第35卷 第13期 计 算 机 工 程 2009年7月
Vol.35 No.13 Computer Engineering July 2009
·软件技术与数据库· 文章编号：1000—3428(2009)13—0026—02 文献标识码：A 中图分类号：TP391

基于非负矩阵分解的中文文本主题分类

张 磊，冯晓森，项学智
(哈尔滨工程大学信息与通信工程学院，哈尔滨 150001)

摘 要：提出基于非负矩阵分解(NMF)的中文文本主题分类方法，应用NMF算法分解词-文本矩阵获取词之间的相关性，有效地解决同义词、多义词的影响。实验结果表明，与基于奇异值分解的潜在语义索引方法相比，该方法计算速度快、占用存储空间较少。在潜在语义数据降低较大的情况下，NMF方法具有更好的分类精度。

关键词：主题分类；非负矩阵分解；潜在语义索引

Topic Classification of Chinese Document Based on NMF

ZHANG Lei, FENG Xiao-sen, XIANG Xue-zhi
(Information and Communication Engineering College, Harbin Engineering University, Harbin 150001)

【Abstract】This paper presents a method based on Non-negative Matrix Factorization(NMF) for Chinese document topic classification. According to NMF, the term-document matrix is decomposed to reveal the relationship between terms. This method solves the problem of synonym and polysemy effectively. Compared with Latent Semantic Indexing(LSI) based on Singular Value Decomposition(SVD), experimental results show that this method has faster computing speed and less memory occupancy. It can improve classification precision when the number of latent semantic index is reduced pronouncedly.

【Key words】topic classification; Non-negative Matrix Factorization(NMF); Latent Semantic Indexing(LSI)

1 概述

文本分类指在事先给定的主题下，根据文本内容自动确定文本类别的过程。向量空间模型[1](Vector Space Model, VSM)是最常用的文本表示方法，但其存在2个问题：(1)VSM仅考虑词频信息，缺乏语义层面的考虑；(2)向量维数过大，造成存储和资源消耗过大。因此，基于语义层面的文本表示方法成为目前文本处理领域研究的热点。目前将词或短语的特征项空间映射成语义空间的主流方法是潜在语义索引(Latent Semantic Indexing, LSI)[2]方法。

LSI采用奇异值分解(Singular Value Decomposition, SVD)作为矩阵分解的方法，将高维的特征项空间映射到语义空间。非负矩阵分解(Non-negative Matrix Factorization, NMF)[3-4]是一种新的矩阵分解方法，本文将NMF应用于文本分类中，将词汇空间映射到语义空间，在解决语义问题的同时达到更好的快速降维效果，并针对不同的潜在语义数目，对比NMF和SVD分解方法的优缺点。

2 文本分类系统框架

文本分类的系统结构如图1所示，整体上可以分为训练过程和分类过程。

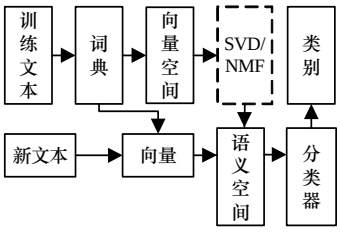

图1 文本分类系统框架

在训练阶段对文本进行预处理得到文本的表征词典，根据该词典，将文本投影到向量空间中，在此空间基础上，利用相应的分解算法将其映射到语义空间，进而构造分类器。在分类过程中，把表示新文本的特征向量同样投影到语义空间后，通过分类器进行分类。

2.1 词典生成

由于中文的词与词之间没有明显的分割符号，因此须对文本进行分词处理。在分词处理后，得到一个词表。为了改善文本表示的质量，缩小表征词典的规模，提高文本分类器的训练和分类效率，要对词表进行“去停用词”和过滤等处理，获得文本的表征词典。

2.2 向量空间模型

为便于计算机的处理，文本必须表示为计算机可以识别的格式，向量空间模型是常用的有效方法之一。目前被广泛采用的VSM权重计算公式是TFIDF方法。该方法均衡考虑词在相应主题的出现次数和在其他类别主题中的出现次数，具有较好的区分特性，计算公式为

$$TFIDF(t_i,d)=TF(t_i,d)\times IDF(t_i) \tag{1}$$

其中，$TF(t_i,d)$表示词t_i在文档d中的出现次数；$IDF(t_i)$为反文档频率，计算公式为

$$IDF(t_i)=\lg(\frac{N}{DF(t_i)}) \tag{2}$$

其中，N表示总文档数；$DF(t_i)$为含有词t_i的文档数。

基金项目：国家自然科学基金资助项目“基于Lattice的汉语语音主题分类方法研究”(60702053)；国家自然科学基金资助项目“基于子词网格的汉语语音检索关键技术研究”(60575030)

