文章编号：1004-4353(2016)03-0231-04

基于 *N* 层向量空间模型的文本相似度计算方法

赵雪， 崔荣一*
（延边大学工学院 计算机技术学科 智能信息处理研究室，吉林 延吉 133002）

摘要：针对向量空间模型忽略词语出现位置和词序的缺点，结合科技文献结构明显分层的特点，本文提出了基于 *N* 层向量空间模型的文本相似度计算方法. 该算法首先用 *N* 层向量空间模型表示查询短语和科技文献，其次在词频角度上和词序角度上分别计算两者间的相似度，最后得出整体的文本相似度. 将本文算法应用于中、朝、英对照科技文献多语种检索模块测试其有效性，测试结果表明，本文设计的文本相似度计算方法算法性能较好，与传统的向量空间模型余弦相似度算法相比，查准率提高了 2.7%，MRR 提高了 2.02%.

关键词：向量空间模型；词频；词序；相似度算法
中图分类号：TP391 **文献标识码**：A

The similarity algorithm of texts based on *N*-VSM model

ZHAO Xue， CUI Rongyi*
（*Intelligence and Information Progress Lab.*, *Dept. of Computer Science & Technology*, *College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: Aiming at the disadvantages of vector space model that ignores the occurrence and order of the words, combing with science and technical literature clearly layered structure features, this paper puts forward the similarity algorithm based on *N*-layer vector space model. First we establish the query phrase and science and technical literature *N*-layer vector space model. Then we figure out the similarity between texts in word frequency. Next we figure out the similarity between texts in word order. At last we get the final similarity. The algorithm is applied in Chinese-Korean-English science and technical literature multilingual retrieval module to test the validity. Testing results show that compared with the traditional vector space model cosine similarity algorithm, the new algorithm improves the precision of 2.7%, MRR increases by 2.02%.

Keywords: vector space model; word frequency; word order; similarity algorithm

随着信息时代的飞速发展，如何快速准确地获取所需要的科技文献信息越来越受到研究者的关注[1]. 实现互联网的信息检索需依赖搜索引擎，引擎将搜索到的相关信息按照内容与查询的相似度降序排序并返回结果给用户，因此能否准确地计算检索结果与查询的相似度是满足用户检索需求的关键问题.

实现科技文献的检索需要依赖于信息检索模型的建立，常用的信息检索模型是向量空间模型. 已有许多学者对向量空间模型进行了研究，并取得了一些研究成果，例如：谭静对类似表格结构的文本相似度进行研究时，针对文本段查准率不高的问题，提出了自助加权的文本段向量空间模型[2]，提高了准确率；操卫平设计一种结构化向量

收稿日期：2016-04-23
* **通信作者**：崔荣一(1962—)，博士，教授，研究方向为模式识别、智能计算.

空间模型,利用向量间余弦相似度表示文本间相似度,提高了检索效率,并成功地应用于中文信息检索系统中[3];徐亮提出改进的基于向量空间模型的句子相似度算法,通过计算获取较为准确的句子与标准答案的相似度值,提高了系统的准确率,并成功应用于中文问答系统中[4]. 目前为止,尽管关于相似度的研究很多,但针对传统的向量空间模型忽略词语位置以及词序的问题没有进行有效地改进,在计算查询与科技文献的相似度值时效果仍不理想. 针对这种情况,本文提出了基于 N 层向量空间模型的文本相似度计算方法,改进了传统向量空间模型的余弦相似度计算方法,将词频和词序两方面都纳入相似度的计算中,并通过实验验证了本文方法的有效性.

1 向量空间模型

向量空间模型将文档映射作为特征向量,利用向量之间的距离来逼近文本之间的语义,常用于文本过滤、信息检索、索引以及相关性排名等方面. 向量空间模型假设特征项在文档中出现的频数在某种意义上可以代表文档的含义,并不考虑它们在文档中的位置和顺序. 在衡量 2 个文本之间的相似度时,通过特征项以及特征项的词频来计算 2 个文本的相似性.

