基于领域本体的文档向量空间模型构建

郝文宁，冯　波，陈　刚，靳大尉，赵水宁
（解放军理工大学 工程兵工程学院，南京 210007）

摘　要：针对 VSM 不能揭示隐藏在不同特征词后面的相同概念语义、反映文档中的潜在语义关系、在相似度计算中精度较低的问题，提出一种基于领域本体的文档向量空间模型 DOBVSM（domain ontology-based vector space model）。该模型把领域本体中的概念扩展为文档特征词，并通过概念间的语义关系对特征词权重进行调整，最终建立包含语义关系的文档 DOBVSM。通过实验分析表明：DOBVSM 计算的文档相似度值更加发散，与专家评价值最为接近，能够较好地反映文档之间的相似情况。

关键词：领域本体；向量空间模型；相似度；语义关系；基于领域本体向量空间模型
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2013)03-0764-04
doi:10.3969/j.issn.1001-3695.2013.03.031

Document vector space model construction based on domain ontology

HAO Wen-ning, FENG Bo, CHEN Gang, JIN Da-wei, ZHAO Shui-ning
(*Engineering Institute of Corps of Engineers, PLA University of Science & Technology, Nanjing* 210007, *China*)

Abstract: As VSM couldn't reveal the same conceptual semantics hidden in different feature words and the latent semantic relations in the documents and had a low accuracy in document similarity calculation, this paper presented DOBVSM. This model added the conceptions in domain ontology to the feature word set, and then adjusted the weight of feature words using the relation between conceptions, constructed the document DOBVSM contains semantic relations. Experimental analysis show that the documents' similarity computed by DOBVSM was more divergent, closest to the expert evaluation and can perform batter in document similarity.

Key words: domain ontology; VSM model; similarity; semantic relation; DOBVSM

在信息化时代，文档是信息的主要载体，由于以自然语言形式存在的文档很难被计算机理解，因此要实现计算机对文档的处理，必须定义文档的表示模型。国内外学者对文档表示进行研究，提出了布尔模型、向量空间模型、概率模型、概念模型等，其中使用最多的是 1969 年 Salton 等人提出的向量空间模型（vector space model，VSM）。它的基本思想是将一个文档 d_j 描述为关于特征词的一个向量（(t_1, w_{j1})，(t_2, w_{j2})，…，(t_n, w_{jn})），每个特征词被认为是一个潜在的属性，采用统计的方法计算特征词 t_i 在 d_j 中的权重 w_{ji}。利用 VSM 的优势在于其知识表示方法[1]：文档被形式化为向量空间中的向量，因此文档间的运算就转换成向量间的数学运算，大大降低了问题的复杂度。但该方法认为各特征词之间的语义关系是独立的，互不相交的，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，导致了模型在计算文档相似度时的准确性不高的现象。因此，上述方法不能揭示隐藏在不同特征词后面的相同概念语义和反映文档中的潜在语义，不能满足像军事等这样的专业领域需要建立文档间的语义关系，以及在相似度计算中需要较高精度的要求。

1　相关研究

近几年，潜在语义标引（LSI）成为最为流行的文档维度归约算法，它通过 SVD 分解文档矩阵，找出最具代表性的特征，挖掘文档中的潜在语义关系，最终实现文档的降维[2]。该算法认为：具有较高词频信息的特征词之间具有潜在语义关系；这种使用统计的方法而不是从特征词的含义本身出发建立的文档语义关系并不能真实地揭示隐藏在不同特征词后面的相同概念语义、反映文档中的潜在语义关系。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。本体是特定领域客观存在的概念与概念之间关系的描述。在文本表示中，可以将一个文本分解为多个概念，利用本体得到概念之间的语义关系。基于本体的文本表示方法揭示了文本语义层面的知识，无疑会带来更好的文本表示效果。目前许多学者都进行了相关的研究，Oleshchuk 等人[3]认为文档在本体中所对应的部分实际上是一个子本体，那么文档之间的相似度可以通过子本体之间的相似度表示，但其计算方法过于简单，反映不出本体中的语义关系。2008 年 Jing 等人[4]采用基于本体的互信息测度来计算特征项之间的相似度，利用 WordNet 本体得到两个特征项

收稿日期：2012-07-24；**修回日期**：2012-08-27

作者简介：郝文宁（1971-），男，山西运城人，教授，博士，主要研究方向为海量高维数据归约、作战效能评估；冯波（1987-），男，四川广安人，硕士研究生，主要研究方向为数据挖掘（fengbogjk@163.com）；陈刚（1974-），男，重庆人，副教授，硕士，主要研究方向为作战指挥训练模拟；靳大尉（1979-），男，河北保定人，讲师，硕士，主要研究方向为作战效能评估；赵水宁（1971-），男，江西南昌人，讲师，博士，主要研究方向为军用数据及知识工程.

