的术语扩充方法

6.1. 扩充方法过程

如图 2 所示

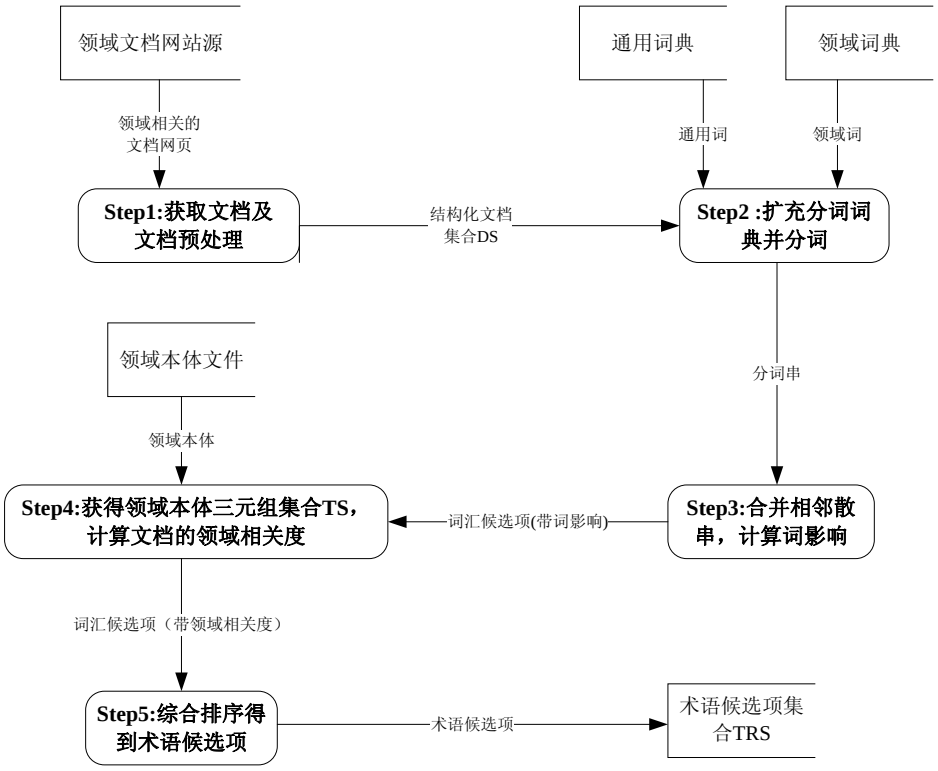

图 2 术语扩充流程图

6.2. 获取文档及文档预处理

新术语要从领域相关文献中挖掘出来，网络上文献更新速度快，能代表时代特点。但从网上获取的文档，带有网页格式信息，有一些格式信息和广告信息，和文档内容无关，需要过滤掉。一些是可以利用信息，如网页关键字、文本中的词语链接，可以根据需要保留。

6.3. 扩充分词词典并分词

一般的分词方法所使用的词典为通用词典，所以要配合领域词典一起分词，才能够一次就识别出已有的领域术语，领域词典通常从权威机构获得或由领域专家提供。在扩展 ICTCLAS 分词词典时，调用程序中插词函数，同时为了优先识别已有领域术语，减少歧义错分，需要尽可能调高这些词语的优先级。

6.4. 合并相邻散串，计算词影响

对于分词初步结果合并相邻散串，得到部分候选合成词。首先利用基于中文语法规则的方法，通过词性信息，建立一系列过滤模板，如术语一般是名词或名词短语，然后计算散串的互信息 M，即考虑规则 6 的影响。

互信息是用来度量一个消息中两个信号之间的相互依赖程度。二元互信息是两个事件的概率函数：

$$M(X,Y)=\log_2\frac{P(X,Y)}{P(X)*P(Y)} \qquad (1)$$

互信息值越高，X 和 Y 组成短语的可能性越大；反之，互信息值越低，X 和 Y 之间存在短语边界的可能性越大。

描述性的文档，词的重要性不一样，通过计算，尽量找到对于一篇文档比较重要的词。综合考虑规则 3-5，主要计算候选项的 TF-IDF 的值，再考虑词的位置信息。

$$I_{ij}=\text{Position}(w_{ij})*\text{f}(w_{ij})*\log_2\frac{n(DS)}{\sum_j g\left(w_{ij}\right)} \qquad (2)$$