向量空间模型的建模思路是假设文本集 $\boldsymbol{D}$ 中有 n 个文档,每个文档 $\boldsymbol{d}$ 有 m 个不同的特征项,特征项之间彼此独立[2],即 $\boldsymbol{d}=(t_1,t_2,\cdots,t_m)$. 给特征项赋予不同的权重后,文本 $\boldsymbol{d}$ 可以抽象为特征向量 $\boldsymbol{V}_d$:

$$\boldsymbol{V}_d=(w_1,w_2,\cdots,w_m). \tag{1}$$

文本集合 $\boldsymbol{D}$ 则可以表示为

$$\boldsymbol{D}=\begin{bmatrix}\boldsymbol{V}_{d1}\\ \boldsymbol{V}_{d2}\\ \vdots\\ \boldsymbol{V}_{dm}\end{bmatrix}=(t_1,t_2,\cdots,t_m)=\begin{pmatrix}w_{11} & \cdots & w_{1m}\\ \vdots & \ddots & \vdots\\ w_{n1} & \cdots & w_{nm}\end{pmatrix}. \tag{2}$$

利用式(2)可以构建文本集合 $\boldsymbol{D}$ 的向量空间模型,通过计算向量之间的余弦相似度就可以获得不同文档的相似度. 向量空间模型特征项的选择方法包括信息增益(IG)、互信息(MI)、统计、期望交叉熵、特征熵、特征权、文本证据权以及几率比等. 选择特征项后,需要为特征项加权. 在加权算法中,$tf\text{-}idf$ 是最为常用的方法,其计算公式如下:

$$tf\times idf=tf_{t,\boldsymbol{Q}}\times\log(M/n_t). \tag{3}$$

式(3)同时考虑到 2 个因素:tf 因子体现了在文本中频繁出现的特征项在模型中赋予较高的权重,idf 因子则是加大了不同文本之间的区分度,说明在许多文献中都出现的词语对于区分相关文献和不相关文献的作用较小.

向量空间模型的主要优点是:1) 结构简单,应用方便;2) 通过对特征项的加权可以改进模型的检索效果,模型的部分匹配策略允许检索出与查询条件相接近的文献;3) 相似度的引进使得能够对查询结果进行排序;4) 通过查询扩展和相关反馈可以改善排序结果,且性能较好. 但该模型也存在一些不足,主要是:由于向量空间模型中的特征项被认为是互相独立的,其 $tf\text{-}idf$ 加权模式只考虑特征项的频率,而不考虑其在文本中的位置. 而事实上,文本信息是以自然语言形式表现的,前后语句之间常常具有很强的语义相关性,特征项在文本里的不同位置对文本有不同的重要程度,因此采用特征项加权方法能够提高向量模型检索的效率.

2 基于 N 层向量空间模型的文本相似度计算方法

本文设计的基于 N 层向量空间模型的文本相似度算法包括 2 个创新点:一是引入 N 层向量空间模型,改进了向量空间模型不考虑词语出现位置的缺点;二是计算词序相似度,改进了向量空间模型不考虑词语上下文关系对相似度影响的缺点. 基于 N 层向量空间模型的算法公式如下:

$$\mathrm{Sim}(\boldsymbol{Q},\boldsymbol{W})=\frac{\sum\limits_{t\in T(\boldsymbol{Q})\cap T(\boldsymbol{W})}w(t,\boldsymbol{Q})\cdot w(t,\boldsymbol{W})}{\sqrt{\sum\nolimits^2 w(t,\boldsymbol{Q})\cdot\sum\nolimits^2 w(t,\boldsymbol{W})}}+\frac{n}{m}\cdot\mathrm{dis}(\boldsymbol{Q},\boldsymbol{W}), \tag{4}$$

其中:$\boldsymbol{Q}$ 表示查询短语;$\boldsymbol{W}$ 表示科技文献;$\mathrm{Sim}(\boldsymbol{Q},$

$\boldsymbol{W}$) 表示查询短语 $\boldsymbol{Q}$ 与科技文献 $\boldsymbol{W}$ 之间的相似度；t 表示查询短语分词后的某一词语；$T(\boldsymbol{Q})$ 表示查询词语 t 的集合；$T(\boldsymbol{W})$ 表示所有与查询短语相关的论文的集合；$w(t,\boldsymbol{Q})$ 表示词语 t 在查询词中的权重，权重采用 $tf\text{-}idf$ 计算；$\mathrm{dis}(\boldsymbol{Q},\boldsymbol{W})$ 表示基于向量距离的词序相似性，$\frac{n}{m}$ 表示包含度.

2.1　词频相似度的计算

向量空间模型虽然用 $tf\text{-}idf$ 方法改进了词语权重，但其本质还是只关心词频而不关心词语出现的位置[4]. 本文算法在向量空间模型的相似度计算中，根据对每一层的实际情况赋予不同的比例因子，然后计算词频的相似度值，这样可以较好地度量出文本之间的相似度. 针对科技文献具有明显的标题、段落、摘要 3 层结构的特点，本文中 N 取 3. 经过多次测试，获得科技文献的标题、关键词、摘要的比例因子为 10∶5∶1.

