基于改进 TF-IDF 算法的牛疾病智能诊断系统

杜永兴 牛丽静 秦 岭 李宝山
(内蒙古科技大学信息工程学院 内蒙古 包头 014010)

摘 要 传统的 TF-IDF(Term Frequency & Inverse Documentation Frequency)算法提取的关键词不能合理地代表某疾病的症状,降低智能诊断系统的性能。对此,提出一种改进的 TF-IDF 算法,并将其应用在牛疾病诊断系统中。系统将用户描述的文本内容转换成向量的形式,用 TF-IDF 算法提取关键症状词,利用余弦定理和可信度计算给出可靠的疾病推荐和治疗方案。实验结果表明,该算法在疾病诊断中准确率和可信度两方面都具有更好的效果。与传统 TF-IDF 算法相比,平均可信度提高约 4%。

关键词 智能诊断 TF-IDF 余弦相似度 VSM

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.02.009

CATTLE DISEASE INTELLIGENT DIAGNOSIS SYSTEM BASED ON IMPROVED TF-IDF ALGORITHM

Du Yongxing Niu Lijing Qin Ling Li Baoshan
(*School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, Inner Mongolia, China*)

Abstract The of the keywords extracted by the traditional TF-IDF (Term Frequency & Inverse Documentation Frequency) algorithm can not reasonably represent the symptoms of disease, thus reducing the performance of intelligent diagnostic systems. In response to this situation, an improved TF-IDF algorithm is proposed and applied in the cattle disease diagnosis system. The system converted the text content described by the user into a vector form, extracted the key symptom words by TF-IDF algorithm, and used the cosine theorem and credibility calculation to give a reliable disease recommendations and treatment plans. The experimental results show that the algorithm has better effects in both disease accuracy and credibility. The average credibility is improved by about 4% compared with the traditional TF-IDF algorithm.

Keywords Intelligent diagnosis TF-IDF Cosine similarity VSM

0 引 言

计算文本相似度是研究疾病智能诊断的一种重要的方法。目前 VSM 空间向量模型和 TF-IDF 方法提取关键词技术广泛应用在人病智能导医系统中。林予松等[1]采取了用户关注度来计算症状的权重,设计了人工智能导医系统。徐奕枫等[2]提出基于疾病类间分布的症状权重改进算法,改善了传统 TF-IDF 算法提取疾病的效果,取得了不错的成绩。Teshnehlab 等[3]首先通过主成分分析减少特征,然后使用基于深度神经网络算法对结肠癌分类,其分类准确度为 0.6。Cheng[4]通过物联网和人工智能自动化设计一个可以及时解决动物园里的动物身体出现异常情况的系统,帮助动物管理员更系统地管理照顾动物。以上系统的设计都有着显著的成果,但它们需要用户在系统中选择相应的症状,不能实现对用户所输入的文本内容进行理解。在设计中实现理解用户输入内容的复杂度远高于直接

收稿日期:2019 - 08 - 14。国家自然科学基金项目(61661044);内蒙古科技大学创新基金项目-优秀青年科学基金项目(2017YQL10);内蒙古自治区高等学校青年科技英才计划项目(NJYT-19-A15)。**杜永兴**,教授,主研领域:电磁场与微波技术,计算机仿真,大数据分析。**牛丽静**,硕士。**秦岭**,副教授。**李宝山**,教授。

选择症状。在对用户描述进行关键词提取时,用传统的 TF-IDF 算法在疾病的关键词提取中并未考虑到提取的权重比较高的关键词是否能合理地表示某种疾病。

针对上述问题,本文提出改进的 TF-IDF 算法,并将该算法应用在牛的疾病诊断系统中。首先用已有的方法对用户的输入的文本内容进行分词、提取关键症状词。然后采用向量空间模型 VSM 将提取的关键词用向量的形式表示,用余弦定理计算用户输入的关键词向量和已有的疾病关键词向量的值作为疾病的相似度。最后进行可信度的计算,推断出牛所患的疾病。应用该算法提取的关键症状词可以比较合理地表示疾病的症状,提高了系统的性能,使得该系统有效地实现了对牛所患疾病及时的诊断和治疗,对牧户在畜牧业的管理上也起到一定的指导和决策作用。

