基于语言网络和语义信息的文本相似度计算

詹志建,杨小平
ZHAN Zhijian, YANG Xiaoping

中国人民大学 信息学院 计算机系,北京 100872
Department of Computer, School of Information, Renmin University of China, Beijing 100872, China

ZHAN Zhijian, YANG Xiaoping. Text similarity calculation based on language network and semantic information. Computer Engineering and Applications, 2014, 50(5):33-38.

Abstract: Aiming at the shotcoming of traditional text similarity methods with statistical information of word frequency and semantic information of word in text, it proposes a new text similarity calculation based on language network and word semantic information. This new method extracts feature items based on the feature values of the word nodes in a documental language network. It also considers both the importance of feaure items and the semantic relations among feature items, and proposes to construct a semantic network of document feature items to calculate the similarity of documents. Finally it uses several *K*-means clustering methods for evaluating preformance of the new text document similarity. Experimental results show that the method's *F*-measure is superior to the others' which proves that the proposed method is effictive.

Key words: language network; text clustering; text similarity; term semantic similarity

摘 要: 通过分析已有的基于统计和基于语义分析的文本相似性度量方法的不足,提出了一种新的基于语言网络和词项语义信息的文本相似度计算方法。对文本建立语言网络,计算网络节点综合特征值,选取TOP比例特征词表征文本,有效降低文本表示维度。计算TOP比例特征词间的相似度,以及这些词的综合特征值所占百分比以计算文本之间的相似度。利用提出的相似度计算方法在数据集上进行聚类实验,实验结果表明,提出的文本相似度计算方法,在*F*-度量值标准上优于传统的TF-IDF方法以及另一种基于词项语义信息的相似度量方法。

关键词: 语言网络;文本聚类;文本相似度;词语相似度
文献标志码: A **中图分类号:** TP311 doi:10.3778/j.issn.1002-8331.1308-0263

1 引言

随着科学技术的飞速发展,网络上的资源数量呈指数级增长,其中很大一部分属于文本数据,例如网页文本、XML文档、电子邮件和报表数据等。面对这些文本数据,如何进行科学有效的管理,已经成为信息技术领域的研究热点。如果没有有效的组织和提取方式,普通用户很难快速查找到所需信息。文本聚类是有效管理文本数据的方法之一。文本聚类是指自动地将文本数据划分为不同的类别,同类别中的文本数据非常相似,不同类别中的文本数据基本不相似。文本相似度量又是文本聚类的核心问题之一。文本相似度研究的是两个或多个文本之间匹配程度的一个重要参数[1]。相似度越大,说明文本之间相似程度越高,反之就越低。

文本的相似度量方法在许多领域有着广泛的应用:在自动问答领域,文本相似度量方法被认为是改进问答效果最好的方法之一[2];在图像检索领域,通过计算图像周边文本的相似度可以获得更好的检索效果[3];另外,对于文本分类[4]、信息检索[5]、文本摘要自动生成[6]、情感分析[7]等领域,文本相似度计算也有着广泛的应用基础。

目前国内外对文本相似度计算的研究方法主要分为基于统计的方法和基于语义分析的方法[8]。基于统计的文本相似度量方法是将文本表示为特征词集合,将特

基金项目: 国家自然科学基金(No.70871115)。
作者简介: 詹志建(1982—),男,博士,主要研究领域为信息处理、语义计算、WEB数据管理;杨小平(1956—),男,博士,教授,主要研究领域为信息系统工程、电子政务、网络安全技术。E-mail:zhanzj@ruc.edu.cn
收稿日期: 2013-08-20 **修回日期:** 2013-10-18 **文章编号:** 1002-8331(2014)05-0033-06
CNKI网络优先出版:2013-11-15, http://www.cnki.net/kcms/detail/11.2127.TP.20131115.1124.013.html

征词作为文本的基本元素,建立特征词向量空间,通过计算特征词向量空间之间的相似度来表征文本之间的相似度。这种相似度量方法需要大规模的语料库,忽略了文本中的语法和组织结构等信息,也忽略了文本中的词语语义信息,而且文本表示模型高维稀疏,计算效率低下。基于语义分析的文本相似度量方法则是利用特定领域的知识库来构建文本特征词之间的语义关系。这种方法不需要语料库,但建立完备的知识库则比较费时费力。

