文本聚类中权重计算的对偶性策略*

卜东波, 白 硕, 李国杰

(中国科学院 计算技术研究所,北京 100080)
E-mail: bdb@ncic.ac.cn
http://www.ict.ac.cn

摘要: 在文本聚类/分类处理中,一个重要步骤就是寻找文本的合理表示.在被广泛采用的向量空间模型中,一个文本被表示成一个向量,向量的各维是特征项,而向量空间模型的核心问题就是如何进行特征的抽取和选择.在特征的权重计算中,存在一种对偶性现象.利用迭代的方法来处理和利用这种对偶性,获得了文本的隐含概念.实验结果表明,采用概念空间代替原始词空间来表示文本,能够得到更好的聚类结果.

关 键 词: 文本聚类;向量空间模型;特征抽取;对偶性;隐含概念空间
中图法分类号: TP181 文献标识码: A

文本聚类/分类的目标是将语义相近的文本聚成一堆,最理想的境界自然是能准确地揣测和摹拟人们所理解的语义,把人们认为语义相近的文本聚成一堆.要想进行文本聚类/分类,首要问题就是要对文本进行形式化表示.这种形式化表示应该尽可能多地反映文本所蕴涵的语义信息,同时应该是便于计算的,也就是说,从文本的形式化表示能比较容易地计算出文本所蕴涵的语义信息来.

一个中文文本表现为一个由汉字和标点符号组成的字符串,由字构成词,由词构成短语,进而形成句、段、节、章、篇等结构.但是,直接使用整个字符串作为聚类/分类的原始输入是很不方便的,有必要寻找一种更精练的形式化表示方法.

从文本所蕴涵的信息的角度来看,一个中文文本可以由字、词、短语等语义特征项的频率及其相互之间的顺序来完整表达.如果要表示文本中特征项之间的顺序信息,就必然要使用有向的指针结构,整个文本就变成了一个复杂的图结构,比如树或者网.然而信息检索和文本聚类/分类处理要求定义一种距离函数,以表示文本之间的相似程度.如果使用复杂的图结构表示文本,则很难定义一种合理的距离函数,因为存在这样的问题:怎样的两棵树才能说是很相似?又是什么样的两个网才能说是距离比较小呢?

与使用复杂的网或树结构表示文本相反,向量模型仅仅使用文本中特征项的频率信息,使用一个向量来表示文本.在向量模型中不会遇到上述困难,因为数学中有很多种定义距离的方式可资使用,比如欧式距离、相关系数等.自然,仅仅采用这种频率信息是不能精确反映人们所理解的语义的,不可否认会存在一些特例,其语义是仅仅使用频率所无法精确描述的,然而这种方案却能够很方便地计算和操作,对于信息检索和聚类/分类等应用场合来说,其表达效果还是可以接受的.

G. Salton 提出的 VSM(vector space model)就是使用向量来表示文本的一种模型,并成功应用到 SMART 系统中,是应用最广泛的模型[1,2].向量空间模型实际上是走统计的路线,研究从大规模语料库中发现出来的统计规律,利用文本在一些特征项集合上的分布来近似表达语义.因此,向量空间模型表达效果的优劣直接依赖于特

* 收稿日期: 2001-04-13; 修改日期: 2001-07-13
基金项目: 国家自然科学基金资助项目(69773008)

作者简介: 卜东波(1973－),男,山东微山人,博士,助理研究员,主要研究领域为算法设计与分析,信息检索,生物信息学;白硕(1956－),男,辽宁沈阳人,研究员,博士生导师,主要研究领域为算法设计与分析,计算语言学,信息检索,人工智能;李国杰(1943－),男,湖南邵阳人,研究员,博士生导师,主要研究领域为并行处理,计算机体系结构,人工智能,组合优化,人工神经网,遗传算法.

征项的选择与抽取以及特征项权重的计算.

