

基于领域本体的主观题自动评阅算法的研究

兰富菊,赵志弘,韩永国

(西南科技大学 计算机科学与技术学院,四川 绵阳 621010)

摘要:针对 VSM 不能揭示文档中特征词间的潜在语义关系,相似度计算准确性较低的问题,结合本体模型的结构特点,从语义重合度、语义距离以及本体结构等因素综合考虑概念间的相似度计算,提出了一种基于领域本体的文档向量空间模型。该模型通过构建概念间的语义相似度矩阵对特征词权值进行调整,建立包含语义关系的标准(学生)答案的向量空间模型,并用“VSM 模型+余弦值”算法评估学生答案和标准答案的相似度。实验表明,与传统方法相比,该方法提高了评测效果及准确率。

关键词:领域本体;相似度矩阵;权值;语义关系;向量空间模型

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2014)06-0166-04

doi:10.3969/j.issn.1673-629X.2014.06.041

Research on Subjective Machine Marking Algorithm Based on Domain Ontology

LAN Fu-ju, ZHAO Zhi-hong, HAN Yong-guo

(College of Computer Science and Technology, Southwest University of Science and
Technology, Mianyang 621010, China)

Abstract: In view of the problems that VSM couldn't reveal the latent semantic relations between the key words in a document, and have a low accuracy in document similarity calculation, combined the structural feature of ontology model, considering the concept similarity from the semantic contact ratio, semantic distance and the ontology structure, a document vector space model based on domain ontology is proposed. The model adjusts the weight of feature words by building concept semantic similarity matrix, constructing the standard answer and the student answer VSM contains semantic relation, and making use of the "VSM+cosine" algorithm to assess similarity of the student answer and the standard answer. Compared with the traditional method, experiments show that this method improves the scoring results and accuracy.

Key words: domain ontology; similarity matrices; weighting; semantic relation; VSM

0 引言

近年来随着计算机辅助教学的发展,在线考试系统已广泛应用到各个领域,当前现有国内主观题自动评阅算法主要有:引入单向贴进度算法、谓词演算方法、基于知网的相似度计算方法和 VSM 模型,其中应用最广泛的就是 VSM 模型^[1]。它是一种代数模型,其基本思想是:用向量来表示文本,因此文档间的运算就转换成向量间的数学运算,使得模型具备可计算性。但该方法认为各特征词之间是独立的,忽略了基于语义层次的概念之间的语义关系,如同义词与近义词等

语言现象,导致在文档相似度计算时准确性不高,随着语义网和本体研究的深入,为文本表示提供了新的可能。文献[2]将本体定义为特定领域客观存在的概念与概念之间关系的描述;文献[3]将本体定义为共享概念模型的明确的规范说明。对此有关专家提出了很多方法,Jing 等^[4]采用基于本体的互信息测度来计算文档特征项之间的相似度,利用知网得到两个特征项之间的距离并计算其权重,从而完成文本的聚类;Maedche A 等利用 String Matching 进行语义相似度计算^[5];Kwon Ju-Hum 等从语义距离方面对相似度进行研究^[6];Ehrig M 等提出基于规则的计算方法^[7];刘群

收稿日期:2013-08-20

修回日期:2013-11-25

网络出版时间:2014-02-24

基金项目:四川省工程实验室开放基金项目(11zxwk06)

作者简介:兰富菊(1987-),女,硕士研究生,研究方向为教育软件开发、自然语言处理;韩永国,教授,博士,研究方向为知识工程、教育网格技术。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0915.035.html>

等利用知网通过计算概念间义原间的距离确定概念相似度^[8];朱礼军等通过计算概念间的距离^[9],提出基于RDFS的领域知识本体中的概念相似度的计算方法;朱会峰等人^[10]引入知网,并定义了关键概念集,计算知网中的概念节点及概念间的语义关系,利用文本的关键概念集和概念特征向量计算文本相似度。针对以上前人的研究,文中把本体概念间的相似度计算综合考虑其因素,对VSM模型进行改进,从而构建一种具有语义关系的文档的向量空间模型。

1 基于领域本体的文档向量空间模型构建

相关研究表明^[11-12],本体层次树结构中影响概念节点 X, Y 的相似度的主要是两个节点间的语义距离(最短路径距离),距离越大,语义相似度就越小;两个概念包含相同的上位概念节点在总的节点中的比例,比例越大,相似度就越大;同时语义相似度也受两概念所处的层次深度的影响,因为在本体层次树中,概念所处的层次越深,分类越精细。