其中 Position(w_{ij})为 w_{ij} 在文档中位置所提供的重要程度系数，按题目，首段，末段重要程度排序。f(w_{ij})为 w_{ij} 的词频，$\log_2\frac{n(DS)}{\sum_j g\left(w_{ij}\right)}$ 为 w_{ij} 的逆文档词频。

6.5. 获得领域本体三元组集合 TS，计算文档的领域相关度

文献[5],给出了一种使用本体中的实例标注文档文件的方法,用到了领域相关度的计算，通过计算本体中每个实例所对应三元组,与文档之间的包含关系,获得实例与文档的相关度。针对术语扩充这个目的，本文对其进行改进如下:

a 使用分词算法对每个子句进行分词,借助词典及分词算法减小相邻字歧义,如子句“白色斑点”中，如果不进行分词有可能会匹配到“色斑”，是一个错配，会降低查准率。而对于本体扩充这个目的，可以通过更多的文档提供足够多的术语候选项，所以在查准率和查全率的平衡中，应该更重视查准率，尽量匹配正确。

b 由于文档中描述是使用自然语言，子句往往不足以表达完整的意思，经常出现主语省略或代指情况，所以考察力度提升至句子和段落。

c 对于三元组和句子的匹配度计算：主语对于概念的代表性最强；宾语次之；谓语较为灵活，变体较多，很难完全匹配，而对于匹配上的谓语，因为谓语往往很多领域公用的也往往不一定是与特定领域相关，如“有害”这一属性，可以在文档中出现，却不一定是和农业本体中的“有害”属性一致。

用 RelDoc2Damain(d_j)代表文档 d_j 与特定领域的相关程度，即统计与文档匹配的三元组数和匹配度:

RelDocToDomain(d_j)=($\sum$ RelSentToTriple)/(文档词数*n（TS）) （3）

算法如下:

```
输入：领域文档集合DS，领域本体DomainOnto
输出：每篇文档与领域本体的相关度RelDocToDomain排序
//以子句为粒度的领域相关度计算算法
 文档数目N=n(DS)
 由DomainOnto得到三元组集合TS
 WHILE(N>0)
   BEGIN
   FOR 文档d  IN  DS
       BEGIN
       初始化d ，取出标题TitlePara，分出正文段落，得到段落集合PS
       对TitlePara进行分词，将TitlePara表示为一个词串
       FOR （ 段落P in PS ）
           BEGIN
               将段落按照标点符号分成若干句子S，并对S进行分词，将S表示成词串的形式，得到句子集合SS
           END
       初始化该文章的领域相关度RelDocToDomain=0
       FOR(段落P in PS )
           BEGIN
               FOR( S in SS )
                BEGIN
                   FOR(triple in TS )
                       BEGIN
                           //在S的词串中匹配triple的主体、谓词和客体
                           三者同时有匹配项或者仅主体和客体有匹配项时
                           //计算S与triple的相关度
                           Rel(S-triple)= 主体相关度*主体系数+谓词相关度*谓词系数+客体相关度*客体系数
                           RelDocToDomain = RelDocToDomain + 正文系数* Rel(S-triple)
                       END
                END
           END
           IF(RelDocToDomain>0) THEN
               FOR(三元组triple in TS )
                   BEGIN
                       //在TitlePara的词串中匹配triple的主体、谓词和客体
                       当三者同时有匹配项或者仅主体和客体有匹配项时
                       //计算TitlePara与triple的相关度
                       Rel(TitlePara-triple)=主体相关度*主体系数+谓词相关度*谓词系数+客体相关度*客体系数
                       RelDocToDomain = RelDocToDomain + 标题系数* Rel(TitlePara-triple)
                   END
               RelDocToDomain = RelDocToDomain/(文档词总数*n(TS))
               将RelDocToDomain与d的关系记录到数据库
               N=N-1
       END
   End

将领域文档按RelDocToDomain从大到小排序
```