之间的距离,再通过计算特征项与文档之间的互信息值来得到特征项在该文档中的权值,将文档表示成特征项的权值向量,最终通过计算向量的距离得到文档之间的相似度,从而完成文本聚类。2009 年 Song 等人[5]提出了基于本体的遗传聚类算法,该算法利用 WordNet 本体进行文本表示,并提出一种本体中的概念相似度计算方法,最终采用遗传算法进行聚类。国内,2008 年谢红薇等人[6]在 VSM 中应用本体的概念,将文本中每个特征项与本体匹配,进而调整特征项的权值,得到新的特征向量,改进了 VSM 缺乏语义知识的不足。宋玲等人[7]提出将本体作为背景知识引入到概念及概念间相似度和文档间相似度的计算中,通过图模型表示本体中的概念以及概念间的语义关系,将一个概念和一个文档扩展为一个语义模糊集,并计算模糊集合之间的相似度。2010 年朱会峰等人[8]提出在文本的表示上引入 WordNet,并定义了关键概念集,使用 WordNet 中的概念节点及概念间的语义关系减少文本特征向量维数,算法使用文本的关键概念集和概念特征向量计算文本相似度。

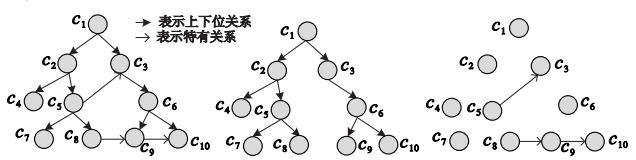
结合上述研究,本文提出了一种文档的改进向量空间模型——DOBVSM。

2　基于领域本体的文档向量空间模型构建

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。因此,本文提出的 DOBVSM 将文档的特征词集合中增加了领域本体的概念,并通过本体中概念间的语义关系和转换策略赋予了所有领域特征词更高的权值,建立了特征词之间的语义关系。DOBVSM 的构建包括领域本体中概念间关系的计算和特征词权值的转换两部分。

2.1　领域本体中概念间语义关系

对于本体 O,定义 $G=(V,E)$。其中 V 是顶点集合,每个顶点代表一个概念;E 是边集,表示概念间的语义关系,这种语义关系是一种等价关系,即满足自反性、对称性以及传递性,如图 1(a)所示。其中,将本体中上下位关系构成的树型结构定义为本体的层次树,如图 1(b)所示;将非上下位关系构成的图定义为本体的相关概念的特有关系如图 1(c)所示。本体中概念间的语义关系包含概念间语义相似度和概念间语义相关度[9]。概念间的语义相似度主要反映在本体中的上下位关系,即本体的层次树中概念间关系;而概念间的相关度主要反映了本体中概念间特有的关系。此处指概念间的非上下位关系,即相关概念的特有关系图中概念节点间的关系。



(a) 本体结构　(b) 本体层次树结构　(c) 相关概念特有关系

图1　本体结构和概念关系

2.1.1　概念间语义相似度计算

相关研究表明[10,11],本体层次树结构中影响概念节点 c_i、c_j 的相似度的主要是 c_i、c_j 两个节点在树中的最短路径距离,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,两概念的相似度就越大;同时两概念所处的层次深度和概念节点的层次差,因为在本体层次树中,概念所处的层次越低,分类越精细,所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,而处于同一层次的概念相似度小于处于不同层次的概念间相似度。

定义 1　语义距离[12]。设 c_i、c_j 为本体层次树中任意两个概念节点,语义距离 $\mathrm{distance}(c_i,c_j)$ 表示从 c_i 到 c_j 所经过的路径长度。