在聚类操作和特征项的权重的关系中,存在一个基本循环:要想得到好的聚类结果必须首先合理设置权重,而按照 Ward 的观点,合理的特征抽取和权重设置应当使得样本类内方差尽量小,同时类间方差尽量大,这样就要求必须首先知道聚类的结果.也就是说,聚类和特征抽取及权重设置互为因果,两者构成一个循环[3].

本文使用迭代的策略来打破这种循环,并且这种迭代操作是收敛的,特征项的权重最终稳定于一个矩阵的特征向量上.为每个特征项赋予迭代计算出的权重,实际上就得到了文本的隐含概念.和直接采用特征项仅仅反映了文本的表层信息相比,这种隐含概念能够更深刻地反映出文本的深层结构.实验结果表明,和直接在特征项空间中表示文本相比,在概念空间中表示文本能够更好地表达文本的语义信息.

1 权重计算中的对偶性策略

文献[3,4]认为,在整个聚类操作中,存在一个基本循环(basic cycle),即要想聚类必须首先进行特征抽取和设置权重,聚类结果的好坏直接依赖于特征抽取和权重设置的合理与否;而合理的特征抽取和权重设置应当使得样本类内方差尽量小,同时类间方差尽量大,这样就要求必须首先知道聚类的结果.也就是说,聚类和特征抽取及权重设置互为因果,两者构成一个循环.这种循环反映在文本聚类中,就是文本聚类和词聚类之间的循环关系——要想把文本聚类做得好,就要首先知道构成文本的词之间的聚类关系,哪些词语义比较相近,哪些词经常共同出现等;而要想把词聚类做好,又必须首先知道包含这些词的那些文本之间的聚类关系,一些文本涉及同一个话题,抱成一团,那么文本所使用的词也被认为比较相近.

上述的循环关系提示我们在文本和词之间可能存在某种对偶性,这种对偶性在以下的权重计算中表现得更为明显:

假设共有 m 篇文本,使用 n 个词.我们为每个文本和每个词都定义一个权值.文本 f_i 的权值 Wf_i 表示该文本对整个文本集语义的反映程度和概括程度,权值越大的文本越重要,概括程度越强.词 t_j 的权重 Wt_j 表示该词对于整个文本集语义的反映程度,权值越大的词就越重要,其反映整个文本集主题的能力也就越强.

文本和词的重要性之间存在着这样的一种对偶关系:

- 一个重要的文本就是包含许多重要词的文本;
- 一个重要的词就是经常出现在重要文本中的词.

这种对偶关系实际上是对重要性的一个循环定义,无法各自独立地定义文本和词的重要性.如何由这种循环定义来定量地计算出文本和词的重要性呢?文献[6]对超文本进行链接分析的技巧给我们以很大的启发:对付这种循环,迭代方法是一件利器.

开始时赋予 Wf,Wt,随机值,这里我们使用 Wf 表示文本权重向量的单位向量,Wt 表示词权重向量的单位向量,即

$$Wf=(Wf_1,Wf_2,...,Wf_m)',$$
$$Wt=(Wt_1,Wt_2,...,Wt_n)'.$$

然后,进行如下的两步迭代过程:

① 使用当前对词权重的估计值 Wt 来改进对文本权重的估计值 Wf,找出当前比较重要的词,包含这些词的文本就是比较重要的文本,因此相应地增加这些文本的权重.具体来说,每个文本更新后的权重 Wf_i 等于它包含的所有词的词频和词权重乘积的总和,也就是词权重向量 Wt 和词频矩阵第 i 行向量的内积.直观地看,包含重要词较多的文本将获得较高的得分.

② 使用当前对文本权重的估计值 Wf 来改进对词权重的估计值 Wt,找出当前比较重要的文本,经常在这些文本中出现的词就是比较重要的词,因此相应地增加这些词的权重.具体地说,每个词更新后的权重 Wt_j 等于所有含有这个词的文本的权重与词频乘积的总和,也就是文本权重向量和词频矩阵第 j 列向量的内积.直观地看,在那些重要文本中经常出现的词将获得较高的得分.

反复进行以上的两步迭代过程,文本权重向量 Wf 和词权重向量 Wt 将稳定在一个不动点上,这个不动点仅仅和 m 行 n 列的词频矩阵 $A_{m\times n}$ 相关.