本文针对上述方法存在的缺陷,提出了一种既能有效降低文本表示模型的维度,又能结合词项语义信息进行相似度计算的方法。给定两个文本,通过本文提出的算法,能够高效、快速地计算出两个文本在语义层次上的相似度,并且能够在较为广泛的应用领域内使用。

2 相关工作

目前,文本相似度计算研究中大多采用向量空间模型(VSM)进行文本表示[9],许多实际运行的信息检索系统也采用了这样的模型。VSM模型的基本思想是将文本表示成一个向量,向量的每一维表示文本的一个特征,该特征通常是一个字或词。VSM模型以其简单的表示方法,良好的文本表示效果获得了学界的青睐。采用VSM模型计算文本相似度,最重要的是计算词语的权重,使用最多的是TF-IDF方法[10]。采用TF-IDF方法,可以排除文本中那些高频低区分度的词,但该方法忽略了词语在文本中不同位置重要性不同的问题,并且在利用VSM模型计算文本间相似度时,认为文本间的词语是相互独立的,没有考虑词语之间的语义关系。文献[11]在上述理论的基础上,对文本进行区域划分,并赋予文本不同区域的关键词以不同权重,给出了含有位置关系的新的权重计算方法,但这一方法仍没有考虑词语间的语义关系。词语间的语义关系,主要考虑词语间的相似度关系,文献[12]通过考察词义网中节点密度、深度和链接类型等因素提出了一种基于词义网边的词语相似度计算方法。文献[13]利用WordNet研究了局部相关性信息以此来确定文本之间的相似性。文献[14]提出了一种基于贪婪匹配和词语语义信息的短文本相似度计算方法。文献[15]提出了一种基于WordNet和GVSM的文本相似度计算方法,通过语义的路径长度和路径深度计算两个词的语义相似度,结合改进的GVSM模型计算文本相似度。但该方法未对向量维数进行处理,计算效率低下。

文献[16]提出文本 D 的主要内容可以用语言网络 G 来表示,即 G 表示了 D 的主题。文本 D 的主题又是由一系列子主题组成,这些子主题对应的就是 G 的连通子图。而连通子图中的中心高频词和连接两个子图的相对低频词,就是对 G 具有关键作用的词语,可以用来表征文本的特征。基于该思想,在图1中子图的中心词语 b、d 和 g,以及子图间起连接作用的词语 f 都将作为 G 的特征词提取出来。

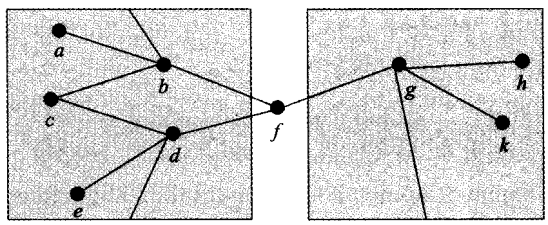

图1 语言网络图示例

基于上述对文本相似性方法的分析以及语言网络的特征,本文提出了一种结合语言网络和词语语义信息的文本相似度计算方法。本文的贡献有以下几个方面:首先通过建立文本的语言网络选取文本中的重要词项,有效降低文本模型的维度,为文本的语义相似度计算提供一个合适的表征模型。其次,通过分析重要词项的语义相似性,给出了文本相似度定义。最后通过几种主流的聚类实验,验证本文提出的文本相似性度量方法是否有效。实验对比了传统的TF-IDF相似度量方法和文献[17]提出的TsemSim文本相度量方法。实验表明,本文的方法在 F-度量指标上优于这两种方法。

