·人工智能及识别技术· 文章编号：1000-3428(2016)08-0160-06 文献标志码：A 中图分类号：TP311

一种面向军事文本的领域特征词向量描述方法

秦 杰，曹 雷，彭 辉，赖 俊

（解放军理工大学 指挥信息系统学院，南京 210007）

摘 要：针对军事文本信息中命名实体多、特征词领域性强的特性，提出一种领域特征词向量描述方法。从优化分词和领域特征词筛选方面压缩向量空间，完善时间、地名、部队名称和武器装备 4 类重要命名实体的提取规则，扩充分词词典库。改进领域相关度和领域一致度相结合的领域特征词筛选算法，突出领域特征词与常用词汇之间的差别，进一步过滤领域特征词。实验结果表明，优化分词后，该方法能够提取出军事文本中的命名实体和部分专有词汇，降低特征词数量，改进后的领域特征词筛选算法将准确率和召回率分别提高 20% 和 16.7%，提出的领域特征词向量描述方法所生成的特征词向量具有较强的领域性。

关键词：军事文本；命名实体；向量空间；分词；领域特征词

中文引用格式：秦 杰，曹 雷，彭 辉，等. 一种面向军事文本的领域特征词向量描述方法[J]. 计算机工程，2016，42(8)：160-165.

英文引用格式：Qin Jie，Cao Lei，Peng Hui，et al. A Domain Feature Word Vector Description Method for Military Texts[J]. Computer Engineering，2016，42(8)：160-165.

A Domain Feature Word Vector Description Method for Military Texts

QIN Jie，CAO Lei，PENG Hui，LAI Jun

（College of Command Information System，PLA University of Science and Technology，Nanjing 210007，China）

【Abstract】 According to the large number of named entities and deep domain of feature words in military text information，this paper proposes a vector description method for domain feature words. It compresses the vector space through the optimization of word segmentation and domain feature word selection，improves the extraction rules for four important types of named entity，including time，place name，troop name and weapon equipment，and extends the word segmentation dictionary library. It improves the domain feature word selecting algorithm combining domain relevance and domain consistency，enlarges the difference between domain words and common words，and further filters domain feature words. Experimental results show that after optimizing word segmentation，the named entities and some specific vocabulary in military texts can be extracted，and the number of feature words can be reduced. The accuracy and recall rate of the improved domain feature word selecting method are increased by 20% and 16.7% respectively. The feature word vector generated by the proposed method has strong domain feature.

【Key words】 military text；named entity；vector space；word segmentation；domain feature word

DOI：10.3969/j.issn.1000-3428.2016.08.029

1 概述

信息资源建设是当前军事信息化建设的核心内容，军事文本信息作为一种主要的信息资源，是维系指挥信息系统高效运转的关键。文本向量化是文本信息常用的处理方式，通过将军事文本信息转化为简洁、完备的特征词向量，有利于实现信息资源的分类、索引、推送等应用，从而促进军事文本信息资源的高效利用。

通常的文本向量化方法所生成的特征词向量往往具有较高维度，既占用大量存储资源，也不利于计算处理。因而，降低特征词向量的维度成为文本处理的核心。特征降维的主要方法包括特征选择和特征提取[1]。其中，特征选择是指从特征词总集中筛选出有用的词条组成特征子集，典型做法就是筛选领域特征词作为特征子集。领域特征词筛选常用的方法包括基于互信息的方法[2]、基于频率的方法[3]、领域相关度和领域一致度相结合的方法[4]等。

作者简介：秦 杰(1990－)，男，硕士研究生，主研方向为文本推荐；曹 雷(通讯作者)，教授；彭 辉，讲师、博士；赖 俊，讲师、硕士。

收稿日期：2015-06-18 修回日期：2015-08-11 E-mail：caolei.nj@163.com

文献[5]指出领域相关度和领域一致度相结合的方法效果较好。但是,通过分析发现,由于部分常见特征词在领域内具有较大数量,且分布均匀,导致特征词领域一致度值较大,很容易被误选入领域特征词集;此外,特征选择方法的对象是特征词集,合理的文本分词是特征选择的基础。军事文本信息中存在大量长特征词,主要包括时间、地名、部队名称、武器装备等命名实体[6]和部分领域专用术语,是常规分词方法很难获取的内容。