1.1 概念间语义相似度计算

定义1:语义距离^[13]。设 X, Y 为本体层次树中任意两个节点,则这两者的语义距离是指在概念节点最短路径的长度,记为: $\text{Distance}(X, Y)$ 。从语言学角度认为:两个词语的语义距离越大,其相似度越低;反之,语义距离越小,其相似度越大。研究表明:以处于层次树中离根较远的概念间相似度要比离根近的概念间相似度大,因为概念节点与根节点距离越远,其概念所表达的意义越具体,由此得出概念 X 和 Y 语义距离的相似度表达方式如下:

$$\text{Sim}(\text{Distance}) = \frac{\partial}{\text{Distance}(X, Y) + \partial} \quad (1)$$

定义2:语义重合度。语义重合度是指本体内部概念之间包含相同上位概念的个数,在本体层次树中,通常通过计算概念间公共父节点的个数来衡量语义重合度。设本体层次树中的根节点为 R, X, Y 是树中的任意两个概念节点, $\text{Count}(S_x)$ 是从 X 出发,向上直到根节点 R 所经过的所有概念节点集合,同理 $\text{Count}(S_y)$ 是从 Y 节点就出发,向上直到根节点 R 所经过的所有概念节点集合,则语义重合度 $\text{Contactratio}(X, Y)$ 可以用 S_x 与 S_y 的交集的节点数 $\text{Count}(S_x, S_y)$ 以及 $\text{Count}(S_x)$ 和 $\text{Count}(S_y)$ 的个数来计算,从而得出概念 X 和 Y 在语义重合度方面的语义相似度计算:

$$\text{Sim}(\text{Contactratio}) = \frac{\text{Count}(S_x \cap S_y) + \varepsilon}{\text{Max}(\text{Count}(S_x), \text{Count}(S_y)) + \varepsilon} \quad (2)$$

定义3:概念的宽度。假设 C 为层次树中任意一个节点,概念 C 的宽度是指本体层次树中,概念 C 所

有子节点的个数,记为 $\text{Width}(C)$ 。概念的宽度越大,说明对该概念进行的细化程度越具体,此概念的地位也越重要,式(3)定义了 X 和 Y 概念宽度的语义相似度计算表达式:

$$\text{Sim}(\text{Width}) = \frac{1}{\text{Width}(X)} \times \frac{1}{\text{Width}(Y)} \quad (3)$$

定义4:树的深度。假设节点 C 为本体层次树中任意一节点,其概念的深度记为 $\text{Depth}(C)$,在层次树中,根节点 Root 的深度为1,其他非根节点 C 的深度为 $\text{Depth}(C) = \text{Depth}(\text{Parent}(C)) + 1$,其中 $\text{Depth}(\text{Tree}^*)$ 为树中所有概念节点的最大深度,即 $\text{Depth}(\text{Tree}^*) = \max(\text{Depth}(C))$ 。因为在本体图中,每一层都是对上一层概念的细化,由此可见,在语义距离相同的前提下,两个节点的深度和越大,概念之间的相似度越大,考虑概念深度的影响,得出概念 X 和 Y 概念深度语义相似度计算表达式如下:

$$\text{Sim}(\text{Depth}) = \frac{\text{Depth}(X) + \text{Depth}(Y)}{\text{Depth}(X^*) + \text{Depth}(Y^*)} \quad (4)$$

定义5:结构影响。在本体层次树中各个概念的宽度、深度对其结构有很大影响,因为在同一领域内,由于对概念的分解与侧重不同,使得概念间的相似度层次树中有很大差异,因此把概念的宽度、概念的深度归结为本体结构对语义相似度的影响。综合式(3)和式(4),得式(5):

$$\text{Sim}(\text{Structure}) = \sqrt[\alpha]{\text{Sim}(\text{Width})} \times \sqrt[\beta]{\text{Sim}(\text{Depth})} \quad (5)$$

其中, $\alpha + \beta = 1$,其各项取值大小视各因素对结构影响大小而定。

1.2 概念间语义关系计算

综合上面提出的各方面的因素,可以得到概念的语义相似度,如式(6)所示:

$$\text{Sim}(X, Y) = \sqrt[a]{\text{Sim}(\text{Contactratio})} \times \sqrt[b]{\text{Sim}(\text{Contactratio})} \times \sqrt[c]{\text{Sim}(\text{Contactratio})} \quad (6)$$