作者简介：张 磊(1973－)，女，副教授、博士，主研方向：语音信号处理，自然语言处理；冯晓森，硕士；项学智，讲师、博士

收稿日期：2009-01-14 **E-mail：**zhanglei@hrbeu.edu.cn

3 语义空间映射

基于上述方法得到的词-文本矩阵是一个高维的稀疏矩阵，对于高维矩阵的存储、处理需要耗费大量的机器资源。另外，稀疏矩阵在匹配过程易受噪声影响。在向量空间中，很难消除多义词和同义词的影响。因此，须将高维稀疏的向量空间矩阵向低维紧凑的语义空间映射。

本文分别利用奇异值分解和非负矩阵分解将词向量空间映射到语义空间，寻求文本中的潜在语义结构，并用这种潜在语义表征文本很好地解决了上述问题。

3.1 奇异值分解

在潜在语义索引方法中，利用 SVD 的方法将文档和词汇对应关系映射到一个低维空间，称为潜在语义空间。其中 SVD 将词-文本矩阵 $\boldsymbol{X}_{m\times n}$ 分解为 3 个矩阵的乘积形式，即：

$$\boldsymbol{X}\approx\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \tag{3}$$

其中，$\boldsymbol{U}$ 是由左奇异值向量 $\boldsymbol{u}_i$ $(1\leqslant i\leqslant m)$ 组成的 $m\times r$ 矩阵；$\boldsymbol{S}$ 为由奇异值构成的 $r\times r$ 的对角阵；$\boldsymbol{V}$ 是由右奇异值向量 $\boldsymbol{v}_j$ $(1\leqslant j\leqslant n)$ 组成的 $n\times r$ 矩阵。通过 SVD，将文本向量映射到由矩阵 $\boldsymbol{U}$ 的前 r 列张成的低维空间中。其中，r 表示潜在语义的数目，r 远小于 m 和 n。这样，每个文档在 r 维空间上的坐标就由矩阵 $\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$ 的列向量给出。

在分类时，测试文本 $\boldsymbol{q}$ 在低维空间上的投影可表示为

$$\boldsymbol{q}^{'}=\boldsymbol{q}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{S}^{-1} \tag{4}$$

基于 SVD 的 LSI，在一定程度上解决了同义词和多义词的影响，并且有效解决了维数过大和数据稀疏等问题。

3.2 非负矩阵分解

NMF 是近年来一种新的基于语义的矩阵分解算法，它将 1 个非负的矩阵分解为左右 2 个非负矩阵的乘积。由于分解前后的矩阵中仅包含非负的元素，因此原矩阵中的一列向量可以解释为对左矩阵中所有列向量(称基向量)的加权和，而权重系数为右矩阵中对应列向量中的元素。这种基于基向量组合的表示形式具有很直观的语义解释，它反映了人类思维中“局部构成整体”的概念。另外，基于简单迭代计算的 NMF 方法具有收敛速度快、左右非负矩阵存储空间小、语义解释性强的特点，因此，适用于处理大规模文本。

在文本分析中，单词的词频统计总是非负的。因此，可以把 NMF 应用于文本主题分类，但目前对于此方面的研究，尤其是中文文本主题分类的研究还很少。

3.2.1 NMF 的基本原理

$\boldsymbol{X}_{m\times n}=[x_1,x_2,\cdots,x_n]$ 为任意给定的一个非负矩阵，NMF 算法就是寻找 2 个非负矩阵 $\boldsymbol{B}_{m\times r}$ 和 $\boldsymbol{H}_{r\times n}$，使 $\boldsymbol{X}$ 可以近似分解成这 2 个矩阵的乘积。即：

$$\boldsymbol{X}\approx\boldsymbol{B}\boldsymbol{H} \tag{5}$$

一般 r 满足 $(n+m)r<nm$。式(5)可写为

$$\boldsymbol{x}_j\approx h_{1j}b_1+h_{2j}b_2+\cdots+h_{rj}b_r \tag{6}$$

每个文本向量 $\boldsymbol{x}_j$ 可以表示成矩阵 $\boldsymbol{B}$ 中的列向量与向量 $\boldsymbol{h}_j$ 分量的近似线性相加。因此，$\boldsymbol{B}$ 中的列向量可以看成是一组基向量，这组基向量构成了一个 r 维空间。测试文本 $\boldsymbol{q}$ 和词-文本矩阵在新的语义空间上的投影可以表示为 $\boldsymbol{q}^{'}$ 和 $\boldsymbol{X}^{'}$，公式如下：

$$\boldsymbol{q}^{'}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{q},\quad \boldsymbol{X}^{'}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X} \tag{7}$$