3 基于语言网络和词项语义信息的文本相似度计算

3.1 语言网络构建

文本中包含有原始的丰富信息,在对文本建立语言网络前,需要对文本进行预处理,包括分词、取词和过滤掉一些停用词等无意义的词。

文本内容预处理完后,对词项建立语言网络[18],语言网络就是将每个词项作为一个节点,词项在同一个句子中共同出现构成节点之间的连边。连边的收集,若只采集两个紧邻的词项之间的联系,则可能会丢失一些长程的关联,同时却提高某些无用词在网络中的重要性。因而需要确定词项在句子中的关联跨度。若跨度太短,很多重要的关联无法记录;若跨度太大,可能产生许多冗余信息。本文采用文献[19]中所遵循的规则,即取关联跨度最大为2进行研究,因为这种长度的关联在语言网络中最为常见和重要。例如,对于句子“文本相似度计算过程”,通过分词产生“文本”、“相似度”、“计算”、“过程”四个词语,从而可以建立如图2所示的语言网络。对于整个词条正文的语言网络,则可以通过合并各个句子语言网络中的相同节点与连边来产生。

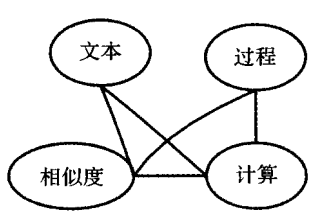

图2 一个句子的语言网络

语言网络可以用符号表示为:$G=(V,E)$,其中 $V=\{v_i|i=1,2,\cdots,n\}$ 为顶点的集合,n 为语言网络中节点的个数。$E=\{(v_i,v_j)|v_i,v_j\in V\}$ 表示边的集合。

3.2 文本特征词项选择

目前,学界公认的语言网络的重要特征有度分布、平均最短路径、聚集度和聚集系数[20]。本文利用语言网络中节点的度和聚集性质提取文本的特征词项。

节点 v_i 的度定义:

$$D_i=|\{(v_i,v_j):(v_i,v_j)\in E,v_i,v_j\in V\}|$$

节点 v_i 的聚集度定义:

$$K_i=|\{(v_j,v_k):(v_i,v_j)\in E,(v_j,v_k)\in E,v_i,v_j,v_k\in E\}|$$

节点 v_i 的聚集系数定义:

$$C_i=\frac{K_i}{D_i(D_i-1)/2}=\frac{2K_i}{D_i(D_i-1)}$$

节点的度体现该节点与其他节点的关联情况,节点的聚集度和聚集系数体现在此节点局部范围内的节点相互连接密度。节点的度和聚集系数体现该节点在局部范围内的重要性。

节点 v_i 的聚类系数定义:

$$G_i=\sum_{i\neq j\neq k}\frac{g(i)_{jk}}{g_{ik}}$$

$g(i)_{jk}$ 表示语言网络 G 中连接节点 v_j 和 v_k,并且经过节点 v_i 的最短路径的条数。g_{jk} 表示连接节点 v_j 和 v_k 的所有最短路径的条数。节点聚类系数表示网络中任意两点的最短路径经过该节点的数量比例,反映了节点对整个网络信息流动的影响。网络中聚类系数比较大的节点,处于较重要的位置。

定义语言网络节点的综合特征值 CF_i:

$$CF_i=0.5\times\frac{C_i}{\sum_{i=1}^{n}C_i}+0.5\times\frac{G_i}{n} \tag{1}$$

对节点词项按 CF 值从大到小进行排序,从中选取 TOP 比例的词项作为文本特征词项,以此特征词向量作为词条正文的特征表示。与传统的词频向量相比,一篇正文的特征词向量维度下降了 $1-TOP$,这在效率上是一个较大的提高。

3.3 文本相似度计算

特征词向量代表了一篇文本中最重要的信息,文本间的相似度计算就可以转化为特征词向量之间的相似度计算。另外,由于每篇文本长度不同,因而表征每篇文本的特征词向量的维度也不一样,必须统一标准,尽量消除这些影响,使得特征词向量间的相似度满足基本的相似度量标准。

设 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 是两篇不同文本的特征词向量,其中 $\boldsymbol{v}_i=(w_{i1},w_{i2},\cdots,w_{im})$,$v_j=(w_{j1},w_{j2},\cdots,w_{jn})$。定义文本之间的相似度为:

$$TSim(\boldsymbol{v}_i,\boldsymbol{v}_j)=cf\times VectSim(\boldsymbol{v}_i,\boldsymbol{v}_j)+(1-cf)\times CosSim(\boldsymbol{v}_i,\boldsymbol{v}_j) \tag{2}$$