6.6. 综合排序得到术语候选项的推荐度

考虑规则 6，综合领域相关度和词候选项的贡献，计算出每个术语候选项的推荐程度

$$\mathrm{Recom}(w_i)=\sum_j\left(I_{ij}*\mathrm{Re}lDocToDomain\left(d_j\right)\right) \qquad (4)$$

$\mathrm{Recom}(w_i)$为综合了词影响 I_{ij} 和文档的领域相关度 $\mathrm{RelDocToDomain}(d_j)$之后，再在整个文档空间 DS 的范围内统计获得，最后将术语候选项排序后，送领域专家做最终判断，确定是否收录，优化后的排序更加合理。

7. 实验设计及结果分析

7.1. 实验过程

本文实验使用的领域文档来自于较权威的全国优秀农业政府网站：安徽农网。首先使用开源工具 Wget 下载安徽农网农业科技频道的网页，再通过正文提取得到农业领域文档共 3468 篇。实验提取 AGROVOC(联合国粮食及农业组织和欧共体开发的多语种结构的叙词表)中的 37060 个农业中文词汇作为领域词典扩充了 ICTCLAS4J[10]的分词词典并使用 ICTCLAS4J 对文档进行分词处理。实验中使用的测试本体由本体编辑工具 Protégé 辅助生成，并使用 Jena 对其进行解析操作，本体的 owl 文件片段如图 3 所示。

```
<Agricultural_Chemical rdf:ID="Carbofuran">
   <cures>
     <Pest_And_Disease rdf:ID="Root_nematode">
       <Harm>
         <Agricultural_Crop rdf:ID="Soybean">
           <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">大豆</rdfs:label>
         </Agricultural_Crop>
       </Harm>
       <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">根线虫病</rdfs:label>
     </Pest_And_Disease>
   </cures>
   <cures>
     <Pest_And_Disease rdf:ID="Bud_blight">
       <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string" >芽枯病</rdfs:label>
       <Harm rdf:resource="#Soybean"/>
     </Pest_And_Disease>
   </cures>
   <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string" >呋喃丹</rdfs:label>
 </Agricultural_Chemical>
```

图 1 本体的 owl 文件片段

设定出词率和正确率作为实验指标:

出词率指程序得出的新词个数与所有领域文档的长度之和的比值，用 TermProductivity 表示:

$$\text{TermProductivity} = \frac{n(\text{术语候选项集合})}{n(DS)} \qquad (5)$$

正确率指取前 200 个词的正确率指得到新词按领域相关度从大到小排序后取得的前 200 个词中含有的正确且和领域相关的词的比率，用 TermPrecision 表示:

$$\text{TermPrecision} = \frac{n(\text{正确术语候选项})}{n(\text{术语候选项集合})} \qquad (6)$$