定义 2　语义重合度。设本体层次树的根为 R,c_i、c_j 是树中的任意两个概念节点,$\mathrm{NodeSet}(c_i)$ 是从 c_i 出发,向上直到根 R 所经过的概念节点集合。$\mathrm{NodeSet}(c_i)\cap\mathrm{NodeSet}(c_j)$ 表示 c_i 和 c_j 到 R 共同经过的节点集合。$\mathrm{NodeSet}(c_i)\cup\mathrm{NodeSet}(c_j)$ 表示 c_i 到 R 经过的概念节点集合和 c_j 到 R 经过的节点集合的并集;$\dfrac{\mathrm{NodeSet}(c_i)\cap\mathrm{NodeSet}(c_j)}{\mathrm{NodeSet}(c_i)\cup\mathrm{NodeSet}(c_j)}$ 表示概念 c_i、c_j 之间的语义重合度。

定义 3　节点层次。设 c_i、c_j 为本体层次树中任意两个概念节点,$\mathrm{Level}(c_i)$ 表示节点 c_i 在本体层次树中所处的层次,$|\mathrm{Level}(c_i)-\mathrm{Level}(c_j)|$ 表示节点 c_i 和节点 c_j 的层次差。

定义 4　概念节点 c_i、c_j 的相似度 $\mathrm{sim}(c_i,c_j)$ 定义[13]如下:

$$\mathrm{sim}(c_i,c_j)=\begin{cases}1 & c_i=c_j\\ \dfrac{\alpha}{(\mathrm{distance}(c_i,c_j)+\alpha)}\times\dfrac{\beta\times|\mathrm{NodeSet}(c_i)\cap\mathrm{NodeSet}(c_j)|}{|\mathrm{NodeSet}(c_i)\cup\mathrm{NodeSet}(c_j)|}\times & \\ \dfrac{1}{\gamma\times|\mathrm{Level}(c_i)-\mathrm{Level}(c_j)+1|} & c_i\neq c_j\end{cases}\quad(1)$$

当 $c_i=c_j$ 时,$\mathrm{sim}(c_i,c_j)=1$,即概念与概念自身间的相似度为 1。式(1)中,α、β、γ 为三个可调节的参数,分别反映语义距离、语义重合度和节点层次差对语义相似度的贡献程度。由式(1)可以看出,概念节点语义相似度 $\mathrm{sim}(c_i,c_j)$ 在[0,1]之间。

2.1.2　概念间语义相关度计算

概念间的语义相关度主要反映本体中非上下位概念之间的关系,如图 1(b)中,(c_5,c_3) (c_8,c_9) (c_9,c_{10}) 之间存在着非上下位关系。概念间语义相关度[13]采用如下计算公式:

$$\mathrm{rel}(c_i,c_j)=\begin{cases}1 & c_i=c_j\\ \dfrac{\lambda}{\mathrm{Shortest}P(c_i,c_j)+\lambda} & \text{其他关系}\end{cases}\quad(2)$$

其中:$\mathrm{Shortest}P(c_i,c_j)$ 表示 c_i 到 c_j 的最短路径长度,如果 c_i、c_j 不连通,则 $\mathrm{Shortes}P(c_i,c_j)=\infty$,$\lambda$ 是一个可调节的参数。

2.1.3　概念间语义关系计算

概念间的相似度主要反映了本体中的上下位关系,概念间相关度主要反映了本体中定义的特有关系,概念间的语义关系为两者的组合,因此,定义概念间语义关系 $\mathrm{simrel}(c_i,c_j)$ 为

$$\mathrm{simrel}(c_i,c_j)=\mathrm{sim}(c_i,c_j)+\mathrm{rel}(c_i,c_j)-\mathrm{sim}(c_i,c_j)\times\mathrm{rel}(c_i,c_j)\quad(3)$$

利用式(3)计算出本体概念集合中所有概念之间的语义关系[14],形成本体的语义关系矩阵 S,则 $S=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1m}\\ s_{21} & s_{22} & & \cdots\\ \vdots & & \ddots & \vdots\\ s_{m1} & & & s_{mm}\end{bmatrix}$。$s_{ij}=\mathrm{simrel}(c_i,c_j)$ 表示概念 c_i、c_j 之间的语义关系,S 为对称矩阵。