式中 cf 表示特征词向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 之间相似度的加权因子,$VectSim(\boldsymbol{v}_i,\boldsymbol{v}_j)$ 表示特征词向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 之间的向量相似度,$CosSim(\boldsymbol{v}_i,\boldsymbol{v}_j)$ 表示特征词向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 之间的余弦相似度。如果两篇文本中彼此相似度较高的词项越多,词项所占的综合特征值在各自文档中比例越高,说明这些词项更能反映它们在文本中的重要性。因此,先找出满足相似度阈值条件的特征词项,再计算这些特征词项的综合特征值之和,最后求出上述结果值占整篇文本综合特征值总和的比例进行加权,具体的加权因子计算公式由式(3)给出:

$$cf=\frac{1}{2}\times\left\{\frac{\sum_{k\in\Lambda_i}CF_{ik}}{\sum_{k=1}^{m}CF_{ik}}+\frac{\sum_{l\in\Lambda_j}CF_{jl}}{\sum_{j=1}^{n}CF_{jl}}\right\}\times\frac{(|\Lambda_i|+|\Lambda_j|)/2}{\max(m,n)} \tag{3}$$

式(3)中的 CF_{ik} 表示特征词项 w_{ik} 的语言网络综合特征值,右端项分子部分表示满足相似度阈值条件的特征词项的综合特征值之和,分母部分表示所有特征词项综合特征值之和。式(3)中的集合 Λ_i 和 Λ_j 定义如下:

$$\Lambda_i=\{k:1\leqslant k\leqslant m,\max\{Sim(w_{ik},w_{jl})\}\geqslant\mu\}$$

$$\Lambda_j=\{l:1\leqslant l\leqslant n,\max\{Sim(w_{jl},w_{ik})\}\geqslant\mu\}$$

如果特征词向量 $\boldsymbol{v}_i$ 中的某个关键词 w_{ik} 与另一个特征词向量 $\boldsymbol{v}_j$ 中的关键词 $w_{jl}(l=1,2,\cdots,n)$ 的相度阈值超过用户设定的相似度阈值,则将该特征词 w_{ik} 放入到集合 Λ_i 中。集合 Λ_j 包含的元素依据集合 Λ_i 的方法对特征词向量 $\boldsymbol{v}_j$ 中的特征词进行选择。$|\Lambda_i|$、$|\Lambda_j|$ 分别表示集合 Λ_i 和 Λ_j 的元素个数。集合元素越多,说明满足相似度阈值条件的词项个数越多,对相似度的影响越大。$Sim(w_{jl},w_{ik})$ 表示特征词 w_{jl} 和 w_{ik} 之间的语义相似度。

$$VectSim(\boldsymbol{v}_i,\boldsymbol{v}_j)=\frac{1}{2}\left(\frac{1}{m}\sum_{k=1}^{m}\max\{Sim(w_{ik},w_{jl})\}+\frac{1}{n}\sum_{l=1}^{n}\max\{Sim(w_{ik},w_{jl})\}\right) \tag{4}$$

$VectSim(\boldsymbol{v}_i,\boldsymbol{v}_j)$ 由向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 中所包含的词项相似度决定。相似的向量必定包含相似度较高的词项,而不相似的向量则彼此包含的词项相似度较低。

$$CosSim(\boldsymbol{v}_i,\boldsymbol{v}_j)=\frac{\sum_{k=1}^{\max(m,n)}CF_{ik}\times CF_{jl}}{\sqrt{\sum_{k=1}^{m}CF_{ik}^2\times\sum_{l=1}^{n}CF_{jl}^2}} \tag{5}$$

$CosSim(\boldsymbol{v}_i,\boldsymbol{v}_j)$ 表示向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 的余弦相似度。

3.4 算法解释

算法首先对文本进行预处理,包括格式处理、句子识别、分词、词性标注和去除停用词等,再以预处理后的文本词语为节点,为每个句子中跨度为1或2的节点建立连边,将各个句子所组成的网络连接起来,合并相同的节点和连边,形成文本的语言网络。对语言网络中的每个节点,计算网络综合特征值,然后按网络综合特征值大小进行排序,选取 *TOP* 比例的词项作为文本的特征词向量。接着计算不同文本特征词之间的相似度并算出 $VectSim(\boldsymbol{v}_i, \boldsymbol{v}_j)$ 和加权因子 cf,再根据余弦相似度求得 $CosSim(\boldsymbol{v}_i, \boldsymbol{v}_j)$,最后根据式(1)求出文本 i 和 j 的相似度值。