7.2. 实验结果 1： 匹配粒度对于结果的影响

在实验中，使用三种不同的粒度进行三元组匹配：句子粒度、段落粒度和文章粒度。然后在每一粒度下分别产生新词，并计算出词率和正确率。

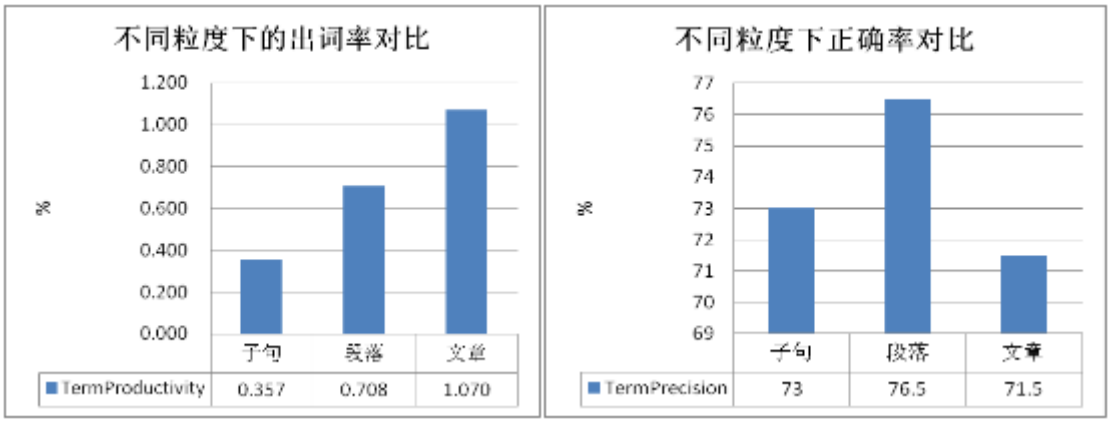

图 2 不同粒度下出词率和正确率的对比图

图 4 反应了在不同粒度下本方法的出词率和正确率的差异。对于出词率而言，以文章为单位进行三元组匹配能达到最大值。而对于正确率而言，各种粒度下都在 70%到 80%之

间，其中段落粒度稍高。综合正确率和出词率，段落粒度在本方法中效果最好。

7.3. 实验结果 2：对于通用词典中尚未收入领域词典的基础词语识别

对实验中以段落为粒度得到的基础词进行统计，得到如下结果：

表格 1 对于通用词典中尚未收入领域词典的基础词语识别

出词率	正确率
0.47%	78.5%

正确结果，即与领域相关的词大致可以分为以下几类：

领域紧密联系词，例如：“灰霉病”、“镇压”、“绿亨”等。其中，“镇压”是农业中对旱地小麦的一种种植处理方式，“绿亨”是一种农用杀菌剂。

领域相关的程度描述词，例如：“偏重”、“严重”、“适宜”等。它们多用来描述农业中与程度相关的行为或者事物，可收录进术语词典表程度。

领域相关的处理类描述词，例如：“选用”、“对照”等。

量纲，例如：“分钟”、“公斤”、“厘米”等。

错误结果，即被认为与领域无关的词可以分为以下几类：

纯数字。这类纯数字不能表示实在的意义，不能算作领域相关词。

时间，例如：“9 月”、“2005 年”等。

普通词，例如：“可以”、“流行”等。由于这类词非常通用，所以不能作为领域相关词。

在使用模板过滤掉纯数字与领域无关时间候选项之后：

表格 2 对于通用词典中尚未收入领域词典的基础词语识别（带数字过滤模板）

出词率	正确率
0.42%	87.4%

由于过滤掉无意义候选项，出词率会有所下降，但正确率上升效果明显。

7.4. 实验结果 3：互信息的对于合成词结果的影响

在进行词语合成时，互信息越大说明两个词能够合成为一个新词的概率就越大。

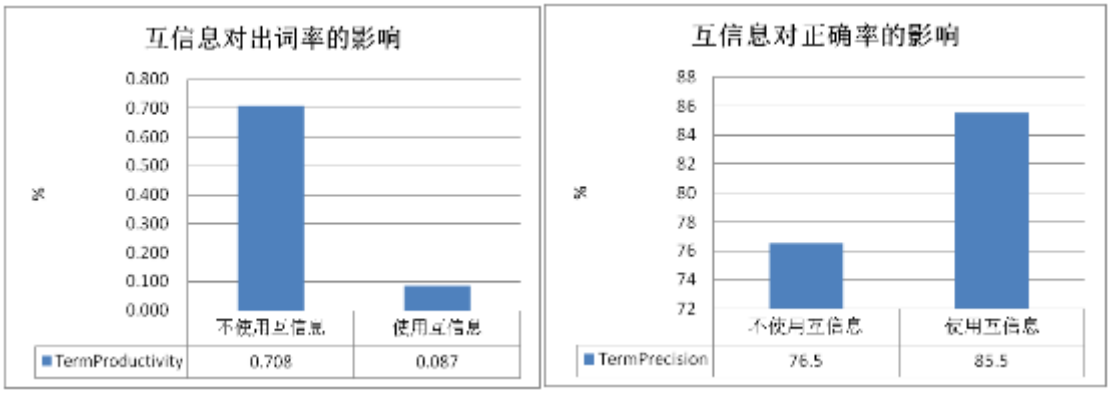

图 3 互信息对出词率和正确率的影响

图 5 “使用互信息”表示如果组成新词的两个源词之间的互信息小于给定的阈值就抛弃这个词。实验中根据不同的领域文档特点，找到合适的互信息阈值，能够提高正确率，但会降低出词率。如图 5 中即为对比结果，出词率降低了一个数量级，正确率提高了 9% 。