1　数据来源

应用 Python 框架和手工录入方式获取了 451 种关于牛的疾病,采用 jieba 分词[5]和手工整理的方式对病因、症状、诊断、治疗和预防等属性拆分,并将其对应的症状进行规范化处理,构造关键症状词语料库。

2　方法设计

2.1　空间向量模型及相似度计算

目前常用空间向量模型的方法来衡量两个文本之间的相似度[6]。向量空间模型(VSM)是把输入的文本和已有的文本都转换成向量的形式进行计算,提高了文本内容的计算性和可操作性,同时该模型也是目前应用最为成熟和广泛的模型之一[7-8]。

假设某用户描述用 D(Document)表示,首先运用自然语言处理已有的技术对用户的描述进行分词、去停用词、计算权重、提取关键症状词。特征项一般由症状关键词组构成,指在文档中能反映用户描述的基本语言单位,用 T(Term)表示。用户描述和关键症状特征可以使用集合表示为 $D(T_1,T_2,\cdots,T_n)$,其中 T_k 是关键症状特征词($1\leqslant k\leqslant n$)。生成向量空间模型的流程如图 1 所示。

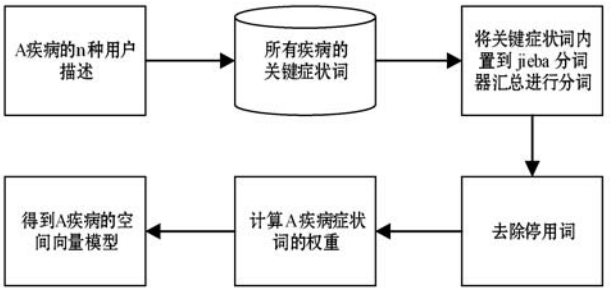

图 1　生成向量空间模型的流程

利用空间向量模型将文本内容转换成向量可以这样表示:对用户输入的文本 m 中的每个词,用 $W_{i,m}$ 表示 m 中第 i 个词的权重,$\boldsymbol{m}=(W_{1,m},W_{2,m},\cdots,W_{t,m})$ 表示用户输入文本 m 的词权重向量;同理,用 $W_{i,n}$ 表示已有文本 n 中第 i 个词的权重,用 $\boldsymbol{n}=(W_{1,n},W_{2,n},\cdots,W_{t,n})$ 表示已有文本 n 的词权重向量,然后通过余弦定理计算 $\boldsymbol{m}$ 和 $\boldsymbol{n}$ 之间的相似度值作为两个文本之间的相似度[9-11]。其中症状权重 W 是根据 TF-IDF 原理计算出来。在本系统中 $\boldsymbol{m}$ 表示带匹配的疾病,$\boldsymbol{n}$ 表示用户输入的描述。相似度计算如下:

$$\sin(\boldsymbol{m},\boldsymbol{n})=\cos\theta=\frac{\boldsymbol{m}\cdot\boldsymbol{n}}{\|\boldsymbol{m}\|\cdot\|\boldsymbol{n}\|}=\frac{\sum_{i=1}^{t}W_{i,m}\times W_{i,n}}{\sqrt{\sum_{i=1}^{t}W_{i,m}^2}\times\sqrt{\sum_{i=1}^{t}W_{i,n}^2}}\tag{1}$$

2.2　TF-IDF 算法分析及改进

在利用 VSM 计算两种疾病的相似性度时,最重要的步骤是用 TF-IDF 算法计算关键症状词的权重,提取关键词[12-13]。TF-IDF 算法的原理是 TF × IDF,其中 TF 表示某个症状词在文档出现的频率,计算中发现像“的”“了”等这些不重要的停用词出现的次数比较高。为避免这种问题,引入逆文档频率 IDF。包含当前词的文档个数越多,IDF 的值越小,说明该词越不重要。其主要思想是如果某个特征项在一个文本中出现频率很高,且在其他文本中出现很少,说明此特征项具有很好的类别区分能力,应该给予较高的权重[14]。TF 计算如下:

$$TF=\frac{C_{in}}{M_n}\tag{2}$$

式中:C_{in} 表示疾病特征词 i 在 n 种描述中出现的次数;M_n 表示 n 种描述中总症状词数。IDF 计算如下:

$$IDF=\lg\left(\frac{N}{n+1}\right)\tag{3}$$

式中:将每个描述看成一个文档,N 为文档总数;n 为包含某项症状词的文档总数。TF-IDF 计算公式如下:

$$TF\text{-}IDF=TF\times IDF\tag{4}$$

将传统的 TF-IDF 算法应用在提取牛疾病的关键症状词时,发现“带有”“比较”“基本”“而”等词计算出来的权重很高。显然这些词作为疾病的关键症状是不合理的。分析原因如下:在计算某一种疾病的关键症状时,将同一种疾病的不同种医案描述作为不同的文档来计算关键症状。此时的用户描述除了停用词出现的次数比较多之外,剩下的是症状描述,比如“精神倦息”“不反刍”等症状词,虽然很重要但是由于在每个用户描述中几乎都有出现,根据 TF-IDF 原理就把此类经常出现的症状词当成停用词处理了。针对上述问题,本文提出基于改进的 TF-IDF 算法,可以有效解决这个问题。改进的 TF-IDF 计算式如下:

$$TF\text{-}IDF = TF \times IDF \times W_{ij}$$

$$W_{ij} = \begin{cases} 10 & \text{该词此在关键症状词典中} \\ 1 & \text{该词不在关键症状词典中} \end{cases} \quad (5)$$

式中:W_{ij}代表第 j 种疾病的 i 个症状。首先通过传统的算法算出关键词的权重,然后将提取的关键词和牛疾病症状词典进行匹配。如果该词在症状词典中,则将该词相应的权重乘以 10;如果该词不在症状词典中,保留其原始的权重不变。最后将关键词的权重重新排序,选择权重较高的前 20 个关键词作为疾病的关键症状词。使每种疾病的关键症状权重更具有代表性,实现了相同症状在不同疾病占有不同的权重,更适用于疾病诊断。

2.3 可信度计算

单纯将相似度作为最后的结果返回给牧民是不够准确的。把可能度和相似度相结合作为疾病可信度计算结果,然后将可信度的结果按照从高到低的次序返回给牧民,增强结果的可靠性。可能度的计算使用的是不确定的推理,当用户输入描述时,将相应的症状权重相加。可能度的计算如下:

$$knd_j = W_{1j}x_1 + W_{2j}x_2 + \cdots + W_{ij}x_n \quad (6)$$

$$x_i = \begin{cases} 1 & \text{牧民选择了第 } i \text{ 个症状} \\ 0 & \text{牧民选择了第 } i \text{ 个症状} \end{cases}$$

式中:knd_j 代表患某种疾病的可能度。将选中的疾病索引到对应的权重进行加权求和,最后进行可信度的计算如下:

$$CF = \alpha knd_j + \beta \mathrm{sim}(m,n)$$
$$\alpha + \beta = 1 \quad (7)$$

式中:α 取 0.2,β 取 0.8 进行最后的可信度计算。

2.4 系统设计流程

牛的疾病诊断系统主要运用智能化方式辅助兽医诊断。牧民在使用此系统时,输入相应症状的文本内容,系统首先会对输入的文本内容进行理解,然后计算出输入内容与系统内所有疾病的相似度,最后计算可信度。将查询结果按照可信度从大到小的返回给牧民,并给出相应的诊疗方案。牛疾病智能诊疗系统主要包括自然语言处理、疾病匹配处理、疾病可信度计算三个部分。具体系统设计流程如图 2 所示。

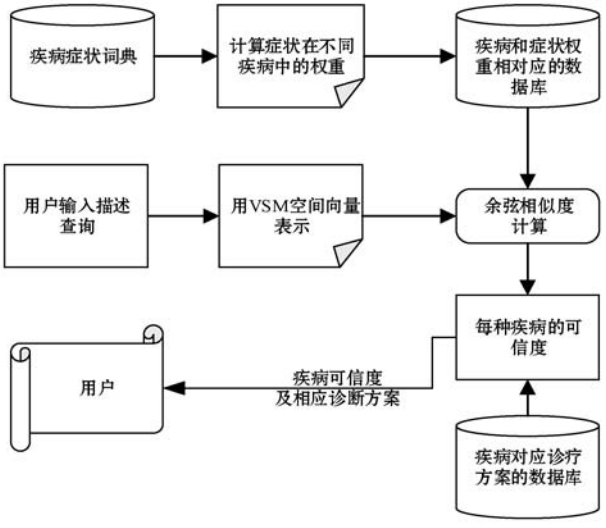

图 2 牛的疾病诊断流程

3 实 验

3.1 实验环境和实验数据

实验的运行环境为 Windows XP 操作系统,CPU 主频 3.7 GHz,内存 16 GB,数据库 Microsoft MySQL 2018,开发工具为 PyCharm 2018,编程语言为 Python。实验数据采取随机抽取 30 头牛的病历样本进行实验验证,且这些疾病兽医已经给出正确的诊断结果。