3.5 基本流程

输入:两篇文本 i 和 j,词语相似度阈值 μ

输出:文本 i 和 j 的相似度值

步骤1 对文本 i 和 j 进行预处理,建立相应的语言网络,根据式(1)计算语言网络中各个节点的综合特征值 CF。

步骤2 对各节点的 CF 值按大小进行排序,选取 *TOP* 比例的词项作为文本的特征词向量,得到文本 i 和 j 的特征词向量 $\boldsymbol{v}_i=(w_{i1}, w_{i2}, \cdots, w_{im})$ 和 $\boldsymbol{v}_j=(w_{j1}, w_{j2}, \cdots, w_{jn})$。

步骤3 从向量 $\boldsymbol{v}_i$ 的词项 w_{i1} 开始,寻找向量 $\boldsymbol{v}_j$ 中与 w_{i1} 相似度最高的词项 w_{jk},记录 w_{i1} 和 w_{jk} 之间的相似度值 θ,并判断 θ 与 μ 大小。如果 $\theta>\mu$,则将 w_{i1} 放入到集合 Λ_i 中。

步骤4 重复步骤3的过程,直至向量 $\boldsymbol{v}_i$ 中所有的词项都在向量 $\boldsymbol{v}_j$ 中找到各自相似度最大的词项,同时记录相似度值并调整集合 Λ_i。

步骤5 统计步骤3和步骤4的相似度值总和,除以向量 $\boldsymbol{v}_i$ 中词项的数量,以此作为向量 $\boldsymbol{v}_i$ 对 $\boldsymbol{v}_j$ 的相似度值 $Sim(\boldsymbol{v}_i, \boldsymbol{v}_j)$。

步骤6 同理求得 Λ_j 和 $Sim(\boldsymbol{v}_j, \boldsymbol{v}_i)$。

步骤7 根据步骤5和步骤6的计算结果,再根据式(4)求得 $VectSim(\boldsymbol{v}_i, \boldsymbol{v}_j)$。

步骤8 计算集合 Λ_i 中所有词项的综合特征值总和,以及集合 Λ_j 中所有词项的综合特征值总和,并根据式(3)计算加权因子 cf。

步骤9 根据式(5)求得向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 的余弦相似度值。

步骤10 根据步骤11、步骤12和步骤13的计算结果,再结合式(2)求得文本 i 和 j 的相似度值。

4 实验

本文实验数据选用复旦大学自然语言处理小组收集与整理的部分文本分类语料库。该部分语料库分10个类别,共2 815篇。每个类别中根据文本内容又细分为不同子类别,如表1所示。

表1 实验数据摘要

类别	文本数量	子类别数量	子类别中最少文本数量	子类别中最多文本数量	子类别平均文本数量
环境	200	6	8	25	33
计算机	200	5	10	22	40
交通	214	8	7	20	26
教育	220	6	6	16	37
经济	325	5	11	14	65
军事	249	8	12	20	31
体育	450	9	9	22	50
医药	204	6	8	20	34
艺术	248	5	7	23	50
政治	505	10	7	18	50

该语料库中各个类别的子类别可以作为各个文本集的标准聚类结果。实验首先采用中科院分词软件ICTCLAS对文本集合进行预处理,然后建立文本语言网络,计算各个分词的综合特征值 CF,从中选取TOP比例的特征词作为文本的特征向量。特征词之间的相似度计算采用文献[21]的方法,再结合本文提出的文本相似度计算方法,对文本数据集进行相似度计算,得到文本相似度计算矩阵。

为了验证本文算法的有效性,同时实现了基于TF-IDF方法的文本相似度矩阵[10]和文献[17]提出的结合词项语义信息的文本相似度计算方法TSemSim,进行聚类结果的比较。采用CLUTO工具包(http://www.ce.umn.edu/~karypis/cluto)进行聚类实验,并实现了CLUTO工具包中 K-均值(DKM)、二分 K-均值(BKM)和凝聚 K-均值(AKM)的聚类算法。

