文章编号：1003-0077(2009)06-0072-07

一种基于使用差异的词语领域性分析方法

李素建[1]，宋 涛[1]，高 杰[2]，幺鹏跃[1]，李文捷[3]

（1. 北京大学 计算语言学教育部重点实验室，北京 100871；
2. 山东菏泽学院，山东 菏泽 274000；3. 香港理工大学计算机系，香港）

摘 要：领域知识的表达形式最终体现在词汇的领域性上，因此对领域词及其部件的领域度分析是一个关键。该文在分词的基础上，对各个领域语料进行分析，利用词语之间的关系，引入链接分析方法分析词语在各个领域中的使用重要性，并通过词语在各个领域中的使用差异性计算其领域度，从而达到领域分析的目的，获取某个领域的领域部件词。该文采用以上方法在军事、娱乐等领域进行了实验，实验结果表明该方法相对于当前常用的 tf×idf 方法和 Bootstrapping 方法，可以更有效地进行领域分析获取领域部件词。

关键词：人工智能；自然语言处理；领域性分析；领域词；领域部件词；链接分析；使用差异

中图分类号：TP391 **文献标识码**：A

A Method of Lexical Domain Analysis Based on Usage Discrepancy

LI Sujian[1], SONG Tao[1], GAO Jie[2], YAO Pengyue[1], LI Wenjie[3]

(1. Institute of Computational Linguisitics, Peking Universtiy, Beijing 100871, China;
2. Foreign language Department, Heze University, Heze, Shandong 274105, China;
3. Department of Computing, The Hongkong Polytechnic University, Hongkong, China)

Abstract: The representation of domain knowledge usually focuses on the domain lexicons, and then domain analysis for terms or term components is a natural task. In this paper, we propose a novel domain analysis method based on the discrepancy of lexical usage. Based on the word segmentation result, we introduce a link analysis method to compute the usage degree of each word for several typical domain corpora. Then through analyzing the discrepancy of the word usage in different domains, we can acquire the domain term component with larger usage discrepancy. This method is experimented on several domains such as military, entertainment and so on, achieving better results than the commonly used tf×idf method and Bootstapping method.

Key words: artificial intelligence; natural language processing; domain analysis; domain term; domain term component; link analysis; usage discrepancy

1 引言

随着社会的发展，各个领域的文本信息近年来成指数级增长，一方面给用户提供了大规模的信息来源，同时也给用户搜索有用信息造成了困难。用户获取信息时通常希望在某个领域得到深入可靠的知识，而不只是局限于那些表层的通用的信息，由此领域性知识的研究是当前的一个热点。而不同领域的知识其表达形式最终还是体现在词汇的领域性上，因此不少研究者进行了术语识别、术语部件、领域词典构建等方面的研究[1-8]。由此可见，领域词汇的分析和获取是领域知识深入分析和研究的一个重要基础和关键问题。而且，领域词分析的结果对信息检索、信息抽取、文本分类聚类等应用也将提供有益的帮助。

当前领域词获取方面的研究主要是基于大规模

收稿日期：2009-01-19 **定稿日期**：2009-08-10

基金项目：国家自然科学基金资助项目(60603093,60875042)；国家 973 课题资助项目(2004CB318102)

作者简介：李素建(1975—)，女，博士，副教授，主要研究方向为计算语言学、自然语言处理、信息检索等；宋涛(1986—)，女，学士，主要研究方向为计算语言学、信息检索；高杰(1978—)，女，硕士，讲师，主要研究方向为英语语言学。