针对上述问题,在特征词筛选前,可以对分词进行优化,避免长特征词被切分,从源头降低向量维度。基于此思路,本文提出一种面向军事文本的领域特征词向量描述方法,综合应用优化分词方法和改进后的领域特征词筛选算法,实现军事文本信息中命名实体和部分领域专有词汇提取。基于特征词频次进行非线性运算,放大领域与常用词汇之间的差异,进一步剔除领域无关词,实现军事文本信息特征词向量的优化。

2 分词优化

中文文本分词是中文信息处理的基础,现有的分词算法主要有 3 大类:基于字符串匹配的分词方法,基于理解的分词方法和基于统计的分词方法[7]。这些分词方法将文本分词后得到的词集基数往往较大。其中一个主要原因就是许多不应该切分的长特征词被切分成几个短词汇。在分词前,识别出这类长特征词,将大大降低文本词汇集合大小,同时词汇也更有意义。为了实现长特征词的提取,在正向最大匹配(Forward Maximum Matching,FMM)分词的基础上,考虑军事领域特性,从命名实体识别和词典库扩充两方面进行优化。

2.1 命名实体识别

军事文本信息中广泛存在时间、地名、部队名称、装备型号等命名实体,这几类命名实体在军事领域具有重要意义,它们通常属于长特征词,是常规分词很难获取的内容。

对于命名实体识别,文献[8]对军事文本信息中的时间、地名、部队名称和专用词语用词特点进行了分析;文献[9]采用基于有限状态自动机的命名实体识别方法,实现了对数字、时间、距离和部队名称的识别。对于军事装备类实体,文献[10]提出字母符号串和汉字串的概念,采用定性和定量的方法对军事文献中的字母串和汉字串进行分析,但目前缺乏军事装备类实体名相关提取规则。

2.1.1 识别规则补充

作为军事领域的一类重要实体,武器装备通常由“型号 + 类型”组成,如“95 式步枪”、“DF-21 中程弹道导弹”。其中,型号是反映武器装备类别和特征的一组汉语拼音字母和阿拉伯数字,类型是武器装备的军事术语。除了常规的名称表示外,军事文本中还经常出现一些简称,如“东风-21”、“歼-10”等。型号与类型之间常常还有特定汉字,如“式、级、型、号”等。根据上述对武器装备的分析,总结出武器装备实体识别规则,如表 1 所示。

表 1 武器装备识别规则

规则	例子
<武器>=<型号字符串> <武器装备类型>	DF-21 中程弹道导弹
<武器>=<型号字符串> <特定汉字><武器装备类型>	86 式步兵战车
<武器>=<武器系列> <型号字符串>	东风-21
<武器>=<武器系列><型号字符串><武器装备类型>	红旗-16 舰空导弹

在表 1 中,“型号字符串”通常由字母、数字和短横线组成;“武器装备类型”种类相对较多,本文从文献[11]中获取了武器装备的详细分类,包含了轻武器、火炮、坦克、车辆、地雷、飞机、舰艇、导弹、雷达等 20 个类别的武器装备;“特定汉字”包含“式、级、型、号、厘米、毫米、管”;“武器系列”是成系列的武器装备别名,包含“东风、红旗、巨浪、歼、轰、运、直等”。

2.1.2 武器装备识别

根据表 1 中武器装备的命名规则,设计了识别武器装备的有限状态自动机(Finite State Automata,FSA)[12],如图 1 所示。a,b,c,d 分别代表<武器系列>、<型号字符串>、<特定汉字>、<武器装备类型>输入单元。S_0 代表初始状态,S_1,S_2,S_3,S_4 表示接收到相应输入单元后所达到的状态。

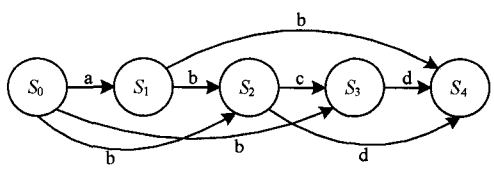

图 1 识别武器装备的 FSA

武器装备实体在文本中可能出现的每一种情况,图 1 中都会有一条从 S_0 到 S_4 的路径与之对应。依据这些路径,就可以实现对武器装备实体的识别。其中,需要的知识库包括:

(1)常用的表示武器系列的词表,如东风、红旗、巨浪、歼、轰、运、直等。

(2)特定汉字表,如式、级、型、号、厘米、毫米、管等。

(3)武器装备类型表,如手枪、自动步枪、战斗机、雷达、空空导弹等。

2.2 扩展的词库

命名实体识别后,采用正向最大匹配(Forward Maximum Matching,FMM)算法[13]对文本信息进行分词。该算法的基本思想是:按最大匹配方式从文本中找出所有出现在词典库中的词,此算法复杂度相对较低,其核心是词典库的容量,词典库的完备性直接影响分词的准确性。

军事文本中涉及大量领域词汇,采用通用词典库分词必然导致漏分、错分的情况发生。为此,本文对通用的词典库进行了扩充,扩充词源来自中国军事科学院 2011 年出版的《中国人民解放军军语》,共 8 587 条词语,包含作战、指挥、训练、体制编制、信息化建设、武器装备和军事技术等方面的领域术语。

2.3 分词基本流程优化

经过优化后,分词包含 2 个步骤:(1)命名实体识别。根据命名实体规则库,识别出时间、地名、部队名称和武器装备 4 类主要实体;(2)正向最大匹配分词。在识别命名实体后,进一步采用正向最大匹配分词方法对文本进行分词处理。最后,将命名实体和分词结果合并作为文本特征词集合。优化分词基本流程如图 2 所示。

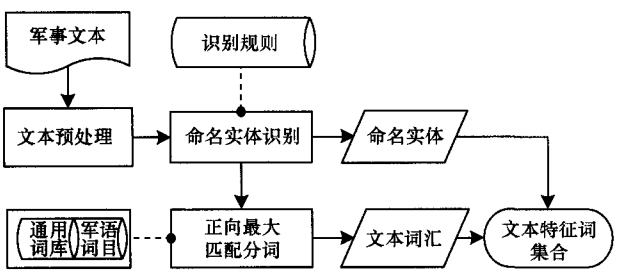

图 2 优化分词基本流程

3 领域特征词筛选

经过分词处理之后,得到的特征词集合中往往含有大量领域无关词。这些无关词在各个领域中均大量出现,严重影响了领域特征分析。因此,剔除领域无关词是特征词向量压缩的重点。

3.1 结合领域相关度与一致度的算法

领域特征词筛选的基本思想是:采用一定算法计算每个特征词对领域的重要程度,并通过预设的阈值,剔除掉低于阈值的特征词。目前,领域一致性和相关性相结合的方法在领域特征词筛选时,应用效果较好[5]。其中,领域相关度和领域一致度的定义如下。

定义 1 领域相关度反映特征词与领域相关程度。假设领域集合 $set=\{D_1,D_2,\cdots,D_n\}$ 为一系列领域,则特征词 t 对于领域 D_i 的相关度为[14]:

$$DR(t,D_i)=\frac{P(t|D_i)}{\sum_{i=1}^{n}P(t|D_i)} \tag{1}$$

其中,条件概率 $P(t|D_i)$ 可用频率估计:

$$P(t|D_i)\approx\frac{f_{t,i}}{\sum_{t'\in D_i}f_{t',i}} \tag{2}$$

其中,$f_{t,i}$ 是特征词 t 在领域 D_i 中的频率;$f_{t',i}$ 是特征词 t' 在领域 D_i 中的频率。

定义 2 领域一致度反映特征词在领域中的分布均匀度。设 d_j 是领域 D_i 中的一篇文档,则特征词对于领域 D_i 的一致度为[14]:

$$DC(t,D_i)=H(t,D_i)=\sum_{d_j\in D_i}\left(P(t,d_j)\cdot\log_a\frac{1}{P(t,d_j)}\right) \tag{3}$$