本文实验采用 F-度量值来衡量文本相似度计算结果。F-度量值是根据准确率 P(Precision)和召回率 R(Recall)给出的一个综合的评价指标,其定义如下:

$$F=\frac{2RP}{R+P}$$

$$\text{准确率}P=\frac{\text{返回正确的文本数量}}{\text{计算结果的所有文本数量}}$$

$$\text{召回率}R=\frac{\text{返回正确的文本数量}}{\text{子类中的文本数量}}$$

全局聚类的 F-度量值定义为:

$$F=\sum_i \frac{n_i}{n}\max_j(F)$$

式中,n_i 表示各个子类别的文本数目,n 表示所有文本数量,j 表示计算的聚类结果。F 值越大,聚类结果越好。

实验第一步要确定选取不同TOP比例的语言网络特征词对文本相似度计算的影响,并设定词语相似度阈值 $\mu=0$,即所有的特征词项同等重要。本文选取了3种 K 均值算法中最直接的DKM算法进行聚类,图3给出了

在利用DKM聚类算法进行聚类的条件下,选取不同比例的TOP特征词项对文本相似度影响的实验结果。从图中可以直观地看出,如果选取文本中TOP 40%的比例特征词,能够取得较好的聚类效果。低于这个比例时,因为选取的特征词较少,不能充分表达文本的特征信息,从而使得聚类效果欠佳;而超过这个比例时,选取的特征词过多,特别是当特征词选取比例超过50%时,*F*-度量值有明显的下降。这是因为过多比例的特征词使得无关的词语之间相似度计算结果较低,拉低了文本之间的相似度,使得聚类效果反而随着特征词数量增加而变得不理想。

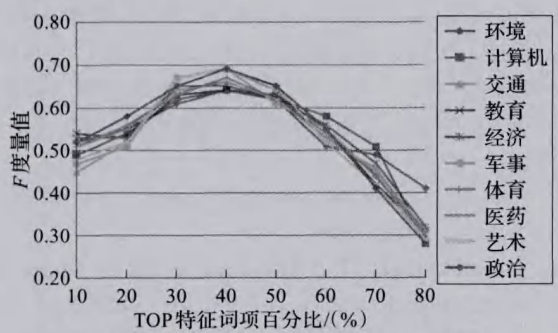

图3 TOP特征词项对聚类结果的影响

实验第二步要确定词语相似度阈值 μ 的大小对聚类结果的影响。图4给出了在选取TOP 40%比例特征词作为文本特征向量时,利用DKM聚类算法进行的聚类条件下,同一聚类中的特征词项相似度阈值 μ 的大小对聚类结果的影响。从图中可以看出,聚类效果随着 μ 值呈现抛物线模型变化,当 μ 落值在区间[0.65,0.70]之间时,达到最好的聚类效果。经过分析,可以得出 μ 值过小或者过大,使得选取的特征词集合 Λ_i 和 Λ_j 元素数量产生变化,而影响到聚类效果。

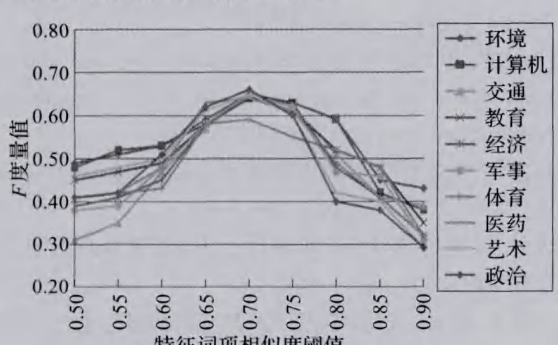

图4 特征词相似度阀值μ对聚类结果的影响

根据以上实验结果,本文选取TOP 40%比例的特征词作为文本特征向量,并取词语相似度阈值 μ 为0.7,采用本文算法与TF-IDF以及TSemSim算法进行比较,比较结果如图5所示。从图中可以看出,采用本文算法,文本相似度计算结果在3种经典的聚类算法下都比采用传统的TF-IDF算法以及TSemSim算法有着更好的*F*-度量值。这表明文本算法能够有效改进聚类效果。