3.2.2 NMF 算法实现

NMF 是个 NP 问题，可以用迭代方法交替求解 $\boldsymbol{B}$ 和 $\boldsymbol{H}$。判断迭代收敛性的目标函数一般有欧氏距离和 KL 离散度。这里采用 KL 离散度作为目标函数，定义如下：

$$D(\boldsymbol{X}\,\|\,\boldsymbol{B}\boldsymbol{H})=\sum_i\sum_j[x_{ij}\lg(x_{ij}/y_{ij})-x_{ij}+y_{ij}] \tag{8}$$

其中，y_{ij} 为矩阵 $\boldsymbol{Y}=\boldsymbol{B}\boldsymbol{H}$ 的元素。这样，NMF 求解问题，演变为求非负矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{H}$，使上式的离散度达到最小。

按式(9)、式(10)的规则不断迭代，可以确保 $\boldsymbol{B}$ 和 $\boldsymbol{H}$ 收敛到一个局部最优解[5]：

$$h_{au}\leftarrow h_{au}\frac{\sum_{i=1}^{m}b_{ai}x_{iu}/y_{iu}}{\sum_{k=1}^{m}b_{ka}} \tag{9}$$

$$b_{ia}\leftarrow b_{ia}\frac{\sum_{u=1}^{n}h_{au}x_{iu}/y_{iu}}{\sum_{v=1}^{n}h_{av}} \tag{10}$$

4 分类器设计

利用 SVD 和 NMF 特征抽取方法完成降维之后，训练集和测试集中的文本均可在语义空间中用一个低维的向量来表示。如果对于相同主题的文本进一步聚类，生成相应的聚类中心，这样的分类过程可以看做是测试文本向量和不同聚类中心向量之间的距离求解问题。这就是 Rocchio 分类方法的基本思想。在聚类中心的选取上，Rocchio 方法按照算术平均为每类文本集生成一个代表该类的中心向量，然后在新文本来到时，确定新文本向量，计算该向量与每类中心向量建的距离(相似度)，最后判定文本属于文本距离最近的类。

计算新文本特征向量和每类中心向量间的相似度，用向量间的夹角的余弦来度量，公式如下：

$$Sim(\boldsymbol{d}_i,\boldsymbol{d}_j)=\cos(\boldsymbol{d}_i,\boldsymbol{d}_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}} \tag{11}$$

其中，$\boldsymbol{d}_i$ 为新文本的特征向量；$\boldsymbol{d}_j$ 为第 j 类的中心向量；M 为特征向量的维数。由上式可知余弦越大，相似度越大。

5 实验及分析

本文对基于 NMF 和 SVD 降维的文本分类方法进行对比实验。在实验中，使用中文自然语言处理开放平台提供的文本分类语料库作为实验语料。该语料库分为测试集和训练集，分属 20 个类别。本文选取环境、教育、经济、体育和政治 5 个主题。语料库经过 NoteTab Light 分词软件分词，并去掉停用词。采用 NMF+Rocchio 和 SVD+Rocchio 在潜在语义数目 r 分别为 10 和 20 的情况下作对比实验，将结果进行比较，如表 1、表 2 所示。

表 1　r=10 时的实验测试结果　(%)

主题	NMF			SVD		
	准确率	召回率	$F1$	准确率	召回率	$F1$
环境	100.0	90.0	94.7	100.0	70.0	82.4
教育	100.0	90.0	94.7	100.0	70.0	82.4
经济	71.4	100.0	83.3	58.8	100.0	74.1
体育	100.0	80.0	88.9	100.0	90.0	94.7
政治	100.0	100.0	100.0	100.0	100.0	100.0
平均	94.3	92.0	92.3	91.8	86.0	86.7

在 VSM 模型中，采用 TFIDF 作为权值计量方法。评价的方法选择传统的准确率(Precision)、召回率(Recall)和 $F1$ 值的方法。一般来说，计算代价及文本分类的准确率与 r 成正比。NMF 的 $F1$ 值在表 1 和表 2 中分别为 92.3%和 98.0%。

(下转第 54 页)