其中,联合概率 $P(t,d_j)$ 可用频率估计:

$$P(t,d_j)\approx\frac{f_{t,d_j}}{\sum_{d_j\in D_i}f_{t,d_j}} \tag{4}$$

其中,f_{t,d_j} 表示特征词 t 在领域文档 d_j 中出现的频率。

领域相关度和一致度都是反映特征词对领域的重要程度,值越大,越有可能为领域特征词。领域相关度和一致度相结合的特征词筛选算法通过对两者加权平均,实现对特征词的综合评判,综合指标如下:

$$DW(t,D_i)=\alpha\cdot DR(t,D_i)+\beta\cdot DC(t,D_i) \tag{5}$$

当综合评定值 $DW(t,D_i)$ 大于等于预设阈值 θ 时,将 t 选作领域 D_i 特征词。参数 α 和 β 反映相关度和一致度各自所占的比重。文献[15]实验结果表明,α 取值 0.9 左右,β 取 0.25 ~ 0.35 时过滤效果较好。

依据领域相关度与领域一致度相结合的算法,采用复旦大学提供的全部测试文本语料库进行实验,将军事文本作为领域语料,取 $\alpha=0.9$,$\beta=0.3$。

实验筛选出 500 个特征词,其中,领域特征词占 324 个。分析发现,大量入选的非领域特征词都有一个共同点,即在领域中出现次数相对较多,且分布均匀,大部分属于常用词汇。这一类特征词的领域一致度值较大,导致综合值也较高,很容易被误选入领域特征词集中。

3.2 算法改进

领域相关度与领域一致度相结合的特征词筛选算法,在处理各领域出现均较多的常见词汇时,准确率不高。其主要缺陷是算法对领域特征词与常用特征词的区分不够明显,在领域一致度计算时,算法考虑的是特征词的分布均匀度。此时,领域特征词和常用词汇具有同样的特征。而特征词的领域一致度

值往往大于领域相关度值，所以部分常用词汇就很容易被误选入领域特征词集。

改进的基本思路是：利用特征词频次对特征词综合值进行非线性运算，从而实现领域特征词与常见特征词之间的差别放大。此时，可对综合值计算公式进行如下修正：

$$DW(t,D_i)=\alpha\cdot DR(t,D_i)+\frac{\lambda\cdot persent}{\lambda^{persent}}\cdot\beta\cdot DC(t,D_i)+persent^{-1}\cdot\log_a(persent^{-1}+1) \quad (6)$$

$$persent=\frac{all}{all_{\text{domain}}} \quad (7)$$

其中，all 表示特征词 t 在所有领域出现的频次；all_{domain} 表示特征词在测试领域出现的频次；λ 为参数，根据实际放大需要选定；$\lambda\cdot persent/\lambda^{persent}$ 通过对特征词在领域与非领域出现的频次操作，实现了对特征词领域一致度的非线性计算，放大了领域特征词与高频常用词汇之间的差别；最后一项 $persent^{-1}\cdot\log_a(persent^{-1}+1)$ 主要用于修正 $persent$ 等于 1 时的综合值，当某特征词仅在领域中出现时，$persent$ 等于 1。这类特征词通常有 3 种：领域特征词，计量词和命名实体。其中，领域特征词出现次数一般高于 1 次；计量词和命名实体出现次数多为 1 次，如“3 架”、“2011 年 7 月 4 日”等。

4　军事文本领域特征词向量描述方法

文本信息向量空间描述就是将文本表示成一个特征词向量，向量的每一个维度由特征词与其权值组成。为突显命名实体在军事领域的地位，进一步剔除掉无关词汇，在结合上述优化分词和改进后领域特征词筛选算法的基础上，提出一种面向军事文本的领域特征词向量描述方法。具体算法如下，其中，d_j 为某文本内容，k 为整数，θ 为预设阈值。

(1)识别 d_j 中命名实体，形成词集 s_1，剩余文本内容为 d'_j；

(2)扩展词典库 D，形成词库 D'；

(3)借助词库 D'，对 d'_j 正向最大匹配分词，形成词集 s_2；

(4)采用改进后领域相关度与一致度相结合的算法计算 s_2 中特征词的综合值；

(5)筛选出 s_2 中综合值大于预设阈值 θ 的特征词，形成 s_3；

(6)剔除 s_3 中特征词领域出现次数低于 k 的词，形成 s_4；

(7)获取 s_4 中特征词的最大综合值 ω；

(8)设置 s_1 中所有特征词的综合值为 ω；

(9)将词集 s_1 与 s_4 合并形成词集 s_5；

(10)将 s_5 中特征词表示成向量。

时间、地名、部队名称和武器装备 4 类命名实体在军事领域有着极其重要的意义，算法在分词前进行命名实体提取，并强制将其加入到特征词向量，有效提高了分词的准确率，加强了特征词向量的领域特性。在此基础上，针对仅在领域中出现的特征词（$persent$ 等于 1），算法在强制选取命名实体后，通过低频阈值 k 对计量词和命名实体进行清除，进一步提高了特征词的领域度。