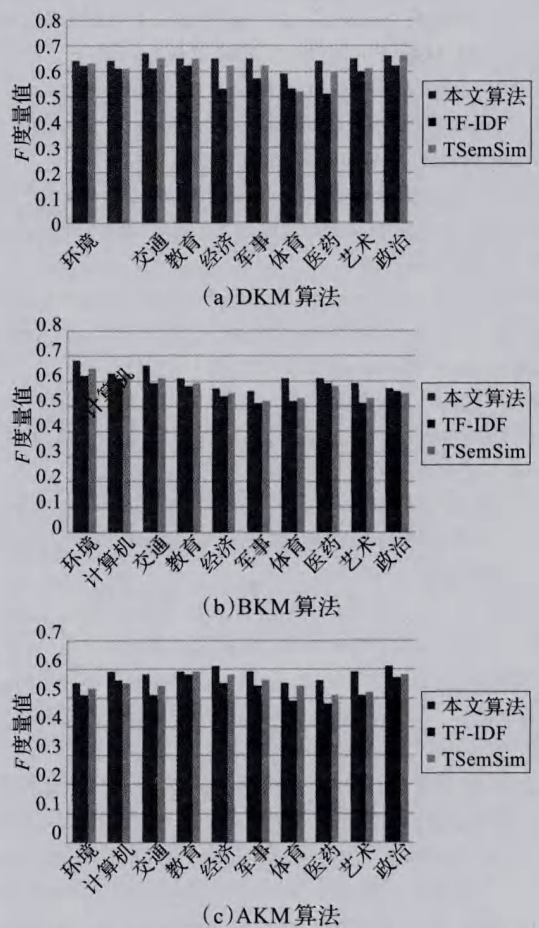

图5 本文算法、TF-IDF算法和TSemSim算法在DKM、BKM、AKM聚类算法上的*F*-度量值比较

5 结论

本文首先通过分析已有的基于统计的和基于语义分析的文本相似性度量方法的不足,提出了一种新的基于语言网络和词项语义信息的文本相似度计算方法。与传统的计算方法相比,本文提出的算法既能够有效降低文本表示模型的维度,又结合词语间的语义相似度计算文本间的相似度。并通过经典的聚类算法实验,验证了本文算法的有效性。

本文的后续工作将在现有语言网络和词项语义信息分析的基础上,进一步深入分析文本中不同位置、不同权重的词语对文本相似度计算结果的影响,综合考虑文本中词语的位置权重、段落篇章结构等信息,更好地提高文本相似度计算精度。

参考文献:

[1] Corley C, Mihalcea R.Measuring the semantic similarity of texts[C]//Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005:13-18.

[2] 熊大平,王健,林鸿飞.一种基于LDA的社区问句相似度计算方法[J].中文信息学报,2012,26(5):40-45.