余弦相似度是借助于几何学角度，利用向量之间的相似性来逼近文本之间的相似性. 如果 2 个文本的特征向量为 $\boldsymbol{V}_{d_i}=(w_{i,1},w_{i,2},\cdots,w_{i,n})$ 和 $\boldsymbol{V}_{d_j}=(w_{j,1},w_{j,2},\cdots,w_{j,n})$，且 2 个向量在空间中的夹角为 θ，则它们之间的余弦相似度计算公式为

$$S(d_i,d_j)=\cos\theta=\frac{\sum_{k=1}w_{ik}\cdot w_{jk}}{\sqrt{\sum_{k=1}^{2}w_{ik}\cdot\sum_{k=1}^{2}w_{jk}}}. \tag{5}$$

用 2 个向量的夹角余弦值可以衡量 2 个文本间的文本相似度，S 值越接近于 1，表明 2 个文本越相似. 向量的权重利用公式[5] $w(i,j)=tf_{i,j}\times idf=tf_{i,j}\times\log(M/n_i)$ 计算. 用 $tf\text{-}idf$ 方法计算词频相似度时，利用比例因子加权来表示每一层模型中的词语对文本的重要程度. 加权后得 $tf_{tj}=\sum_{k}^{N}(l_k\times tf_{tjk})$，其中 l_k 是第 k 层的比例系数，N 为 3，tf_{tj} 表示词语 t 在第 j 篇科技文献中的频率，tf_{tjk} 表示词语 t 在第 j 篇科技文献的第 k 层的加权后频率.

2.2　词序相似度的计算

由于词序对文本相似度值也有着重要的影响，所以引入包含度改进基于向量距离的词序相似度计算方法. 包含度用以度量 2 个文本之间的重合情况，由 $\frac{n}{m}$ 来表示，n 表示查询短语 $\boldsymbol{Q}$ 和科技文献 $\boldsymbol{W}$ 标题的公共向量个数，m 表示科技文献 $\boldsymbol{W}$ 标题的向量元素个数[6]. 包含度度量出公共词语在文献标题中所占有的密度以及公共词语和文献标题的重合情况. 在基于向量距离的词序相似性计算中，只提取公共向量，比较向量间公共元素距离，而忽略向量中的其他元素，这与实际检索的情况不符，融合包含度的向量词序相似度计算方法可以较好地弥补向量空间模型不考虑词语顺序的缺点. 基于向量距离的词序相似性[7] $\mathrm{dis}(\boldsymbol{Q},\boldsymbol{W})$ 的计算公式如下：

$$\mathrm{dis}(\boldsymbol{Q},\boldsymbol{W})=\begin{cases}1-\dfrac{\mathrm{distance}(\boldsymbol{Q},\boldsymbol{W})}{\mathrm{maxDistance}}, & \text{if } n>1;\\ 1, & \text{if } n=1;\\ 0, & \text{if } n=0.\end{cases} \tag{6}$$

其中 $\mathrm{distance}(\boldsymbol{Q},\boldsymbol{W})$ 表示查询短语向量 $\boldsymbol{Q}$ 与科技文献 $\boldsymbol{W}$ 标题之间基于向量距离的词序相似度大小，maxDistance 表示 2 个向量之间的最大距离，计算公式为 $\frac{n^2}{2}$.

基于向量距离的词序相似度算法，将查询和科技文献标题映射到向量空间模型中，提取公共部分作为位置映射，利用位置向量之间的距离计算词序相似度. 由查询短语向量 $\boldsymbol{Q}$ 与科技文献 $\boldsymbol{W}$ 标题的重复词语组成公共向量 $\boldsymbol{C}$，其中 n 为公共向量包含的元素个数. 将查询短语向量 $\boldsymbol{Q}$ 中的公共词语隐射到向量 $\boldsymbol{v}=(v_1,v_2,\cdots,v_n)$，其中 $1\leqslant v_i\leqslant n$，$i=1,\cdots,n$.

对于 3 个连续自然数组成的序列，规定 1、2、3 为标准序列，则向量距离是其他序列到标准序列的距离，计算时利用两组排列中对应位置上的元素之差的绝对值总和. maxDistance 为 $\mathrm{distance}(\boldsymbol{v},\boldsymbol{v}')$ 的最大值. 词序相似度的计算过程的算法效率完全取决于 $\mathrm{distance}(\boldsymbol{v},\boldsymbol{v}')$，向量距离计算的时间复杂度为 $O(n)$，空间复杂度为 $O(1)$.