5　实验结果与分析

为验证本文方法的有效性，分别进行了分词优化实验、领域特征词筛选实验和特征词向量生成实验。

5.1　分词优化实验

本文实验用于验证优化分词对降低特征词数量的效果，分词方法基于 Lucene[16] 中文分词组件 je-analysis-1.5.1.zip 实现，该组件采用正向最大匹配方式分词，包含词目共 217 413 条。

为便于量化分词优化的效果，定义特征词压缩率 P：

$$P=\frac{W_{\text{FMM}}-W_{\text{I-FMM}}}{W_{\text{FMM}}}\times 100\% \quad (8)$$

其中，W_{FMM} 为采用 FMM 分词所得的特征词数量；$W_{\text{I-FMM}}$ 为分词优化（记为 I-FMM）后所得的特征词数量。

实验分别选择军事新闻报道、军事科技、武器装备介绍和作战文书 4 类军事文本各 1 篇文本进行分词实验。分词结果如表 2 所示。

表 2　分词优化结果

文本类型	FMM	I-FMM	压缩率/%
军事新闻	216	198	8.3
军事科技	364	282	22.5
武器介绍	348	265	23.9
作战文书	312	214	31.4

从分词结果可以发现，经过分词优化，所得到的特征词数量明显降低，特征词向量维度有了较大压缩。其原因在于命名实体识别和词库的扩展避免了大量的领域长特征词被误分。其中，军事科技、武器介绍和作战文书压缩效果尤其明显，是因为军事科技文本中涉及大量的军事领域术语，而武器介绍文本中有大量武器信息，作战文书中有时间、地名和部队名称等命名实体。

5.2　领域特征词筛选实验

为验证改进后领域概念筛选算法的有效性，采用复旦大学测试语料库[17]作为支撑，进行了算法实验。语料库包含军事、交通、历史、经济等 20 个领

域,共计 9 833 篇文档。将军事文本作为领域语料,其他类文本作为过滤语料。根据式(5)和式(6),计算出军事领域文本语料中的特征词(t)的综合值并绘图,结果如图 3、图 4 所示。其中,综合值超过 2 的特征词用方框标注,综合值小于或等于 2 的特征词用圆圈标注。

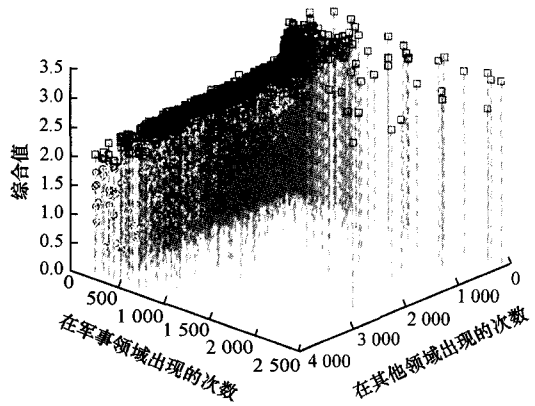

图 3 改进前综合值计算结果

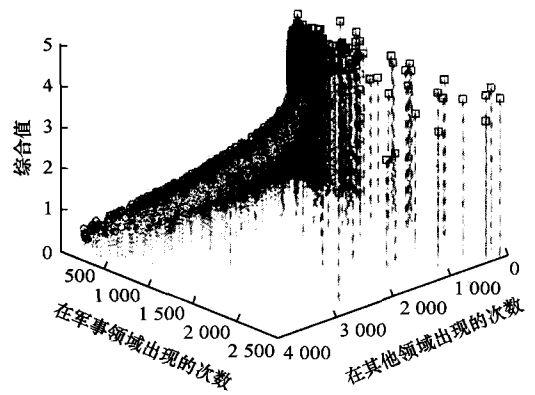

图 4 改进后综合值计算结果

从图 3 可以看出,改进前的领域特征词筛选算法计算综合值时,领域特征词与非领域特征词的综合值分布差别不够明显。有较多的特征词在其他领域分布较多,也有较高的综合值,这类特征词大多属于常用词汇;图 4 表明,改进后的筛选算法能较好地降低常用词汇的综合值,并且没有对领域特征词综合值产生大的影响,有助于筛选领域特征词。