上述对权重向量的求解过程可以使用算法语言描述如下:

算法 **1**. 求解权重向量的迭代算法.

使用随机值初始化 Wf, Wt;

Repeat /* 执行迭代过程 */

for i =1 to n do

$$Wt_j = \sum_{i=1}^{m} A_{ij} * Wf_i$$

for j= 1 to m do

$$Wf_i = \sum_{j=1}^{n} A_{ij} * Wt_j$$

求出 Wf 和 Wt 的单位向量,并以其代替 Wf 和 Wt.

Until 向量 Wf 和 Wt 稳定.

我们使用 $Wf^{(0)}$和 $Wt^{(0)}$分别表示向量 Wf 和 Wt 的初始值,使用 $Wf^{(k)}$和 $Wt^{(k)}$分别表示经过 k 次迭代之后得到的改进值,Wf^*和 Wt^*表示最终的稳定值,

图 1 形象地描述了迭代求解的过程.

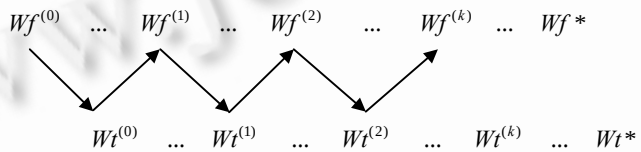

Fig.1 The iteration process of the vector

图 1 向量迭代过程

使用线性代数可以更清楚地分析迭代过程,每次迭代操作实际上是在做向量和矩阵的乘法运算,即

$$Wt^{(k+1)} = (A_{m\times n})^T \times Wf^{(k)},$$

$$Wf^{(k+1)} = A_{m\times n} \times Wt^{(k+1)}.$$

对于任意给定的初始值,这种迭代过程都是收敛的,并且最后的稳定值恰好分别是矩阵 $A*A^T$ 和 A^T*A 的某个特征向量.

对于权重向量的方向,即单位权重向量求解方面,文献[6]中证明了如下定理和推论:

定理 **1**. 矩阵 $A*A^T$ 和矩阵 A^T*A 有相同的非零特征值.

这个看起来不起眼的定理的作用却很重要.在很多应用场合下,我们需要求出 $A*A^T$ 的特征值或者特征向量,但是有时方阵 $A*A^T$ 的维数特别高,而求特征向量过程的时间复杂度是 $O(n^3)$ 的,非常耗时.如果方阵 A^T*A 的维数较低的话,一个变通的方法就是先求出 A^T*A 的特征值或特征向量,然后再依据此定理求出 $A*A^T$ 的特征值或特征向量.这样不仅能够节省大量的时间,更重要的是可以避免大规模运算带来的误差累积,使得结果更加准确.

定理 **2**. 对于任意给定的初始向量 $Wf^{(0)}$ 和 $Wt^{(0)}$,迭代过程都是收敛的. Wf 将稳定于矩阵 $A*A^T$ 的某个特征向量上, Wt 将稳定在 A^T*A 的某个特征向量上.

熟悉线性代数的人马上就可以看出,上述过程就是幂法求矩阵特征值和特征向量的过程.

推论 **1**. Wf 和 Wt 的稳定值 Wf^* 和 Wt^* 满足下面的关系式:

$$Wt^* = \frac{A^T \times Wf^*}{\left\| A^T \times Wf^* \right\|},$$

$$Wf^* = \frac{A \times Wt^*}{\left\| A \times Wt^* \right\|}.$$

上述推论实际上说明了这样一种关系:文本集合在词向量空间中表现成一群点,每个文件在这个空间中的坐标构成矩阵 A 的一个行向量,而 Wt^* 表示词向量空间中的一个方向,Wf^* 表示的则是这些文件在这个方向上的投影.