其中, $a + b + c = 1$,其各项取值大小视各因素对语义相似度影响大小而定。利用式(6)计算出领域本体中所有概念之间的语义关系,形成具有语义关联的相似度矩阵 S , $\text{Sim}(X, Y)$ 表示概念 X, Y 之间的语义关系, S 为对称矩阵。

1.3 构建基于领域本体的文档向量空间模型

基本思想是:首先抽取出自文档中的特征词集合,将领域本体概念集中没有出现在特征词集合中的概念增加到特征词集合中作为文档的特征词,形成一个扩展的文档特征词集合;对于扩展后的文档特征词集合中特征词的权重计算,采用如下规则:若特征词出现在文档中且不属于领域本体中的概念,采用tf-idf方法^[14]计算特征词权重,若特征词属于本体中的概念且

出现在文档中,采用 tf-idf 方法计算特征词权重,并将得到的权重作为初始值,结合本体概念的相似度矩阵 S ,计算出所有的本体中概念词的权重;最终得到基于领域本体的扩展向量空间模型。设文档 d 采用自然语言处理技术得到文档的特征词集合 $T = \{t_1, t_2, \dots, t_n\}$, 本体库中的概念集合 $C = \{c_1, c_2, \dots, c_m\}$, 则集合 T 与 C 的交集构成的集合 $T' = T \cap C$, 令 T' 为文档 d 扩展后的特征词集合,若扩展集合中的特征词 $t_i \in C$, 则称 t_i 为领域类特征词。

基于领域本体扩展的文档特征词集合 T' 中特征词 t_i 的权重 W_i 采用两次计算来得到,首先是基于统计的方法计算出特征词 t_i 的初始权重,然后利用领域本体的概念间语义关系矩阵对领域类特征词 t_i 的权重进行优化,得到包含语义关系的特征词权重,最终建立文档的向量空间模型。具体实现分为两个步骤。

a) 特征词权重的初始计算。对于集合 T' 中特征词 t_i 的权重 W_i 根据如下函数得到

$$W_i = f(t_i) = \text{tf}_i \times \text{idf}_i \quad (7)$$

其中, tf_i 为特征词 t_i 在文档 d 中出现的频率,反映了特征项对文档的重要程度; idf_i 为反文档频率,是特征项在文档集合中分布情况的量化。

b) 文档中领域类特征词权重的调整。经过步骤 a 的特征词权重计算后,建立文档 VSM, 其中领域类特征词 t_i 的权重完全是采用统计的方法,而忽略这类特征词之间的语义关系,简单依靠统计方法是不合理的,因此需要对领域类特征词权重进行调整,建立特征词之间的语义关系。文中采用如下策略进行调整:

根据概念相似度矩阵 S 建立对应的文档领域类特征词向量 $((t_1, w_1), (t_2, w_2), \dots, (t_m, w_m))$, 其中特征词的权值向量 $W = (w_1, w_2, \dots, w_m)$, 令调整后的领域类特征词权值向量为 W' , 则

$$W' = S \times W = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & \cdots & \cdots & s_{mm} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \quad (8)$$

由式(8)可知,对于每一个属于领域类特征词,它的权重与它关联的其他所有领域类特征词共同决定,从而建立特征词之间的语义关联;传统的文档向量空间模型中特征词不管是否属于领域概念,都是简单依据统计信息给出权重,说明领域类特征词和非领域类特征词对文档的贡献度是一样的,然而对于专业领域的文档而言,前者对于文档来说更为重要,因此,需要提高领域类特征词的权重,通过式(8)调整后的权值由于没有采取归一化处理,这一过程使得 $W > W'$, 增大领域类特征词的权重,从而提高了领域类特征词对于文档的贡献程度,构建的文档更具有现实意义。

1.4 算法的描述

在传统的 VSM 中,特征词对文档的重要程度不能简单地由特征词在文档中的统计信息确定,文档中特征词之间存在语义关系。一个特征词对文档的重要性还取决于与其有语义关系的其他特征词的相关信息,同时对于专业领域的文档而言,属于领域中特征词对于其他非领域中的特征词在区分文档方面具有更高的区分度,应该赋予更高的权重。