进一步,以百度百科关于歼-10 的简介作为测试文本,采用人工标记的方法,标记了文本中的 24 个领域特征词。将本文改进后的筛选算法、改进前的筛选算法以及 TFIDF 特征词选择算法进行对比分析,设置相关度参数 α 等于 0.9,一致度参数 β 等于 0.3,低频阈值 k 取 0,放大参数 λ 取 2。实验结果如表 3 所示。评价指标包括准确率 $precision$、召回率 $recall$、$F_{measure}$ 值,具体计算方式如下[4]:

$$precision=\frac{correct_{extracted}}{all_{extracted}} \tag{9}$$

$$recall=\frac{correct_{extracted}}{all_{corpus}} \tag{10}$$

$$F_{measure}=\frac{2precision\cdot recall}{precision+reall} \tag{11}$$

其中,$correct_{extracted}$ 表示筛选得到的领域特征词数;$all_{extracted}$ 表示得到的所有特征词数;all_{corpus} 表示文本中所有领域特征词数。

从表 3 中各指标可以发现,TFIDF 与改进前的领域特征词筛选算法具有相当的选择效果,领域特征词筛选性能较差,但造成此结果的原因却不同。其中,TFIDF 是由于关注点为特征词的文档性质,对领域性关注较弱,而改进前的领域特征词筛选算法误选了常用词汇。改进后的筛选算法在领域特征词的准确率、召回率和 F 值都有显著提高。但是,3 个指标仍然偏低,其原因主要有 2 个:(1)文本中部分特征词被切分,如“涡扇发动机”、“单座单发”、“中国航空博物馆”等;(2)实验采用的军事领域语料主要是相关新闻报道,对军事领域涵盖不够全面,可从完备词典库和扩充领域语料库 2 个方面改善召回率偏低的问题。

表 3 筛选结果对比

参数	TFIDF	原筛选算法	改进后的算法
all_{corpus}	24	24	24
$all_{extracted}$	20	20	20
$correct_{extracted}$	13	12	16
准确率/%	65.0	60.0	80.0
召回率/%	54.2	50.0	66.7
F 值/%	59.1	54.5	72.7

5.3 特征词向量生成

在上述参数设置条件下,基于本文提出的军事文本信息特征词向量描述方法,对“歼-10”文本内容进行建模。总共筛选出了 20 个特征词组成测试文本内容的特征词向量,具体结果如下:<(歼-10 战斗机,2.05)(歼-10,2.05)(中国人民解放军空军第 44 师132 团,2.05)(2004 年 1 月,2.05)(2009 年 11 月 5 日,2.05)(J-10,1.9)(F-10,1.9)(单座,2.05)(火鸟,1.9)(1001 号,1.9)(战斗机,1.0)(中航,0.86)(展板,0.85)(北约,0.83)(首飞,0.81)(大推力,0.79)(单发,0.76)(全天候,0.73)(空军,0.69)(装备,0.64)(图文,0.61)>。

从中可以看出:本文提出的军事文本信息特征词向量描述方法能够识别出时间、部队名称和装备等命名实体,并将其加入到向量特征词。但存在一定不足,无法对特征词之间存在同义关系进行处理,如“歼-10 战斗机”、“J-10”、“F-10”、“歼-10”实际表达相同语义,但算法未进行合并。

6 结束语

本文提出一种面向军事文本信息的特征词向量描述方法,采用先优化分词结果再筛选特征词的思路构建特征词向量。通过命名实体识别和扩充词典库,优化军事文本信息的分词结果,利用领域特征词与常用词汇在测试领域和其他领域出现频次的差异,改进领域特征词筛选算法,剔除误选常用词汇。实验结果表明,本文方法所生成的特征词向量简洁、完备,能够识别出军事文本中的命名实体,领域特征词筛选效果明显。在此基础上,下一步工作可以针对特征词之间的语义问题,通过构建领域本体,找出特征词之间的同义、包含、从属等语义关系,降低向量维度。

参考文献

[1] 杨杰明,刘元宁,曲朝阳,等. 文本分类中基于综合度量的特征选择方法[J]. 吉林大学学报:理学版,2013,51(5):887-893.