Wf^* 是文本权重向量,它的各个分量表示相应的文本对整个文本集合语义的概括程度,权重越大的文本越重要,然而,这种重要性只是从某个侧面看的结果,因为 Wf^* 是这些文件在 Wt^* 方向上的投影,它仅仅反映了从 Wt^* 方向上对各个文件重要性的衡量.Dumais 提出的 LSI(latent semantic index)[6,7]技术中将这种方向称为"隐含的概念",这种概念不是仅由某一个词就能完整表达的,而是由一类词共同拥有的语义或者经常共同出现来表达的.因此,Wt^* 只是反映了文本集合中的某一个"隐含的概念",或者说某一个主题,Wf^* 则表示了各个文本对这个主题的贡献的大小,从这个主题来看各个文本的重要与不重要.

在 Clever 系统中使用这种技巧来进行超文本的链接分析.但是和我们在这里的应用不同的是,在 Clever 系统中为每个页面赋予两个权重,分别表示页面内容的权威程度和引用程度,它处理的矩阵仅仅是 m 个节点之间的关联矩阵,是一个 m 阶方阵;而且,矩阵的每一个元素都是 0/1 二值,以表示两个节点之间是否有链接关系[6].

2 概念空间以及特征选择

很少会出现仅仅有一个主题的文本集,通常的文本集都会有多个主题或曰"隐含的概念".比如,随着 Wt 选取不同的初始值,Wt^* 会得到不同的稳定值 $\xi_1,\xi_2,...,\xi_r$. ξ_1 反映了文本集合中的一个概念,ξ_2 则反映了 ξ_1 所不能表达的另一个概念,而 ξ_3 则反映了 ξ_1 和 ξ_2 都不能表达的某个概念…每个 ξ_i 都反映了文本集各不相同的主题.任意两个稳定值 ξ_i 和 ξ_j 都是两两正交的,直观地说,某个 ξ_i 对文本集合主题的反映作用是不能被其他的 ξ_j 所完全代替的.

针对某个特定的主题 ξ_i,可以定义各个文件对这个主题的反映程度,也就是文本的重要程度.对于一个文本,我们使用其在 ξ_i 方向上的投影来定量地刻画该文本对主题 ξ_i 的反映程度,投影为正数的文件可以看作是对这个主题的赞同,投影为负数的文本可以视为对这个主题的否定,而投影的绝对值大的那些文本对反映这个主题的作用也比较大,绝对值小的文本的反映力也较小.所有文本的权重合起来恰好就是与 ξ_i 对应的向量 η_i.

各个"隐含概念" ξ_i 有着不同的重要性,即概念之间也有主次之分.这种重要性可以使用所有文本在概念方向 ξ_i 上投影的方差来定量刻画,方差越大则该概念越重要,反之,方差越小则该概念越不重要.从信息的角度来看,方差的大小表达了概念 ξ_i 蕴涵的信息量的多少.它表示投影的散布情况,散布越大,蕴涵的信息量就越大;散布越小,蕴涵的信息量就越小.

设隐含概念 ξ_i 对应的特征值为 λ_i,即 $(A^T*A)*\xi_i=\lambda_i*\xi_i$,且 ξ_i 为单位向量.对于所有文本在隐含概念方向上的投影向量,即 $A*\xi_i$ 的模长和方差方面,我们证明了如下定理成立:

定理 **3**. $\|A*\xi_i\|=\sqrt{\lambda_i}$.

证明:

$$\begin{aligned}\left(\|A*\xi_i\|\right)^2&=(A*\xi_i)^T*(A*\xi_i)\\&=\xi_i^T*A^T*A*\xi_i\\&=\xi_i^T*(A^T*A)*\xi_i\\&=\xi_i^T*\lambda_i*\xi_i\\&=\lambda_i*\xi_i^T*\xi_i\\&=\lambda_i,\end{aligned}$$

故有 $\|A*\xi_i\|=\sqrt{\lambda_i}$ 成立. □

推论 **2**. $D(A*\xi_i)=\lambda_i*D(\eta_i)$.

如果我们以各个“隐含概念” ξ_i 为坐标轴,一个文本的坐标是其在概念方向上的投影,定义一个新的坐标系来表示所有文本,这个新的空间可以称为概念空间.

