一种基于 NMF_{SC} 的文本聚类方法①

王永贵，高　月

(辽宁工程技术大学 软件学院，葫芦岛 125105)

摘　要：通过分析文本的特征，提出了一种基于稀疏约束非负矩阵分解（NMF_{SC}）的文本聚类新方法。该方法用 NMFSC 分解词-文本矩阵来降低特征空间的维度，并依照稀疏约束更好地控制稀疏度，然后利用簇中文本的相似性进一步细化簇。实验表明，与基于 k-means 的文本聚类方法和基于 NMF 的文本聚类方法相比，此方法具有较高的归一化互信息值（NMI），从而具有良好的聚类性能。

关键词：文本聚类；细化簇；非负矩阵分解；稀疏表示；归一化互信息值

Document Clustering Method Based on NMF_{SC}

WANG Yong-Gui, GAO Yue

(College of Software Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Through analyzing the characteristics of the text, a novel text clustering approach based on Non-negative Matrix Factorization with sparseness constraint (NMFSC) is presented. The method uses NMFSC decomposing word-text matrix to reduce the dimension of the feature space, and better controls sparsity with sparseness constraint, and then further refines clusters by using the similarity of documents in clusters. Compared with text clustering method based on k-means and text clustering method based on NMF, the results of experiment show that the method has high value of the normalized mutual information, thus it has good clustering performance.

Key words: text clustering; refine clusters; non-negative matrix factorization; sparse representation; normalized mutual information

1　引言

随着互联网络的发展，各种电子文本资源每天都在以惊人的速度迅速增长。对于这种半结构或无结构化数据，如何从中获取特定内容的信息和知识成为摆在人们面前的一道难题。文本聚类正是解决这一难题的重要方法之一。它主要目的是在语义空间中将文本按照主题划分为不同的类，使同类文本之间具有最大的相关性, 而不同类文本之间具有最大的相异性。因此文本聚类作为一种对大规模文本信息进行有效地组织、导航、检索和概括汇总的关键技术而日益受到关注，大体上，文本聚类方法可分为划分聚类算法和层次聚类算法[1]。近年来，随着文本聚类越来越得到人们的重视，一些新的文本聚类方法也逐渐呈现出来，如基于图的方法、基于机器学习的方法和基于矩阵分解的方法等。其中，基于图的方法[2]是用一个无向图来表示给定的文本集合，其中每个顶点代表一个文本；基于机器学习的方法[3]是利用簇成员的先验知识来开发一个新的半监督文本聚类模型；而基于矩阵分解的方法[4]则是用文本集合中的语义特征来进行文本聚类。

非负矩阵分解（Non-negative Matrix Factorization, NMF）[5,6]是近年来一种新的基于语义的矩阵分解算法，它将原始矩阵分解成左右两个非负矩阵的稀疏分布表示，且分解前后的矩阵中仅包含非负的元素，因此原始矩阵中的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

① 收稿时间:2010-12-12;收到修改稿时间:2011-04-10

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。但由于 NMF 的稀疏能力比较弱，并难以控制稀疏表示的稀疏程度，所以本文利用了带有稀疏约束的非负矩阵分解（NMFSC）。可是它仍不能保证每次都能从任意数据对象集合中成功地分解出语义特征。所以我们在基于 NMFSC 的文本聚类结果基础上，利用簇中文本的相似性进一步细化簇。这样既可以通过 NMFSC 确定关于簇的主要主题和次要主题，又能使用簇中文本的相似性移除簇中的不相似文档。

2 文本表示

设 D 为给定的文档集合，即 $D=(d_1, d_2, \cdots, d_n)$，其中 d_j 为文档集合中的第 j 篇文档，且 $1<=j<=n$。文档表示为：$d_j=(t_1, t_2, \cdots, t_m)$，其中 t_i 是特征项，$1<=i<=m$。对于某篇文档而言，t_i 是经过文本集预处理后得到的能代表文档内容的基本单位，一般选择其作为文本的特征。通常每个特征项会被赋予一定的权重，即 $dj=(t_1, w_1; t_2, w_2; \cdots, t_m, w_m)$，简记为：$dj=(w_1, w_2, \cdots, w_m)$，其中 w_m 是 t_m 的权重。然后用矩阵模型化文档集合，在这个矩阵中，每个文本占矩阵的一列，每列的维数大小由文本集合的特征数决定，这样文档集合 D 被表示成 m×n 的词-文本矩阵 V，其中 V 中的每一个元素 V_{ij} 表示特征项 t_i 在文本 d_j 上的权重，即

$$v_{ij}=w_{ij}=tfidf_{i,j}=tf_{ij}\times\log\frac{N}{n_i},\quad i=1,\ldots,m;\ j=1,\ldots,n \tag{1}$$