2.2 文档的 DOBVSM 构建

基于领域本体的文档向量空间模型其基本思想是:首先抽取出文档中的特征词集合,将领域本体概念集合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[15]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 DOBVSM。

设文档 d 采用自然语言处理技术得到文档的特征词集合 $T=\{t_1,t_2,\cdots,t_n\}$,本体库中的概念集合 $C=\{c_1,c_2,\cdots,c_m\}$,则集合 T 与 C 的交集构成的集合 $T'=T\cup C$,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 $t_i\subset C$,则称 t_i 为领域类特征词。为了方便问题的描述,对于具有各种特点的特征词作如下的规定:特征词集合 T 中不属于领域本体概念的特征词构成的集合为 T_1,则 $T_1=T-C$;集合 T 与 C 的交集构成的集合 $T_2=T\cap C$ 中的特征词表示出现在文档中又属于领域本体中的概念;集合 T 与 C 的交集构成的集合 $T_3=C-T$,表示只出现在领域本体的概念集合中而没有出现在文档中的特征词,则文档的特征词扩展集合 $T'=T\cup C=T_1\cup T_2\cup T_3$。

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。具体过程如下:

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

$$w_i=f(t_i)=\begin{cases}0 & t_i\in T_3\\ tf_i\times idf_i & t_i\in T\end{cases}\tag{4}$$

其中:tf_i 为特征词 t_i 在文档 d 中出现的频率,反映了特征项在文档中的重要程度,值越大,表示该词对文档的描述能力越高,更能准确地反映文章的内容;idf_i 为反文档频率,特征项在文档集合中分布情况的量化,是一个普遍重要性的度量。$idf_i=\log(\frac{N}{df_i}+L)$,$df_i$ 为在整个文档集合中包含特征词 t_i 的文档数,N 为文档集中文档总数,L 的取值通过实验确定,一般取值为 0.01。当 $t_i\in T$ 时,特征词 t_i 权重 w_i 为

$$w_i=tf_i\times idf_i=\frac{tf_i\times\log(N/df_i+0.01)}{\sqrt{\sum_{t_i\in T}|tf_i\times\log(N/df_i+0.01)|^2}}\tag{5}$$

其中:N 为总文档数,分母是归一化因子,它是为了减少文档长度不同带来的负面影响,对权重作的泛化处理。

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。本文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$((t_1,w_1),(t_2,w_2),\cdots,(t_m,w_m))$,其中特征词的权值向量 $W=(w_1,w_2,\cdots,w_m)^{\mathrm{T}}$,令调整后的领域类特征词权值向量为 W',则

$$W'=S\cdot W=\begin{bmatrix}S_{11}&S_{12}&\cdots&S_{1m}\\S_{21}&S_{22}&\cdots&S_{2m}\\\vdots&&\ddots&\vdots\\S_{m1}&S_{m2}&&S_{mm}\end{bmatrix}\begin{bmatrix}w_1\\w_2\\\vdots\\w_m\end{bmatrix}=\begin{bmatrix}w_1'\\w_2'\\\vdots\\w_m'\end{bmatrix}\tag{6}$$

由式(6)可知,对于每一个属于领域类特征词,它的权重与他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,前者对于文档来说是更为重要,因此,需要增加领域类特征词的权重。对于领域类特征词,通过式(6)调整后的权值并没有采取归一化处理,这一过程使得 $w_i'>w_i$,增大领域类特征词的权重,提高了领域类特征词对于文档的贡献程度,构建的文档 DOBVSM 更具有现实意义。

2.3 模型构建的算法描述

文档的 DOBVSM 构建包含两个过程:首先是计算出领域本体中概念间的相似度,得到相似度矩阵,具体过程如图 2 所示;然后对文档进行预处理,进行特征词的抽取,并进行特征词集合的扩展和基于概念间语义关系的权值调整。描述如下:

输入:专业领域文档集,领域本体库。

输出:文档 DOBVSM 表示。

a)文档进行预处理,得到文档特征词 T;采用 Jena 包解析本体文件,得到领域概念集合 C;

b)得到文档扩展特征词集合 T',$T'=T\cup C$;

c)for(每一个特征词 t_i)

d) 采用式(4)计算各特征词 t_i 的权重 w_i,作为初始权重;

e) for(每一个领域类特征词 t_j)

f) 采用式(6)调整 t_j 的权重值得到 w_i';

g)得到文档的 DOBVSM。

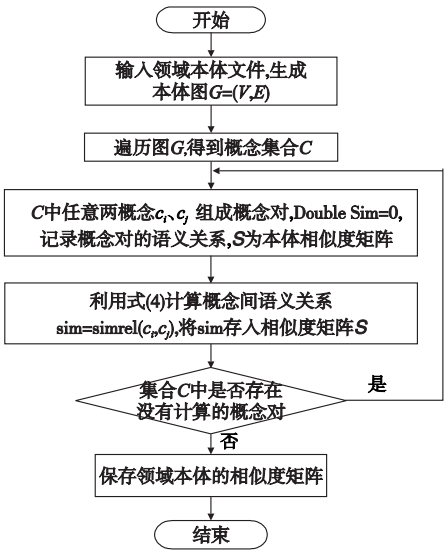


图2 领域本体中概念相似度计算流程

3 实验分析

3.1 实验描述

为了验证基于领域本体的文档向量空间模型在解决文档间语义相关问题中的效果,选择计算文档相似度作为参考标准。使用军事领域专家已经构建的武器装备本体为基础,根据

式(3)计算武器装备领域的概念词之间的语义关系,建立武器装备领域中概念词之间的语义关系矩阵。同时在描述武器装备性能的文档库中选取了 10 篇文档作为研究对象,分别编号为 $d_1,d_2,\cdots,d_{10}$,10 篇文档中编号分别为 1、2、3 的文档描述的是世界各国航母的相关性能,编号为 4、5 的文档描述的是现代主战坦克的相关信息,编号为 6、7 的文档描述的是世界武装直升机的性能,编号为 8、9、10 的文档主要描述的是舰载机的相关数据。对选取的研究文档采用中国科学院的 ICTCLAS 分词系统对文档进行分词,去停用词等预处理操作得到文档的特征词集合,构建文档的 DOBVSM,并利用余弦函数计算文档的相似度。为了验证算法结果的可比性,采用了传统的"VSM + 余弦定理"计算相似度,同时将实验文档分发给五位军事领域专家,让专家根据领域背景知识来判断文档的相似程度,并给出[0,1]内的数值作为相似度度量,取其中的平均值作为最终结果。

3.2　实验结果及分析

根据实验要求计算出文档间的相似度,选择其中部分文档间相似度值生成折线图,如图 3 所示。其中专家评价的相似度值能够较为真实地反映文档的相似度大小,因此,以专家评价的相似度值为标准。观察传统的 VSM 的相似度曲线可知,该方法计算出的相似度取值主要集中在区间[0.1,0.4]内,且当相似度较大时,VSM 计算的相似度值与专家评价的值差距较大,而在相似度取值较小的区域,VSM 计算的相似度值与专家评价的值较为接近,如在 $\mathrm{sim}(d_4,d_5)$、$\mathrm{sim}(d_1,d_2)$、$\mathrm{sim}(d_8,d_9)$ 点处相似度取值较大,在 $\mathrm{sim}(d_4,d_9)$ 处相似度都在 0.2 附近。因为传统的 VSM 没有考虑特征词之间的语义关系,只是以特征词在文档集合中的统计信息给出权值,计算出的相似度值过于集中,取值区间较小,不能真实地反映文档间的语义关系。本文提出的 DOBVSM 计算出的相似度值与专家评价的文档相似度值较为接近,曲线的总体趋势是一致的,较好地反映出了文档之间的相似度;相似度值的波动范围为[0.1,0.7],相对于传统的 VSM,改进方法计算出的相似度更加发散,具有更强的区分度。因为 DOBVSM 通过领域本体结构图计算出领域特征词之间的语义关系,并在计算文档领域特征词权值时,以特征词之间的语义关系对采用统计方法得到的权值进行调整。该过程中不仅建立了特征词之间的语义关联,同时还赋予了领域特征词更高的权重,计算出的语义相似度更加科学、合理,更加精确地反映出文档之间的相似度。