图 2 表示了一个概念空间,我们对于包含 15 个文本的文本集,求出各个隐含概念向量,并在隐含概念空间中重新表示各个文本,为了绘图方便起见,我们只使用两个特征向量,只描绘了二维概念空间.

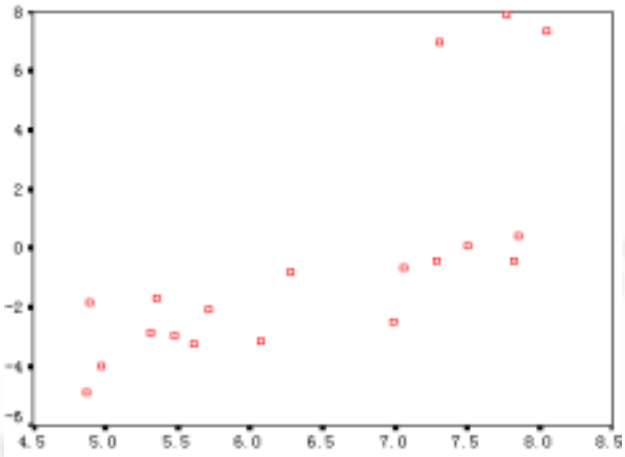

Fig.2 The map of text in the concept space

图 2 文本在概念空间中的投影

在从词空间转化到概念空间的过程中,我们可以忽略一些不重要的概念.那些重要性特别低的概念不是整个文本集合意图的重点,蕴涵的信息量比较小,忽略掉并不会影响大局,因此可以作为噪声过滤掉.这里概念的重要性度量采用文本集合在该概念方向上投影的方差来表示.

我们首先求出 A^*A^T 的所有特征值 λ_i 以及相应的特征向量 η_i (设共有 r 个特征向量),然后按照 $\lambda_i * D(\eta_i)$ 由大到小排列.可以只选择前 k 个重要的方向,使用如下准则进行 k 的选取:求最小的 k 满足

$$\frac{\sum_{i=1}^{k}\lambda_i * D(\eta_i)}{\sum_{i=1}^{r}\lambda_i * D(\eta_i)} \geq t.$$

其中 t 是一个预先设定的阈值,表示信息损失的多少,一般取 $t = 0.80 \sim 0.90$.也就是说,忽略掉一些重要性特别低的概念会造成信息的损失,如果损失不超过 0.10~0.20 这个限度,我们则认为是可以接受的.

使用概念空间代替原始词空间有如下几个好处:

(1) 概念空间的各维是正交的,这和直观上是一致的,而各个词之间大量地存在着线性相关关系,词空间不是一个正交空间.另外,由于概念空间是一个正交空间,因此可以使用欧式距离来定义各个样本之间的远近关系.

(2) 使用深层的概念而不是仅仅使用表象的词,能够更深入地描述文本之间的关系,有利于挖掘文本集的深层结构.

(3) 能够过滤噪声.在概念空间的某些维上,所有文本的表现大致相同,差别很小以至于可以作为噪声被忽略掉.

(4) 可以降维.原始的词频矩阵是一个 m 行 n 列的矩阵,而我们只选取了前 k 个主要概念构成概念空间,变换后的矩阵是 m 行 k 列的.在我们的实验中,取信息损失上限为 0.15,常常能够将数千维的原始词空间降低到数十维,从而使后续处理步骤大为简化.

3 概念空间在文本聚类中的应用

如上所述,文本聚类/分类的目标就是将语义相近的文本聚成一堆,最理想的境界自然是能准确揣测和摹拟

人们所理解的语义,把先验知识规定的同类文本聚成一堆.先验知识把文本分成几类,对这个文本集合聚类的结果就应当是几类.然而在实际计算中很少达到这么精确的结果,常常出现的情形是把先验知识规定的类拆分成一些子类,这些子类都是聚类操作得到的相互之间最相似的一团文本.因此,子类数目的多少就能表示聚类结果与先验知识的协调程度.特征项选择得越好,对先验知识摹拟得越准确,子类数越少;反之,就会把先验知识规定的类拆分得很碎,子类数目就越多.