语料和机器学习的方法。文献[9]在大规模语料的基础上提出一些判断领域词边界的度量指标，如unithood 和 termhood 等，在一定程度上提高了领域词获取的性能。由于直接获取领域词的性能到达了一个瓶颈，有一些研究[6-7]提出了术语部件的概念，希望在部件词构建的基础上准确识别领域词。文献[6-7]对计算机领域进行了术语部件的属性描述，但主要工作还在于人工总结经验，那么移植到其他领域获取领域部件词还需要大量的人工劳动。为了判断词语的领域度，常采用 tf(term frequency)×idf(inverse document frequency)及其改进公式[10]，词语的频次描述了词语的使用性，而倒排文档频度反映了词语在不同文档中的使用差异性，因此领域度和词语在领域中出现的频次 tf 成正比，和包含该词语的文档数 df 成反比。tf×idf 方法主要利用了单个词语在使用中的频度和差异判断领域度，而没有考虑词语之间的关联性。由于词语之间不是独立存在的，同一领域的词语经常一起出现，一个词语的领域性可以由与它同现的词语得到体现，因此在分析词语领域性时应考虑词语之间的关联。文献[2]根据词语的频数和与已知领域词的共现频数来发现新的领域词。进一步，文献[4]预先针对某个领域提供一些种子词作为领域词，根据其他词语与已有领域词的关系，采用 Bootstrapping 方法不断自动扩充领域词典的规模以获取领域词汇。在文献[4]中主要考虑候选词语与已选领域词的关系，其实候选词语之间也存在着互相推荐的关系，因此本文希望计算领域性时更全面地考虑词语之间的关系。

通过对当前领域词及其部件词获取方法的研究和分析，本文拟将在词语切分的基础上，对领域部件词的识别展开研究。领域词通常由通用词典中的一个或多个词语构成，例如金融领域的领域词“花旗银行”包括通用词典中的“花旗”“银行”①两个词语，这些词语在本文中称作领域部件词。如何确定哪些词语是某个领域的领域部件词是本文的研究重点。我们首先对词语之间的关联关系进行了分析和计算，由此获取词语在各个领域的使用性差异，进而根据在不同领域中的使用性差异分析词语在某个特定领域的领域性。这里通用词典的数量级在八万左右，本文在基于通用词典的切分基础上，对多个领域的文本进行领域性分析，可以同时获取多个领域的领域部件词，从而可以降低目前领域部件词描述的工作量，为提高领域词识别的性能奠定基础。由于各个领域的文本语料容易收集，同时基于通用词典的分词程序目前也比较成熟，采用以上方法，可以利用较少的资源，同时获取多个领域的领域部件词。

本文其他内容组织如下：第二部分将界定本文工作的相关概念，并概括介绍整个方案的设计。第三部分将详细介绍基于通用词典，利用词语的关联性获得词语的使用度，并进一步根据它们在不同领域的使用差异性进行领域性分析，从而获得领域部件词。第四部分主要在多个领域的语料上进行实验，通过我们提出的领域性分析方法获得各个领域的领域部件词，并与其他方法进行比较。最后第五部分对全文进行总结，并展望进一步的研究工作。

2　领域性分析相关定义及方案设计

本节主要界定论文中涉及到的相关概念，其中包括领域性分析、领域词和领域部件词的定义，并概括介绍本文获取领域部件词的整体方案。

2.1　相关定义

本文主要涉及的是词汇级的领域性分析，下面通过几个概念的定义，明确本文中所界定的领域性分析范围，以及领域性分析最终要获得的结果。

定义 1：领域性分析是指通过对一个或多个领域的语料进行分析，从而获取表示各个领域内概念、特征或关系的领域词或领域部件词，以能够区别某个领域与其他领域的不同。

由此可见，本文领域性分析的目的是要获得能够表示各个领域所特有概念和特征的领域词或领域部件词。下面分别对领域词和领域部件词进行定义。

定义 2：领域词是指某个领域中可以表示一个完整概念的词或词组，代表了某个领域的质心特征，专指性强、区别度高，能在一定程度上将该领域与其他领域区分开来。

领域词一般也称作某个领域的术语，如体育类的“世界拳击理事会，拳王”。领域词的构成通常包括一个或多个常用词语，例如“世界拳击理事会”则由“世界”、“拳击”和“理事会”三个词语构成。而构成领域词的词语可以看作是领域部件词，根据文献[7]对于术语部件的描述，我们对领域部件词定义如下：