3.2 评价指标

为验证改进后的算法在牛疾病诊疗系统中的准确率和可信度,本文采用基于传统的空间向量模型的相似度算法和本文提出改进的 TF-IDF 算法进行对比实验。实验采用随机抽取 30 头牛的病历样本进行实验验证,采用 S@n(success at n) 方法进行结果评测[15],其表示正确疾病推荐结果在前 n 个推荐结果中所占比重。

3.3 结果分析

将实验数据采用 S@n 方法进行结果评测,两种算法的对比结果如表 1 所示。可以看出,当 n 取 1、2、3 时,本文算法的正确率明显高于传统算法。

表 1　算法准确率对比

权重计算方法	S@1	S@2	S@3
传统算法	0.70	0.73	0.80
本文算法	0.76	0.80	0.86

通过上述计算相似度及可信度的方法,使用两种算法对同一实验数据计算出相似度和可信度的对比如图 3 和图 4 所示。可以看出,改进算法相似度和可信度较传统算法都有提高,其中可信度平均提高约 4 个百分点,说明本文算法在牛疾病诊断中更具有可行性。

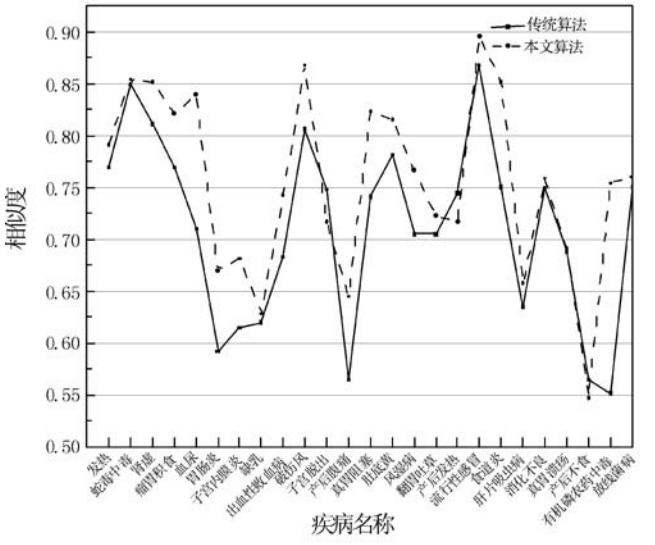

图 3　相似度结果对比图

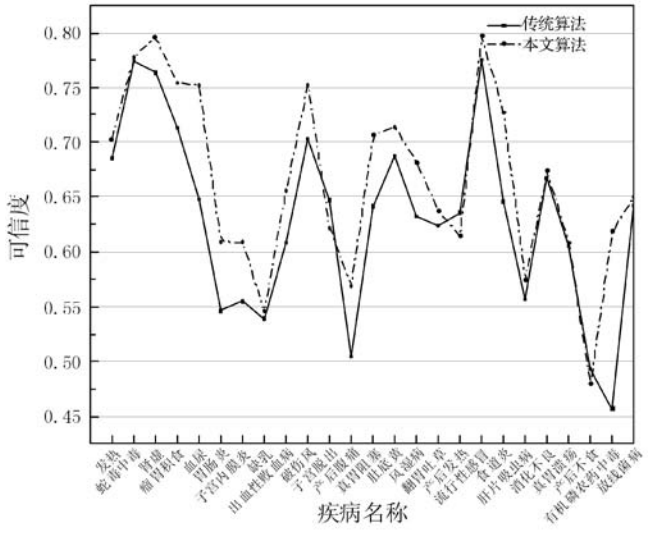

图 4　可信度结果对比图

4　结　语

针对传统 TF-IDF 算法提取关键词不能合理地描述疾病的关键症状,本文提出一种改进的 TF-IDF 算法提取关键症状词并设计了牛疾病智能诊断系统。通过实验对比验证了该算法的有效性。该方法的不足是在提取关键症状词时依赖疾病症状词的语料库。下一步研究将重点考虑在不依赖疾病症状词语料库的基础上更加智能地实现疾病诊断。

参　考　文　献

[1] 林予松,梁璐,崔勇,等. 基于 VSM 权重改进算法的智能导医系统[J]. 计算机应用与软件,2015,32(9):81 - 83.