3　实验与分析

模块中使用基于 N 层向量空间模型的文本相似度算法计算查询短语和科技文献间的相似度，排序后返回给用户查询结果. 将改进后的 N 层向量空间模型与传统向量空间模型的余弦相似

度算法进行了比对,得到如图 1—图 3 的实验结果.

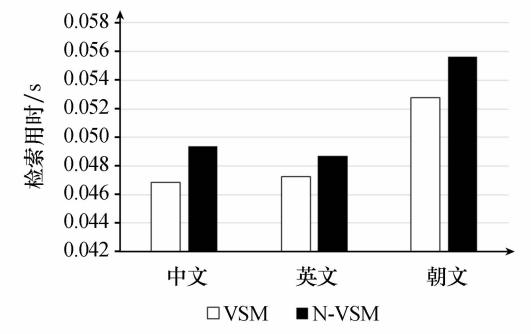

图 1 VSM 与 *N*-VSM 检索的用时对比

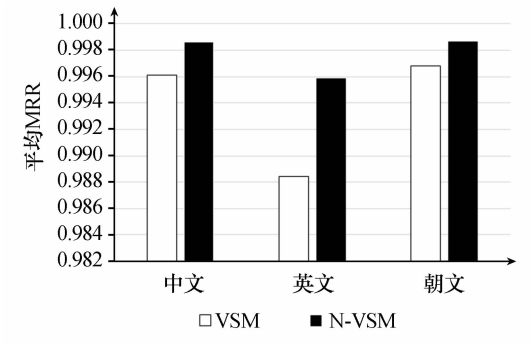

图 2 VSM 与 *N*-VSM 的 MRR 对比

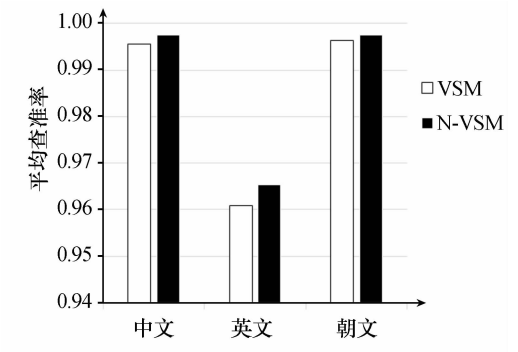

图 3 VSM 与 *N*-VSM 的查准率对比

由图 1—图 3 可知:新算法的检索用时大于传统算法的检索用时,这是因为新算法在以余弦相似度为基础上又计算包含度以及词序相似度,增加了程序的开销.与传统的向量空间模型余弦相似度算法相比较,新算法的查准率平均提高了 2.7%,MRR 平均提高了 2.02%.

4 结论

本文针对向量空间模型不考虑词语位置的缺点,引入了 N 层向量空间模型,将科技文献概要的标题、关键词和摘要赋予不同的比例因子计算权重;针对向量空间模型不考虑词语上下文关系的缺点,引入了关于词序相似度的计算,考虑了词语间的上下文关系对于相似度值的影响.测试结果证明,本文提出的算法比传统的向量空间余弦相似度算法在查准率和 MRR 上分别提升了 2.7%和2.02%,因此本文提出的算法更为有效.本文仅计算了查询短语与科技文献标题间的词序相似度,下一步我们将抽取科技文献的标题、关键词以及摘要中的主题信息构成能够代表科技文献的短句,计算查询短语与短句之间的相似度值,以进一步完善基于 N 层向量空间模型的文本相似度算法.

参考文献:

[1] 宋余庆,陆琳.基于层次模型的搜索引擎评价研究[J].图书情报研究,2014,1(7):32-39.
[2] 谭静.基于向量空间模型的文本相似度算法研究[D].成都:西南石油大学,2015:11-14.
[3] 操卫平.基于结构化向量空间模型的中文信息检索系统研究与设计[D].北京:北京工业大学,2008:13-15.
[4] 高珊.信息检索中的查询扩展及相关技术研究[D].武汉:华中师范大学,2008:20-21.
[5] 施聪莺,徐朝军,杨晓江.TFIDF 算法研究综述[J].计算机应用,2009(29):167-170.
[6] 梁亮.形式概念分析上概念间的包含度理论研究[D].山西:山西大学,2011:7-9.
[7] 董刊生,方金云.基于向量距离的词序相似度算法[J].中文信息学报,2009,23(3):45-48.