我们对 5 个样本集合进行实验.对于每一个测试文本集,首先求出原始词频矩阵,使用上述迭代过程获取各项的权重,然后忽略掉那些不重要的概念,求出文本集在概念空间中的表示,最后在概念空间中进行聚类操作,得到的结果见表 1.

Table 1 Results of clustering in the concept space

表 **1** 在概念空间中的聚类结果

Test set①	Number of documents②	Number of categories③	Term space④		Concept space⑤	
			Number of dimensions⑥	Number of categories⑦	Number of dimensions	Number of categories
Test 1	19	4	1 231	5	7	5
Test 2	40	4	3 094	4	4	4
Test 3	66	4	2 703	7	5	4
Test 4	214	12	3 728	29	42	18
Test 5	403	5	2 890	39	41	33

①测试集,②文档数目,③先验知识规定的类数,④原始项空间,⑤概念空间,⑥维数,⑦子类数目.

从表 1 中可以看出,采用迭代操作得到的概念不仅能够大幅度地降低维数,而且能够减小选取特征和先验知识之间的不协调性,更好地表示和摹拟人们所理解的语义,换句话说,在求得的概念空间中进行聚类,聚类结果更贴近先验知识,即求得的概念能够更好地表示先验知识.

4 结束语

无论是对于文本聚类/分类,还是针对文本的信息检索,特征项的选取都是一个基础性的工作.特征选取的优劣将直接决定最终结果的好坏.从实验结果中可以看出,相对于原始的词空间而言,使用迭代加权过程挖掘出的文本集蕴涵的概念,能够更加有效地反映出文本集的主题,进而有助于文本的聚类/分类和文摘.

References:

[1] Salton, G. Automatic Text Processing. Addison-Wesley Publishing Company, 1988.

[2] Huang, Xuan-jing. Research on retrieval, classification and summarization for large scale text [Ph.D. Thesis]. Shanghai: Fudan University, 1998 (in Chinese).

[3] Fang, Kai-tai, Pan, En-pei. Clustering Analysis. Beijing: Geography Press, 1982 (in Chinese).

[4] Hartigan, J.A. Clustering Algorithms, Yale University, John Wiley&Sons, New York, London, 1975.

[5] Kleinberg, J. Authoritative sources in a hyperlinked environment, In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms. 1998. http://www.cs.cornell.edu/home/kleinber/.

[6] Dumais, S.T. LSI meets TREC: a status report. In: Harman, D., ed. Proceedings of the 1st Text Retrieval Conference (TREC1). National Institute of Standards and Technology, 1993. 137~152.

[7] Dumais, S.T. Latent semantic indexing (LSI) and TREC-2. In: Harman, D., ed. Proceedings of the 2nd Text Retrieval Conference (TREC2). National Institute of Standards and Technology, 1994. 105~116.

附中文参考文献:

[2] 黄萱菁.大规模中文文本的检索、分类与摘要研究[博士学位论文].上海:复旦大学,1998.

[3] 方开泰,潘恩沛.聚类分析.北京:地质出版社,1982.

The Duplex Strategy of Term Weighting in Text Clustering*

BU Dong-bo, BAI Shuo, LI Guo-jie

(*Institute of Computing Technology*, *The Chinese Academy of Sciences*, *Beijing* 100080, *China*)

E-mail: bdb@ncic.ac.cn

http://www.ict.ac.cn

Abstract: An important step in text mining is to find a reasonable representation of the text. In the popular VSM (vector space module), where a text is represented as a vector, the coral problem is to term extraction, selection and weighting. An iteration method is proposed to deal with the duplex phenomena found in term weighting and compute out the latent concept. Experimental results show that the latent concept could help to get better clustering results.

Key words: text clustering; vector space module; term extraction; duplex; latent concept

* Received April 13, 2001; accepted July 13, 2001

Supported by the National Natural Science Foundation of China under Grant No.69773008