定义 3：领域部件词是用来描述表层的、领域词

① 领域词切分后得到落单的字，也看作是通用词典中的词语。

内部构成的字或词。领域词由一个或多个领域部件词构成。

领域词的自动整理和收集经常从领域部件词的识别开始着手,由以上例子可以看出领域词可以由多个词语构成,哪些词语在某一个领域中可以用于构成显示本领域特异性的领域词是领域性分析的一个重要任务。领域部件词分为两类,第一种需要和其他词语共同构成领域词;第二种可以独立构成领域词。本文的领域性分析任务中对于两类领域部件词不进行区分。

2.2 方案设计

由于在领域性分析过程中,可用的资源通常只有通用词典,一般在对某个领域语料分词的基础上,通过计算几个词语之间的关联程度,判断和识别出领域词。由于每个切分单位为通用词或者为单字构成的词,如何判断这些切分单位是否具有领域性,即是否为领域部件词,则是本文中领域性分析的主要任务。由此,我们选择了几个不同领域的语料,对通用词在各个不同领域中的使用差异进行分析,通常来说通用词在各个领域中没有显著的使用差异性,而领域词或领域部件词会在某个特定领域体现出与其他领域的使用差异。例如,"是""了"之类的通用词在所有领域都具有很强的使用度,因此在各个领域中并没有显著的使用差异,而"公积金""贷款"等则在金融领域比在其他领域中具有较强的使用度,从而体现出使用差异性。

首先,我们主要考虑到频度和搭配获取每个词语在某个领域的使用度,使用度说明了一个语言单位在语言使用中所处的地位,一般来说当出现次数较多,和其他语言单位搭配越频繁的词语具有较高的使用度,并且每个词语的使用度也受到与其搭配的语言单位的使用度的影响。因此每个词语通过上下文中的其他词语表现出其使用度,我们把每个词语看作一个节点,其使用度受到相关联节点(即其上下文中的词语)的影响,由此形成了各词语的使用度关联网络,本文中在链接图基础上计算每个词语的使用度。当某个词语在各个领域中都具有差异不大的使用度时,通常为一个通用词,而只有在某个领域显示出与其他领域不同的使用度,该语言单位具有较强的领域性,可以作为该领域中的领域部件词。由此,我们通过计算词语在各个领域中的使用差异,获得某个领域中的领域度。领域度为领域性的体现,分值越高说明领域性越强。

通过以上分析,我们的系统在进行简单的预处理之后,领域性分析主要分为以词语链接图为基础的使用度计算和基于使用差异的领域度计算两大模块,如图1所示。首先在预处理模块中,选择不同领域的语料收集和整理,并利用通用词典对各个领域的文本进行分词处理。然后分别计算词语在各个领域中的使用度,根据每个领域的切分语料构建该领域的词语链接图,每个词语看做一个节点,词语在上下文中的关联看做节点之间的边,在词语链接图基础之上利用Pagerank算法[11]的思想,获得每个词语的使用度,并在每个领域中根据使用度的大小对词语进行使用排名。利用词语在不同领域中的使用排名情况,计算词语在各个领域中的使用差异,进而获得在某个特定领域中的领域度。最后根据领域度的大小,我们可以为每个领域选择其领域部件词。

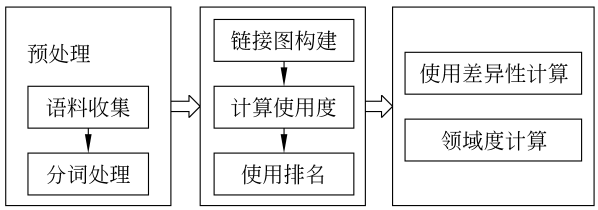

图1 系统流程图

3 基于使用差异的领域性分析

3.1 基于链接分析的使用度计算

这里主要利用每个词语在文本中的出现频度和搭配情况,来获取每个词语在某个领域的使用度。对文本切分之后,每个词语看作一个节点,如图2所示,每个节点具有一个相关的分数表示节点的使用度。和一个词语在使用上关联最密切的为文本中出现在该词语前后的两个词语,因此这些词语在使用上具有一种"投票"或"推荐"关系。两个节点之间的链接可以看作是一个节点对另一个节点在使用上的"推荐"。因此,每个节点使用度的计算则由这些推荐以及推荐节点的使用度所决定。