据此根据本体模型的结构特点,文档相似度的计算包含两个过程:首先是计算出领域本体中概念间的相似度,得到语义相似度矩阵 S ;然后对学生(参考)答案进行预处理,进行特征词的抽取,将领域本体概念集中没有出现在特征词集合中的概念增加到特征词集合中,形成一个扩展的文档特征词集合,先利用公式(7)计算扩展的特征词集合初始权重 W ,再结合本体概念的相似度矩阵 S ,对初始的权值 W 进行优化,分别得到标准答案和学生答案的空间向量 D_1, D_2 ,再用余弦值计算两者的相似性,具体的流程图如图 1 所示。

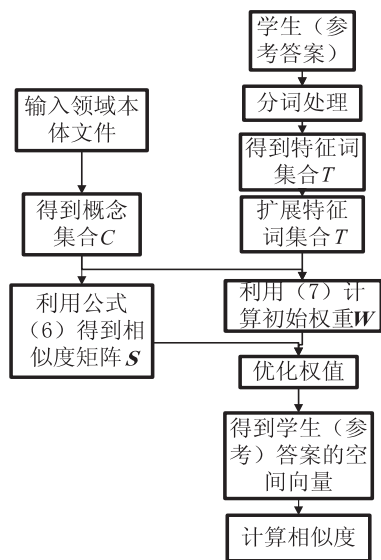


图 1 算法流程图

2 实验结果及分析

2.1 实验准备

以操作系统课程的测试试卷作为实验素材,选取某该测试的 200 份试卷,将其主观题部分录入计算机,同时录入人工阅卷结果。在录入人工阅卷结果时,将每道试题由两位教师协调给出得分,以提高人工评阅的准确性。在每道试题的 200 份答案中,随机选择其中 20 份(以该套试题中一题为例,满分为 20 分,系统给出的分数采用四舍五入)作为测试用例,其余 180 份作为该系统用的训练集,从准确性、速度两方面将该系统与对比系统(对比系统采用的是基于统计的方法)进行比较和分析。

2.2 实验结果及分析

为了便于对实验结果进行分析评价,将人工批改与两种自动批改的分值结果用折线图表示,结果对比如图2所示。

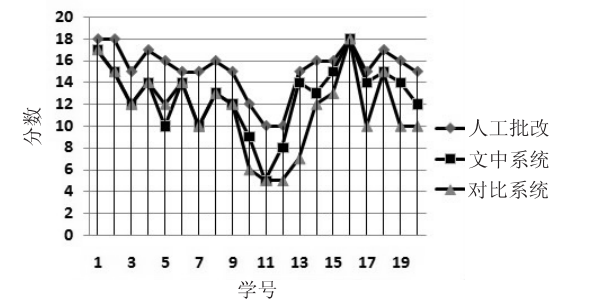


图2 人工批改和两种自动批改对照折线图

(1) 准确性评价。

准确性是评价文中所述自动批改系统可操作性的 重要参数。文中采用如公式(9)、(10)所示计算方法, 综合评价整个系统的评阅准确率:

单个答案准确率(V_i) = 1 -
$$\frac{|\text{人工评分} - \text{机器评分}|}{\text{题目满分}} \tag{9}$$

综合准确率 =
$$\sum_{i=1}^N V_i / N \tag{10}$$

其中, V_i 代表单个答案的评阅准确率; N 表示试卷 总数,此处 N 为 20。

对实验的统计结果如表1所示:由公式(10)计算 得到文中系统综合准确率为 86.75%,对比系统的综 合准确率为 80.05%,实验表明:改进的 VSM 计算的相 似度和教师的评阅更为接近,更加精确地反映出文档 之间的相似度。

表1 准确率统计结果

准确率	≥90%	≥80%	≥70%	≥60%	<60%
文中系统	5	11	3	1	0
对比系统	3	7	6	3	0

此外对上述图表中所示的分数结果的差异情况进 行仔细研究分析,发现人工评阅和机器评阅在4处(考 生5、7、11、12)出现较大偏差,其中有5、7两处是因为 分词词典中缺少某些词汇的相关信息,导致系统无法 对学生答案的某句进行正确的分词处理,最后导致阅 卷结果出现偏差;另外两处11、12偏差主要是学生卷 面整洁,给教师的主观印象较好,给考生提高了分数, 导致了自动批改的分数较低,此处正好说明了主观题 自动批改系统的价值和实际意义。