可以将正确结果大致做以下分类：

领域紧密联系词，如：“小麦白粉病”、“僵茵”、“全能铁甲”、“世玛”等。其中“僵茵”和“世玛”是除草剂，“全能铁甲”是一种砧木品种；

农业相关机构名称，如：“气象部门”、“黄瓜研究所”、“面粉厂”等；

农业地理相关的词，如：“长江中下游”、“长江流域”等

错误结果可以分为以下几类:

分词歧义，如:“侧蔓均”一词，分词时程序将其分为“侧+蔓均”，而实际上原文是“…黄瓜主侧蔓均有…”。

合词不足，如:“粉锈”一词，原文是“使用…粉锈宁可湿性粉剂拌种…”,分词时中间部分被分为“粉+锈+宁+…”，而本方法只能进行两个词合成，所以造成了这种情况出现。我们下一步的工作将考虑三词或更多词的合成，以便解决此问题。

7.5. 实验结果 4：与 TF-IDF 方法的对比

在领域术语扩充方法中，TF-IDF 是比较经典的一个方法。如果使用 TF-IDF 的方法进行领域术语扩充，领域文档中的任何一个词都会得到一个相关度值，所以出词率无法计算，重点对比正确率。

表格 3 不同方法正确率比较

	段落粒度使用互信息的语义扩充方法		TF-IDF 的扩充方法	
	农业领域	农作物病虫害领域	农业领域	农作物病虫害领域
合成词	85.5%	59%	70%	16.9%
基础词	87.4%	30.8%	65%	23%

实验中使用的文档是农业领域相关的文档,其中和农作物病虫害相关的文档是它的一个子集。为了更加详细的了解本文方法与 TF-IDF 之间的差别，对比时采用两种不同的标准来判断新词的正确与否: 该词是否属于农业领域和该词是否属于农作物病虫害领域。因为实验所使用的本体是农作物病虫害相关的本体,所以使用这样的两种判断标准能更好的反应这两种方法之间的差别。在使用农作物病虫害领域作为判断标准时，TF-IDF 的方法得到的合成词正确率非常不理想,而本文方法得到的结果较好,主要因为本文方法具有领域判断的能力。

8. 小结及未来工作

本文提出了一种基于领域本体的术语词表扩充方法，结合传统的基于统计和规则的方法，并通过识别文档中的领域概念来计算文档的领域相关度，对于术语的推荐排序效果起到较好的作用，此外对于领域文档的获取，领域本体的规模和方向性对于结果也有一定影响，应该尽量找到比较权威的领域相关文档源或语料库，同时根据需要细化领域本体的构建。下一步工作是考虑多领域本体对于新术语扩充的辅助作用,多重合成词产生以及探索在人工确定的基础上考虑迭代扩充的方法，以更提高扩充效率和准确率。

参考文献

[1] GB/T 15387.2-2001—2001 术语数据库开发指南[S]

[2]刘建华,张智雄,徐健等.自动术语识别——对科技文献进行文本挖掘的重要技术方法[J].现代图书情报技术,2008,(8):12

[3]张秦龙,穗志方,丁万松.术语自动提取中的领域度计算方法研究[A].第三届学生计算语言学研讨会论文集[C].2006

[4]Kyo Kageura, Bin Umino. Methods of automatic term recognition: a review [J].Terminology 1996, 3(2):259-289

[5]陈叶旺,李文,彭鑫,赵文耘.基于本体的文档语义标注改进方法[J].东南大学学报(自然科学版),2009,39(6):1109

[6]钱平,郑业鲁农业本体论研究与应用.北京:中国农业科技出版社.2006.1:73

[7]刘群,张华平.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8):1421

[8]刘荣,王丽娟,张志平等.利用高频词和互信息面向特定领域提取多字词表达[J].太原理工大学学报,2009,40(3):210

[9] Sparck Jones, K., Index term weighting, Information Storage and Retrieval, 1973, Vol9, No.11:619

[10] ICTCLAS4J.http://www.ictclas.org/Down_OpenSrc.asp