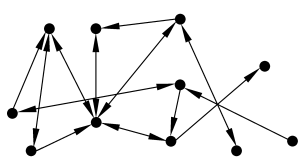

图2 词语间的链接分析图

词语使用程度链接分析图形式化表示为 $G=(V,E)$，其中 V 表示某个领域文档切分后的所有词语，E 表示词语之间相关联的边，为 $V\times V$ 的子集。并且每条边 e_{ij} 具有一个关联分数 f_{ij}，表示了两个词语的使用关联程度。e_{ij} 的设置如下，对于两个词语 i,j，如果 ij 在文中相邻顺序出现，则节点 i 和 j 之间连一条 i 到 j 的有向边。f_{ij} 的设置如下：$f_{ij}=\frac{n_{ij}}{n_i}$，其中 n_{ij} 表示 ij 在文中出现的频次，n_i 表示 i 在文中出现的频次。这样，我们就可以得到每行都被归一化的矩阵 $M_{|V|\times|V|}$ 来表示 G 中节点之间的转移概率，即 $M_{ij}=f_{ij}$。矩阵 M 每行反映了每个词语右相邻词语的使用情况，M 的转置 M^T 则反映了每个词语左相邻的情况。对 M^T 进行归一化得到 $\widetilde{M}^T$。利用矩阵 M 和 $\widetilde{M}^T$，每个词语的使用度可以由和它相邻的其他词语的使用度推导而得到，利用 Pagerank 算法的思想可以得到如下递归的计算公式：

$$U(v_i)=\mu\sum_{j\neq i}U(v_j)\cdot M_{ij}+(1-\mu)\sum_{j\neq i}U(v_j)\cdot\widetilde{M}_{ij}^T \tag{1}$$

这里 $U(v_i)$ 表示词语 v_i 的使用度分数。μ 反映了右相邻的词语对使用度的贡献程度，这里我们认为左右两边具有相同的贡献，设 μ 为 0.5。具体实现中，首先所有词语的使用度分数都设为 1，采用公式(1)，迭代计算所有词语，获得新的使用度分数。迭代过程的收敛条件为任何词语两次迭代过程中的使用度分数之差小于某个给定阈值(本文采用了 0.000 001)。

使用度给出了词语在各个领域中的使用性强弱，但由于采用语料的不同，在各个领域中使用度分数不具有较强的可比性。根据使用度对词语进行使用性排名，而词语在某个领域中的使用排名更能说明其在该领域中的使用重要性。因此对于各个领域，首先获得各个词语在该领域的使用度分数，并进行从高到低的使用性排名，使用度越高，则排名越靠前。

3.2　基于方差的领域度计算

通常某个词语在所有领域中都具有类似的使用性，则在各个领域中的使用排名没有显著差异，那么该词语的领域性则低。也就是说，词语在各个领域中的使用差异性小，则领域性低；反之，则词语的领域性较高。这儿对于使用差异性的衡量，主要利用在各个领域中使用排名的方差得到。

首先计算词语在不同领域中使用排名的总方差，设共有 N 个词语，每个词语为 $P_i(1\leqslant i\leqslant N)$，共有 M 个不同的领域，在每个领域 j 中的使用排名分别为 $Rank_j(P_i)$ $(1\leqslant j\leqslant M)$，其中 $\bar{R}(P_i)$ 表示了 P_i 在 M 个领域中使用排名的平均值，由此可以得到 P_i 的使用排名差异 $Var(P_i)$：

$$Var(P_i)=\frac{\sum_j(Rank_j(P_i)-\bar{R}(P_i))^2}{M}$$
$$\bar{R}(P_i)=\frac{\sum_j Rank_j(P_i)}{M} \tag{2}$$