(2) 速度比较。

实验过程中,对20份样例答案进行评阅,该系统 用时707 800 ms,对比系统用时67 970 ms,文中算法执 行速度比基于统计算法的对比系统慢,两者相差1个

数量级。这主要是因为该系统为提高准确率,在算法 中加入提高阅卷准确率的策略和方法,而对比系统主 要采用执行策略简单的算法。在机器阅卷算法中,只 有在高准确率的前提下如何去研究如何提高阅卷的速 度才有意义。

当然由于文中实验用的数据是小样本数据,而且 只针对操作系统这一门课程,学生在测试前都经过相 关的测试培训,答题大都比较符合规范。所以,只能说 该系统在特定的应用领域达到了一定的评测效果,具 有一定的实用性,在接下来的系统设计过程中继续加 大实验样本,使测试结果更加准确完善。

3 结束语

针对标准(学生)答案文档的信息量较少的特点, 文档间的相似度不能简单地以特征词的统计信息确 定,文中提出的基于领域本体的 VSM 模型不仅加强了 领域类特征词之间的语义关联,赋予了领域类特征词 更高的权重,使得构建的向量空间模型更加科学、合 理,更加精确地反映了两者文档的相似度,也能避免 人工批改中一些人为因素的不利影响,促进考试的公 平性、公正性。然而参考(学生)答案文档向量原始特征 的数量很大,要求计算机有较好的处理速度和存储空间, 因此必须对向量空间模型进行降维处理,可考虑采用 SVD 对文档矩阵进行分解将特征向量空间从高维 变换到低维空间。因此,如何对文档进行有效的降维 处理是下一步的研究方向。

参考文献:

[1] 许云,樊孝忠,张锋. 基于知网的语义相关度计算[J]. 北京理工大学学报,2005,25(5):411-414.

[2] Studer R, Benjamins V R, Fensel D. Knowledge engineering, principles and methods[J]. Data and Knowledge Engineering, 1998,25(1-2):161-197.

[3] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition,1993,5(2):199-220.

[4] Jing Liping, Zhou Lixin, Ng M K, et al. Ontology-based distance measure for text clustering[EB/OL]. 2011. <http://www.siam.org/meetings/sdm06/workproceed/TextMining/jing1.pdf>.

[5] Maedche A, Staab S. Measuring similarity between ontologies [C]//Proc of the 13th international conference on knowledge engineering and knowledge management. London, UK: Springer-Verlag,2002:251-263.

[6] Kwon Ju-Hum, Choi O-Hoon, Moon Chang-Joo, et al. Deriving similarity for Semantic Web using similarity graph[J]. Journal of Intelligent Information Systems,2006,26(2):149-166.