[2] 吴海燕. 基于互信息与词语共现的领域术语自动抽取方法研究[J]. 重庆邮电大学学报:自然科学版,2013,25(5):690-693.

[3] 李江华,时 鹏,胡长军. 一种适用于复合术语的本体概念学习方法[J]. 计算机科学,2013,40(5):168-172.

[4] 傅 鹂,黄利强,付春雷 一种改进的面向文本的领域概念筛选算法[J]. 计算机科学,2012,39(6):253-256.

[5] Velardi P, Missikoff M, Basili R. Identification of Relevant Terms to Support the Construction of Domain Ontologies [C]//Proceedings of the Workshop on Human Language Technologies and Knowledge Management. New York, USA: ACM Press, 2001: 1-8.

[6] Sundheim B M. Named Entity Task Definition [C]// Proceedings of the 6th Message Understanding Conference. Berlin, Germany: Springer, 1995: 319-332.

[7] 张冰怡,魏 博,陈建成,等. 基于对偶编码的中文分词算法[J]. 南京理工大学学报,2014,38(4):526-530.

[8] 杨晓冬,邵根富. 基于本体的作战文书分词的关键技术研究[D]. 杭州:杭州电子科技大学,2013.

[9] 张广军,王建宁. 基于 XML 作战文书理解关键技术研究[D]. 南京:南京理工大学,2009.

[10] 杨 森. 军事文献中复杂字母词语的形式分析[J]. 社会纵横,2010(3):315-316.

[11] 张海泉. 武器家谱[J]. 当代军事文摘,2005(3):21-22.

[12] 赵 伟,夏庆锋. 一种基于有限状态自动机的多鱼协作顶球算法[J]. 兵工自动化,2012,31(11):59-62.

[13] 王惠仙,龙 华. 基于改进的正向最大匹配中文分词算法研究[J]. 贵州大学学报:自然科学版,2011,28(5):112-115.

[14] Navigli R, Velardi P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites [J]. Computational Linguistics, 2004, 30(2): 151-179.

[15] 贾秀玲,文敦伟. 面向文本的本体学习中概念提取及关系提取的研究[D]. 长沙:中南大学,2007.

[16] 邱 哲,符滔滔,王学松. 开发自己的搜索引擎 Lucene + Heritrix[M]. 2 版. 北京:人民邮电出版社,2005.

[17] 张玉芳,杨 芬,熊忠阳,等. 基于上下文的领域本体概念和关系的提取[J]. 计算机应用研究,2010,27(1):74-76.

编辑 顾逸斐

(上接第 159 页)

[6] Druzdel M J, Van D G L C. Building Probabilistic Networks: "Where Do the Numbers Come From?" [J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(4): 481-486.

[7] Zhou Yun, Fenton N, Neil M. Bayesian Network Approach to Multinomial Parameter Learning Using Data and Expert Judgments [J]. International Journal of Approximate Reasoning, 2014, 55(5): 1252-1268.

[8] Pan Jialin, Yang Qiang. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[9] 张建军,王士同,王 骏. 迁移学习数据分类中的 ESVM 算法[J]. 计算机工程,2012,38(8):173-176.

[10] Shimodaira H. Improving Predictive Inference Under Covariate Shift by Weighting the Log-likelihood Function[J]. Journal of Statistical Planning and Inference, 2000, 90(2): 227-244.

[11] 张连文,郭海鹏. 贝叶斯网引论[M]. 北京:科学出版社,2006.

[12] Bickel S, Brückner M, Scheffer T. Discriminative Learning for Differing Training and Test Distributions[C]//Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM Press, 2007: 81-88.

[13] Xia Rui, Hu Xuelei, Lu Jianfeng, et al. Instance Selection and Instance Weighting for Cross-domain Sentiment Classification via PU Learning[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: IJCAI Inc., 2013: 2176-2182.

[14] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[15] Huang Jiayuan, Gretton A, Borgwardt K M, el al. Correcting Sample Selection Bias by Unlabeled Data[C]// Proceedings of NIPS' 06. Vancouver, Canada: [s. n.], 2006: 601-608.

编辑 刘 冰