当某个词语的使用排名差异具有较大的值时，则该词语可能为领域词，但具体属于哪个领域，还无法判定。通常一个词语在其他领域中没有显著的使用差异，而只在某个特定领域中具有和其他领域较显著的差异，则认为该词语属于该领域。即，当评定某个词语 P_i 是否属于领域 j 时，在除 j 以外的其他 $M-1$ 个领域中，计算这 $M-1$ 个领域的使用排名差异 $Var'_j(P_i)$：

$$Var'_j(P_i)=\frac{\sum_{k\neq j}(Rank_k(P_i)-\bar{R}_{\neq j}(P_i))^2}{M-1}$$
$$\bar{R}_{\neq j}(P_i)=\frac{\sum_{k\neq j}Rank_k(P_i)}{M-1} \tag{3}$$

由此可知，某个词语的总方差 $Var(P_i)$ 越大，则领域性越强；而不含某个领域 j 的其他领域排名方差 Var'_j 越小，领域性越强。因此当我们衡量某个词语 P_i 在领域 j 中的领域度 $Domain_j(P_i)$ 时，采用如下的计算公式：

$$Domain_j(P_i)=\frac{Var(P_i)}{Var'_j(P_i)} \tag{4}$$

4　实验

4.1　实验数据和评测标准

为了评测本文中提出的方法，我们从一些网站上收集了不同领域的文本语料，其中包括军事、体育、金融、娱乐四个不同的领域，每个领域各有 100 篇文档。首先对收集到的原始文本语料进行处理，把内容无关的标签和文字去除，然后利用分词工具对语料进行切分。在切分语料的基础上，对每个

领域采用以上方法进行领域性分析，并对各个领域中的词语进行领域度的排名。

为了评测和对比本方法，我们采用常用的tf(term frequency 频次) × idf(inverse document frequency 倒排文档频度)方法对领域度进行衡量，这儿利用了tf×idf[10]的改进公式：

$$Domain_j(P_i) = (1+\log(tf_j(P_i))) \times \log(T/df(P_i)) \quad (5)$$

其中 $Domain_j(P_i)$ 给出了词语 P_i 在领域 j 中的领域度，$tf_j(P_i)$ 表示词语 P_i 出现在领域 j 中的频数，T 表示全部的文档数目，$df(P_i)$ 表示出现 P_i 的文档数目。

分别利用本文提出的方法和tf×idf方法，对每个领域的所有词语计算领域度，并根据词语的领域度大小选取前 L 个认为是本领域的领域部件词。由于缺少领域部件词的标准答案进行评测，我们采用人工评测的方法，由5名计算语言学领域的同学或老师对结果进行评定，从每个领域中选出的前 L 个部件词中判定哪些词语为正确的领域部件词，每个结果平均由3人评测。这儿我们制定了两种评测标准：第一种为**严格准确率** $Accuracy_{strict}$，只有三人达成一致意见认为可以作为领域部件词，才为正确的领域部件词。第二种为**宽松准确率** $Accuracy_{lenient}$，三人中只要有多数人即两个人以上认为是领域部件词则为正确的领域部件词。根据每种方法在不同领域中得到的结果，计算其平均值。两种评测指标如下：

$$Accuracy_{strict} = \frac{\sum_j (|S_{j1} \cap S_{j2} \cap S_{j3}|/L)}{M}$$

$$Accuracy_{lenient} = \frac{\sum_j (|S_{j1} \cup S_{j2}| + |S_{j1} \cup S_{j3}| + |S_{j2} \cup S_{j3}|)/L}{M} \quad (6)$$