[3] Coelho T A S, Caladl P P, Souza L V, et al.Image retrieval

using multiple evidence ranking[J].IEEE Trans on Knowledge and Data Engineering,2004,16(4):408-417.
[4] Ko Y,Park J,Seo J.Improving text categorization using the importance of sentences[J].Information Processing and Management,2004,40(1):65-79.
[5] Kumar N.Approximate string matching algorithm[J].International Journal on Computer Science and Engineering,2010,2(3):641-644.
[6] Erkan G,Radev D.Lexrank:graph-based lexical centrality as salience in text summarization[J].Journal of Artificial Intelligence Research,2004,22(7):457-479.
[7] Mitra M,Hadi A,Man L,et al.Sense sentiment similarity:an analysis[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence,2012:1706-1712.
[8] 华秀丽,朱巧明,李培峰.语义分析与词频统计相结合的中文文本相似度量方法研究[J].计算机应用研究,2012,29(3):833-836.
[9] Salton G,Mcgill M J.Introduction to modern information retrieval[M].New York:McGraw-Hill,1983.
[10] Salton G.The SMART retrieval system-experiments in automatic document processing[M].Englewood Cliffs,New Jersey:Prentice Hall Inc,1971.
[11] 谢翠香.基于改进向量空间模型的学术论文相似性辨别系统设计[J].电脑知识与技术,2009,19(5):5103-5105.
[12] Sussna M.Word sense disambiguation for free-text indexing using a massive semantic network[C]//Proceedings of the 2nd International Conference on Information and Knowledge Management(CIKM'93),Washington DC,US,1993:67-74.
[13] Ramage D,Rafferty A N,Manning C D.Random walks for text semantic similarity[C]//Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing,Singapore,2009:23-31.
[14] Lintean M,Rus V.Measuring semantic similarity in short texts through greedy pairing and word semantics[C]//Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference,2012:244-249.
[15] 郑小波,郑诚,尹莉莉.基于GVSM的文本相似度算法研究[J].微型机与应用,2011,30(3):9-11.
[16] 张敏,耿焕同,王煦法.一种利用BC方法的关键词自动提取算法研究[J].小型微型计算机系统,2007,28(1):189-192.
[17] 黄承慧,印鉴,侯昉.一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J].计算机学报,2011,34(5):856-864.
[18] Watts D J.Small worlds[M].Princeton,USA:Princeton University Press,1999.
[19] Cancho R F,Sole R V.The small world of human language[J].Biological Sciences,2011,268(1482):2261-2265.
[20] 赵鹏,蔡庆生,王肖毅,等.一种基于复杂网络特征的中文文档关键词抽取算法[J].模式识别与人工智能,2007,20(6):827-831.
[21] 詹志建,梁丽娜,杨小平.基于百度百科的词语相似度计算[J].计算机科学,2013,40(6):199-202.

(上接32页)

4 结束语

本文面向产品创新设计,结合演化设计方法,对演化设计中的关键技术,即设计知识(产品基因)的定义、获取和表达进行研究。结合生物工程学理论和思想,立足演化设计,对产品实例种群基因进行详细论述。研究了产品实例基因的定义、获取和表达关键技术。通过定义产品实例种群,提出产品设计知识(产品基因),构造产品实例树和产品基因树,最终获得产品基因并进行产品基因表达,以上关键技术是演化设计方法的基础和必要条件,通过演化设计对产品进行遗传、变异、进化等多种操作。产品实例种群基因的提取有利于产品设计知识的规范、传承、积累和重用,与现代产品设计智能化、集成化、自动化发展相辅相成。文章重点研究了产品实例种群基因关键技术,没有涉及到后期具体的演化设计阶段,只对后期的演化设计进行了简要描述。

参考文献:

[1] 宗立成,余隋怀,胡志刚.智能虚拟布局设计关键技术研究与运用[J].计算机工程与应用,2013,49(11):20-23.
[2] Trousse B,Visser W.Use of case-based reasoning techniques for intelligent computer-aided-design systems[C]//Proceedings of the IEEE International Conference on Systems,Man and Cybernetics,1993.
[3] Hsu W,Woon I M Y.Current research in the conceptual design of mechanical products[J].Computer-Aided Design,1998,30(5):377-389.
[4] Pahl G,Beitz W.Engineering design:a systematic approach[M].2nd ed.London:Springer,1996.
[5] Chen D,Burrell P.Case-based reasoning system and artificial neural net works:a review[J].Neural Computing & Applications,2001,10(3):264-276.
[6] 台立刚,钟延修.产品实例种群及产品基因研究[J].上海交通大学学报,2007,41(9):1466-1469.
[7] 滕弘飞,曾威,梁大伟,等.演化设计方法及其应用[J].机械工程学报,2004,40(1):1-6.
[8] 冯培恩,陈泳,张帅,等.基于产品基因的概念设计[J].机械工程学报,2002,38(10):1-6.
[9] 贺淹才.简明基因工程原理[M].北京:科学出版社,1999.
[10] Nossal G J V.塑造完美的生命——遗传工程要旨[M].北京:科学普及出版社,1989.
[11] 张宗炳.遗传与进化[M].北京:人民教育出版社,1981.
[12] 潘正君,唐立山,陈毓屏.演化计算[M].北京:清华大学出版社,1998.
[13] 周明,孙树栋.遗传算法原理及应用[M].北京:国防工业出版社,1999.