[2] 徐奕枫,刘利军,黄青松,等. 智能导医系统中 TF-IDF 权重改进算法研究[J]. 计算机工程与应用,2017,53(4):238 - 243.

[3] Teshnehlab M, Jafarpisheh N. Cancers classification based on deep neural networks and emotional learning approach [J]. IET Systems Biology,2018,12(6):258 - 263.

[4] Cheng Y H. A development architecture for the intelligent animal care and management system based on the internet of things and artificial intelligence[C]//2019 1st International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE,2019:78 - 81.

[5] Su B H, Tseng S P, Lin Y S. Health care spoken dialogue system for diagnostic reasoning and medical product recommendation[C]//2018 International Conference on Orange Technologies (ICOT). IEEE,2018:1 - 4.

[6] Li W H, Deng L F, Yang L, et al. ZigZag: Supporting similarity queries on vector space models[C]//Proceedings of the 2018 International Conference on Management of Data. ACM,2018: 873 - 888.

[7] Jayashree R, Niveditha N. Natural language processing based question answering using vector space model[C]//Proceedings of Sixth International Conference on Soft Computing for Problem Solving. Springer,2016:368 - 375.

[8] 李琳,李辉. 一种基于概念向量空间的文本相似度计算方法[J]. 数据分析与知识发现,2018,2(5):48 - 58.

[9] Gomaa W, Fahmy A. A survey of text similarity approaches [J]. International Journal of Computer Applications,2013,13(6):13 - 68.

[10] 黄文彬,车尚锟. 计算文本相似度的方法体系与应用分析[J]. 情报理论与实践,2019,42(11):128 - 134.

[11] Xu L H, Sun S T, Wang Q. Text similarity algorithm based on semantic vector space model[C]//2016 IEEE/ACIS 15th International Conference on Computer and Information Science. IEEE,2016:1 - 4.

[12] Arroyo-Fernández I, Méndez-Cruz C F, Sierra G, et al. Unsupervised sentence representations as word information series: Revisiting TF-IDF[J]. Computer Speech & Language, 2019,56: 107 - 129.

(下转第 57 页)

surf 函数、mesh 函数三种可视化方法显示的孔的轴线直线度误差的评定结果以及圆周轮廓、实际轴线、理想轴线及轴线直线度误差的包容面。图中的 x 和 y 坐标并不是真实的坐标,而是微小变动局部放大后轮廓的坐标,轮廓局部放大倍数为 1 000,其目的是使轮廓和轴线直线度误差的包容面更清晰,另外,没有对图中的奇异点进行处理。可以看出,三种可视化方法均能较清晰地显示被测孔的实际轴线、理想直线和包容面。但需要注意的是,轴线直线度误差较小时,图 6 和图 7 的方式将不能清晰地显示实际轴线。