其中 M 为所选择领域的个数，S_{j1}，S_{j2}，S_{j3} 表示对领域 j 中提取出来的 L 个词语3个评测人分别判定为正确的领域部件词集合。

4.2 实验结果

利用基于使用差异的领域度计算方法，在军事、体育、金融、娱乐四个领域中进行了实验，表1分别列出了四个领域的领域部件词示例。

表1 本文方法提取的领域部件词示例

金 融	娱 乐	军 事	体 育
财务部	经纪人	伊拉克	公开赛
房地产	好莱坞	核武器	夺冠
华尔街	演艺圈	国防部	犯规
经济体	票房	总参谋部	威廉姆斯
印花税	模特儿	护卫舰	埃蒙斯
赤字	歌迷	航天员	艾弗森
呆坏账	电影节	核爆炸	奥运村
股份制	成名	缔约国	裁判员
美联储	音乐人	轰炸机	篮板球
国资委	参演	空降兵	世界杯
金融业	贺岁	黎巴嫩	世乒赛
进出口	出品人	特种部队	黄牌
花旗	创作	综合国力	半决赛
撤资	第三者	阿富汗	净胜球
百分点	史泰龙	超低空	蝉联
中产阶级	恩爱	渤海湾	亚运会

表2、表3、表4比较了本文方法和tf×idf方法在四个领域中的评测结果，第二列和第三列分别表示金融、娱乐、军事、体育四个领域利用本文方法选出的前 L 个(在3个表中分别为100、200、300)词语中，3人和2人以上认可的领域部件词个数和准确率，第四列和第五列则分别表示利用tf×idf方法，3人和2人以上认可的领域部件词个数和准确率。

从表2、3、4中可以看出，本文方法和tf×idf方法获得领域词的准确率都随着 L 的增大而不断下降。而本文方法对于获取的前 L 个领域部件词，比tf×idf方法的严格准确率高10～12个百分点左右，宽松准确率高18～20个百分点左右。由此可以看出，本文方法比常用的tf×idf方法，在获取领域部件词方面具有较强的优势。

我们对实验结果进行分析，由于tf×idf方法只考虑到词语的频度信息，当领域部件词在某个领域的语料中出现频度较少或均匀出现在各个文档中时，会得到较低的领域度分数。而本文方法除了考虑词语本身的频度信息外，还考虑上下文中和其他词语的推荐关系，如果上下文中其他词语具有较高的使用性，则该词语在本领域中也具有较高的使用性，进而显示出该词语在不同领域具有较大的使用差异性，获得较高的领域度。例如娱乐领域的“出品

表 2　各领域前 100 个领域部件词的评测结果

	本文方法		tf×idf 方法	
	3 人一致	2 人以上一致	3 人一致	2 人以上一致
金融	33	56	33	46
娱乐	39	62	28	38
军事	43	75	33	70
体育	64	84	36	51
总计	179	277	130	205
$Accuracy_{strict}$	44.75%	—	32.5%	—
$Accuracy_{lenient}$	—	69.25%	—	51.25%

表 3　各领域前 200 个领域部件词的评测结果

	本文方法		tf×idf 方法	
	3 人一致	2 人以上一致	3 人一致	2 人以上一致
金融	54	89	43	62
娱乐	62	121	50	71
军事	79	145	52	119
体育	107	146	59	86
总计	302	501	204	338
$Accuracy_{strict}$	37.75%	—	25.5%	—
$Accuracy_{lenient}$	—	62.625%	—	42.25%

表 4　各领域前 300 个领域部件词的评测结果

	本文方法		tf×idf 方法	
	3 人一致	2 人以上一致	3 人一致	2 人以上一致
金融	75	126	54	79
娱乐	83	169	76	112
军事	100	201	70	156
体育	153	208	94	135
总计	411	704	294	482
$Accuracy_{strict}$	34.25%	—	24.5%	—
$Accuracy_{lenient}$	—	58.67%	—	40.17%

人”、金融领域的“花旗”、军事领域的“超低空”等，在我们所使用的语料中，通过 tf×idf 方法计算得到的领域度较低，而采用本文方法可以得到较高的领域度。

同时，表 5 在金融、军事、体育三个专业领域，比较了本文方法和文献[4]中所采用的 Bootstrapping 方法。根据文献[4]的试验结果，其中选取了前 500 词进行评价。文献[4]没有给出计算正确率的方法，我们给出了本文方法的严格和宽松正确率。从表 5 的结果可以看出，只有金融领域的严格正确率低于文献[4]中的结果，这是因为金融领域与通用领域更为密切相关，以至于有些部件词较难判断，使