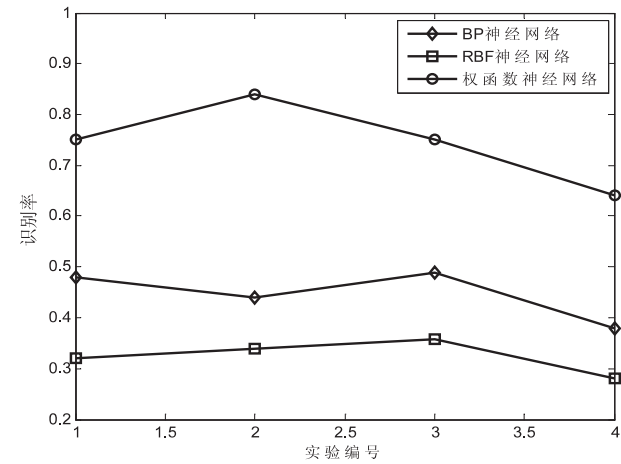


图3 不同网络的识别率

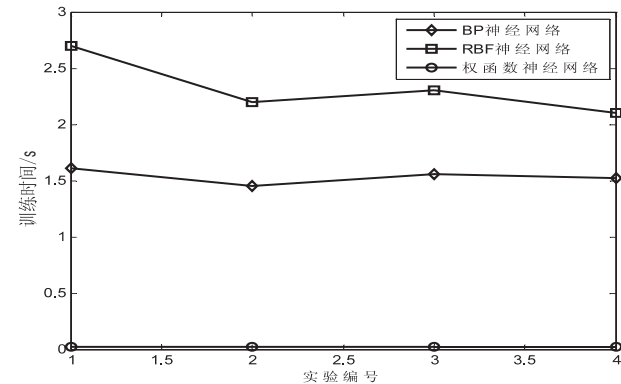


图4 不同网络的训练时间对比

4 结束语

样条权函数神经网络有着十分简单的算法结构,训练后的权值因为是输入样本的函数,所以可以直接反映输入样本的特征。文中通过理论和仿真实验说明样条权函数神经网络算法可以运用到指纹识别技术中。实验结果指出,样条权函数神经网络算法的识别率高于BP、RBF算法。更值得关注的是,在训练时间方面,样条权函数神经网络算法更是远远优于BP、RBF算法,在运算大数据方面,有着广阔的应用前景。

参考文献:

[1] Hagan M T, Demuth H B, Beale M H. 神经网络设计[M]. 北京:机械工业出版社,2003.

[2] 张代远. 神经网络新理论与方法[M]. 北京:清华大学出版社,2006.

[3] 张代远. 样条权函数神经网络的一种新型算法[J]. 系统工程与电子技术,2006,28(9):1434-1436.

[4] 张代远. 新型样条权函数神经网络的云计算研究[J]. 计算机技术与发展,2013,23(7):57-61.

[5] Mehtre B M. Fingerprint image analysis for automatic identification [J]. Machine Vision and Application,1993,6(2-3):124-139.

[6] 潘 滢,谢胜曙,张 群. 基于BP神经网络指纹识别的算法[J]. 邵阳学院学报(自然科学版),2007,4(1):54-57.

[7] Karungaru S, Fukuda K, Fukumi M, et al. Classification of fingerprint images into individual classes using neural networks [C]//Proc of 34th annual conference of industrial electronics. Orlando, FL:IEEE,2008:1857-1862.

[8] Xie Rui, Qi Jin. Continuous fingerprint image quality estimation based on neural network[C]//Proc of international symposium on intelligent signal processing and communication system. Chengdu:IEEE,2010:1-4.

[9] 王崇文. 自动指纹识别方法研究[D]. 重庆:重庆大学,2002.

[10] 祝 恩. 低质量指纹图像的特征提取与识别技术的研究[D]. 长沙:国防科学技术大学,2005.

[11] Abdi H, Williams L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4):433-459.

[12] Pearson K. On lines and planes of closest fit to systems of points in space[J]. Philosophical Magazine,1901,2(6):559-572.

[13] 杨利敏. 图像特征点定位算法研究及其应用[D]. 上海:上海交通大学,2008.

[14] 韩 鹏. 分子三次、分母二次有理样条权函数神经网络研究及应用[D]. 南京:南京邮电大学,2012.

(上接第169页)

[7] Ehrig M, Sure Y. Ontology mapping - an integrated approach [C]//Proc of European semantic web symposium. Berlin, Germany:Springer-Verlag,2004:76-91.

[8] 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.

[9] 朱礼军,陶 兰,刘 慧. 领域本体中的概念相似度计算[J]. 华南理工大学学报(自然科学版),2004,32(Sup):147-150.

[10] 朱会峰,左万利,赫枫龄,等. 一种基于本体的文本聚类方法[J]. 吉林大学学报(理学版),2010,48(2):277-283.

[11] Knappe R, Bulskov H, Andreassen T. Similarity graphs [C]//Proc of the 14th international symposium on methodologies for methodologies for intelligent systems. [s. l.]:[s. n.],2003:668-672.

[12] 徐德智,王怀民. 基于本体的概念间语义相似度计算方法研究[J]. 计算机工程与应用,2007,43(8):154-156.

[13] 李文杰,赵 岩. 基于本体结构的概念间语义相似度算法[J]. 计算机工程,2010,36(23):4-6.

[14] 李 鹏,陶 兰,王弼佐. 一种改进的本体语义相似度计算及其应用[J]. 计算机工程与设计,2007,28(1):227-229.

基于领域本体的主观题自动评阅算法的研究

作者：

[兰富菊](#)，[赵志弘](#)，[韩永国](#)，[LAN Fu-ju](#)，[ZHAO Zhi-hong](#)，[HAN Yong-guo](#)

作者单位：

[西南科技大学 计算机科学与技术学院, 四川 绵阳, 621010](#)

刊名：

[计算机技术与发展](#)

ISTIC

英文刊名：

[Computer Technology and Development](#)

年，卷(期)：

2014(6)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201406041.aspx