图 4 描述的是文档 d_2 与其他 9 篇文档的相似度情况。

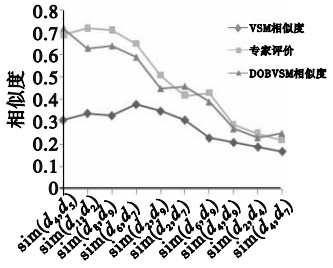


图3　文档集中部分文档相似度

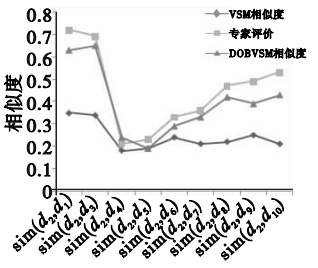


图4　d_2与其他各文档相似度情况

图 4 中,文档 d_2 描述的是航母的相关信息,观察 VSM 相似度曲线,与文档 d_2 相似度较高的为 d_1、d_3,都是描述与航母相关的信息,其余的文档与 d_2 相关度都比较低,约在 0.2 附近。DOBVSM 计算出的相似度与专家评价比较一致,计算出的相似度值具有很好的区分度,能够判断出与 d_2 有较强语义关联的文档是舰载机类文档,相关度值在 0.4 附近,然后是直升机类文档,最后是坦克类文档。综合以上分析可知,本文的文档 DOBVSM 能够揭示特征词中的语义关系,反映文档间的语义关系,用于相似度计算具有较高的准确率和良好的区分效果。

4　结束语

针对领域文档信息量较少的特点,文档间的相似度不能简单地以特征词的统计信息确定,更应该考虑属于领域概念的特征词之间的语义关系来增强相似度计算的准确性。本文提出的基于领域本体的向量空间模型不仅增加了领域类特征词之间的语义关联,同时较非领域特征词,赋予了领域特征词更高的权重,使得构建的向量空间模型更加科学、合理。在用于文档相似度计算时,更加精确地反映了文档间的相似度。实验表明:改进的 VSM 计算的相似度值最为发散,与专家的评价值更为接近,能够较好地反映文档之间的相似度。然而,文档向量原始特征的数量很大,向量处于一个高维空间中,由于计算机内存容量的限制和处理速度的要求,必须对向量空间模型进行降维处理,将特征向量空间从高维变换到低维空间。因此,如何对 DOBVSM 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] HAN Jia-wei, KAMBER M. Data mining concepts and techniques [M]. 2nd ed. [S.l.]: Morgan Kaufmann, 2006.

[3] OLESHCHUK V, PEDERSEN A. Ontology based semantic similarity comparison of documents[C]//Proc of the 14th International Workshop on Database and Expert Systems Applications. 2003:735-738.

[4] JING Li-ping, ZHOU Li-xin, 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] SONG Wei, LI Cheng-hua, PARK S C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures[J]. Expert Systems with Applications, 2009,36(5):9095-9104.

[6] 谢红薇,颜小林,余雪丽. 基于本体的 Web 页面聚类研究[J]. 计算机科学,2008,35(9):153-155.

[7] 宋玲,郭家义,张冬梅,等. 概念与文档的语义相似度计算[J]. 计算机工程与应用,2008,44(35):163-167.

[8] 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:理工版,2010,48(2):277-283.

[9] 甘健侯,姜跃,夏幼明. 本体方法及其应用[M]. 北京:科学出版社,2011.

[10] KNAPPE R, BULSKOV H, ANDREASEN T. Similarity graphs [C]//Proc of the 14th International Symposium on Methodologies for Methodologies for Intelligent Systems. 2003:668-672.

[11] 徐德智,王怀民. 基于本体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.

[12] 秦春秀,赵捧未,刘怀亮. 词语相似度计算研究[J]. 情报理论与实践,2007,30(1):105-108.

[13] 聂卉,龙朝晖. 结合语义相似度和相关度的概念扩展[J]. 情报学报,2007,26(5):728-732.

[14] 张选平,蒋宇,袁明轩. 一种基于概念的信息检索查询扩展[J]. 微电子学与计算机,2006,23(4):110-114.

[15] 李鹏,陶兰,王弼佐. 一种改进的本体语义相似度计算及应用[J]. 计算机工程与设计,2007,28(1):227-229.