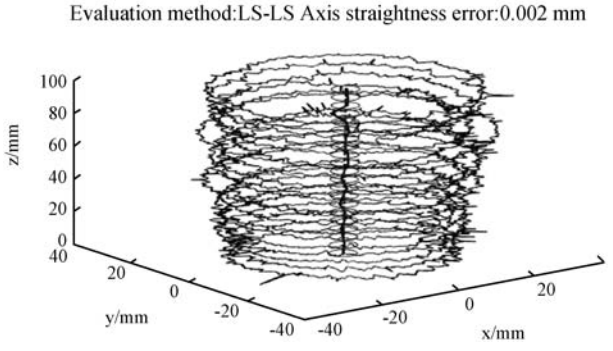

图 5　用椭圆轮廓显示轴线直线度误差的包容面

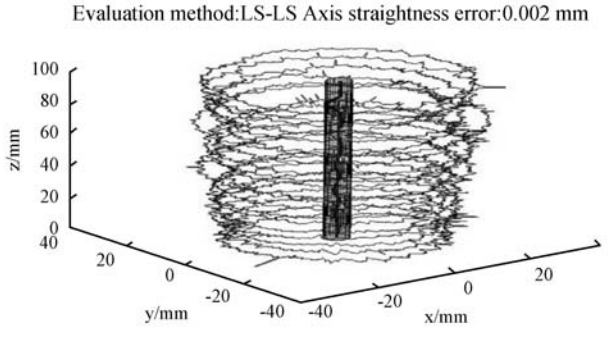

图 6　用 surf 函数显示轴线直线度误差的包容面

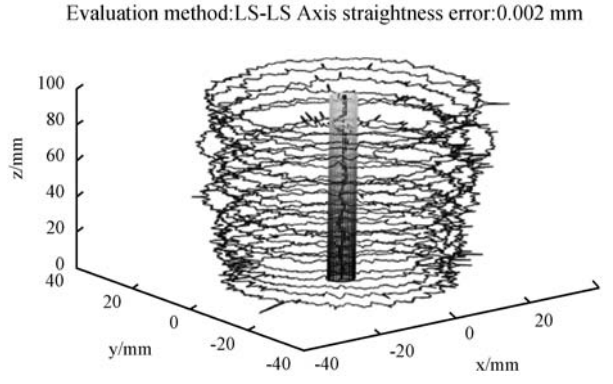

图 7　用 mesh 函数显示轴线直线度误差的包容面

4　结　语

本文以圆周法提取方案对圆柱体的圆周轮廓进行测量,获取圆周轮廓采样点的微小变动量,建立了基于最小二乘圆的圆心坐标的圆柱体轴线直线度误差的评定模型。由圆柱体轴线直线度误差评定模型确定理想轴线参数,建立了圆柱体实际轴线包容面轮廓坐标的生成模型,并提出了圆柱体实际轴线包容面的三种可视化方法。基于所建立的模型,编写了圆柱体最小二乘法轴线直线度误差的评定及其可视化程序,对孔的轴线直线度误差进行了评定及其结果显示。本文提出的圆柱体实际轴线包容面的可视化方法也可用于圆柱度误差评定中的图形显示。

参　考　文　献

[1] 张双双,杨洪涛,刘齐更. 基于 VC 的 4 种圆度误差评定方法的比较[J]. 工具技术,2013,47(11):61 – 64.

[2] 于大国,宁磊,孟晓华. 基于最小二乘法深孔轴线直线度误差评定[J]. 组合机床与自动化加工技术,2014(1):39 – 41,45.

[3] 张新宝,谢江平. 空间直线度误差评定的逼近最小包容圆柱法[J]. 华中科技大学学报(自然科学版),2011,39(12):6 – 9.

[4] Luo J, Wang Q. A method for axis straightness error evaluation based on improved artificial bee colony algorithm[J]. The International Journal of Advanced Manufacturing Technology,2014,71:1501 – 1509.

[5] 张珂,张玮,阎卫增,等. 圆度误差的神经网络评定及测量不确定度研究[J]. 机械科学与技术,2019,38(3):428 – 432.

[6] 史栩屹,李明,韦庆钥. 二次插值鲸鱼优化算法在圆柱度误差评定中的应用[J]. 计量与测试技术,2019,46(2):58 – 60.

[7] 赵艺兵,温秀兰,许有熊. 基于坐标测量机和拟粒子群进化算法的圆柱度误差检测与评定[J]. 中国机械工程,2015,26(18):2432 – 2436.

[8] Zhao Z X, Li B, Zhang G Q, et al. Study on the evaluation of cylinder's global sizes[J]. Precision Engineering,2017,49: 189 – 199.

[9] Zhao Z X, Li B, Zhang G Q, et al. Influence of eccentricity and tilt of cylindrical part's axis on the measurement results of its diameters[J]. Measurement,2019,138:232 – 239.

(上接第 53 页)

[13] Yahav I, Shehory O, Schwartz D. Comments mining with TF-IDF: the inherent bias and its removal[J]. IEEE Transactions on Knowledge and Data Engineering,2019,31(3):437 – 450.

[14] Dai W S. Improvement and implementation of feature weighting algorithm TF-IDF in text classification[C]//2018 International Conference on Network, Communication, Computer Engineering (NCCE),2018.

[15] 邓红莉,杨韬,邵晨曦. 一种对仿真可信度评估的智能专家系统[J]. 计算机仿真,2011,28(8):90 – 93.