得全部评定人没有达成一致，但宽松正确率还是超过了 Bootstrpping 方法。

表 5 本文方法和 Bootstrpping 方法的比较

领域	Bootstrapping 方法	本文方法	
		$Accuracy_{strict}$	$Accuracy_{lenient}$
金融	42.8%	30%	50%
体育	36.6%	41.6%	58.4%
军事	21.6%	35.2%	67.8%

本文方法的结果在一定程度上依赖于所选取的语料，例如，娱乐领域中“上月”和“回归”计算得到较高的领域度，这是因为所选语料为《娱乐快报》(新闻)，在语料中这些词与其他领域部件词联系较多，如“上月某明星…”、“某剧回归…”等。我们认为如果扩大语料规模和范围可以减少这种上下文中的关联，从而提高提取领域部件词的准确率。

此外，由于采用人工方法评测，没有标准答案可依，主要依据评测人的判断，因此对评测结果也造成一定的影响。例如娱乐领域中的“城城”，表示歌星“郭富城”的昵称，但由于标注人对于领域知识的掌握程度不同，会影响到评测领域部件词的准确率。

5 结论

本文提出了一种基于使用差异的领域分析方法，其中利用较少的资源，只需要利用各个领域的文本语料和一个通用分词程序，就可以获取各个领域的领域部件词。该方法主要在词语的基础上利用词语之间的上下文关系，引入链接分析方法获得词语在各个领域中的使用重要性，并通过词语在各个领域中的使用差异计算其领域度，从而达到领域分析的目的，获取某个领域的领域部件词。我们在军事、娱乐等领域进行了实验，实验结果表明该方法相对于当前常用的 tf×idf 方法，可以更有效地进行领域分析。该方法根据词语在多个领域的使用排名进行使用差异性的计算，可以达到同时对多个领域进行领域分析的目的。本文方法的一个缺陷在于，计算词语的领域度时需要预先确定好要分析的几个领域，当词语在一个领域中的使用排名发生变化时，将影响到该词语的领域度结果。此外，目前语料的收集对于领域度也会产生一定的影响，语料规模较小会影响到词语的使用排名，从而影响领域度的计算。不过以上两个问题在语料规模达到一定程度时，领域度的计算也将处于稳定状态。我们在下一步工作中，将进一步扩大语料规模，对该方法进一步的实验和评测。同时，也将进一步研究如何将领域部件词识别的结果应用到领域词的获取中。

参考文献：

[1] 黄玉兰，龚才春，许洪波，程学期. 基于伪相关反馈模型的领域词典生成算法[J]. 中文信息学报，2008，22(1)：111-115.

[2] 凌祺，樊孝忠. 领域词汇自动获取的研究[J]. 微机发展，2005，15(8)：148-150.

[3] 孙霞，郑庆华，王朝静，张素娟. 一种基于生语料的领域词典生成方法[J]. 小型微型计算机系统，2005，26(6)：1088-1092.

[4] 陈文亮，朱靖波，姚天顺，等. 基于 Bootstrapping 的领域词汇自动获取[C]//语言计算与基于内容的文本处理. 北京：清华大学出版社，2003.

[5] 傅骞，魏顺平，王斌，路秋丽. 教育技术领域术语提取研究[J]. 现代教育技术，2008，18(5)：60-65.

[6] 何燕，穗志方，段慧明，俞士汶. 一种结合术语部件库的术语提取方法[J]. 计算机工程与应用，2006，42(33)：4-7.

[7] 吴云芳. 信息科学与技术领域术语部件描述[J]. 语言文字应用，2003，(4)：34-39.

[8] Wilson Wong, Wei Liu, Mohammed Bennamoun. Determining termhood for learning domain ontologies using domain prevalence and tendency[C]//Proceedings of the sixth Australasian conference: Data mining and analytics. Gold Coast, Australia, 2007.

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

[10] Christopher D. Manning. Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval[M]. 2008.

[11] Amy N. Langville, Carl D. Meyer. Deeper inside pagerank [J]. Technical Report, NCSU Center for RES SCI Comp. 2003.