这里，tf_{ij} 表示特征项 t_i 在文档 d_j 中的频率，N 为文档集合中的总文档数，n_i 为包含特征项 t_i 的文档数量。由此我们可以看出特征项在文档中出现的次数越多，与文档的主题就越相关，并且如果某特征项出现在文档集中的大多数文档中，它的重要程度应该削弱。因此，一个特征项的权重较高在于它的词频较高，同时该特征项在整个文档集上的频率较低。

3 NMF算法

3.1 NMF 算法基本理论

NMF 算法可描述如下：对于一个大小为 m×n 的非负矩阵 V，可以将其分解为一个大小为 m×r 的非负矩阵 W 和一个大小为 r×n 的非负矩阵 H 的乘积，即

$$V\approx WH \tag{2}$$

一般情况下，$r<\min(m,n)$，即可依据 $(m+n)r<mn$ 选择合适的 r 值，这样就可以用 H 代替原矩阵，实现对 V 的降维，从而减少存储空间，所以分解后存储空间大小为 r×(m + n)。

NMF 求解问题实际上是一个最优化问题，可利用迭代方法求解 W 和 H。为判断其迭代收敛性，NMF 常用的目标函数有两种，一种是欧几里得距离（Euclidean distance）：

$$\|\mathbf{V}-\mathbf{WH}\|^2=\sum_{ij}(\mathbf{V}_{ij}-(\mathbf{WH})_{ij})^2 \tag{3}$$

当且仅当 V=WH 时达到最小值 0。另一种是 KL 散度（Kullback-Leibler divergence）：

$$D(\mathbf{V}\,\|\,\mathbf{WH})=\sum_{ij}(\mathbf{V}_{ij}\log\frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}}-\mathbf{V}_{ij}+(\mathbf{WH})_{ij}) \tag{4}$$

它不是一个距离，同公式（2）一样当且仅当 V=WH 时达到最小值 0。因此最优化问题可表述为：在约束 W 和 H 均大于等于零的情况下，最小化上述关于 W 和 H 的两个目标函数。由于这两个目标函数都仅对 W 和 H 中的一个是凸的，但不是同时对两个凸，所以此最优化问题只能得到局部最优解。

基于以上的思想，Lee 和 Seung 提出了一种乘性更新规则[5]，对于式（2）的更新规则为:

$$\mathbf{H}_{\mu j}\leftarrow\mathbf{H}_{\mu j}\frac{(\mathbf{W}^{\mathrm{T}}\mathbf{V})_{\mu j}}{(\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H})_{\mu j}},\qquad \mathbf{W}_{i\mu}\leftarrow\mathbf{W}_{i\mu}\frac{(\mathbf{V}\mathbf{H}^{\mathrm{T}})_{i\mu}}{(\mathbf{W}\mathbf{H}\mathbf{H}^{\mathrm{T}})_{i\mu}}. \tag{5}$$

对于式（4）的更新规则为:

$$\mathbf{H}_{\mu j}\leftarrow\mathbf{H}_{\mu j}\frac{\sum_i\mathbf{W}_{i\mu}\mathbf{V}_{ij}/(\mathbf{WH})_{ij}}{\sum_k\mathbf{W}_{k\mu}},\qquad \mathbf{W}_{i\mu}\leftarrow\mathbf{W}_{i\mu}\frac{\sum_j\mathbf{H}_{\mu j}\mathbf{V}_{ij}/(\mathbf{WH})_{ij}}{\sum_v\mathbf{W}_{\mu v}}. \tag{6}$$

3.2 稀疏约束 NMF 算法

NMF 算法可以产生数据的稀疏表示，这样分解后的结果更易于理解，但 NMF 的稀疏能力和程度还是比较弱和难以掌握，因此，本文使用了 Hoyer 在 2004 年提出的基于稀疏约束的 NMF(NMFSC)方法[7]。NMFSC

是一种基于 NMF 的矩阵分解算法，在对原始矩阵 V 进行矩阵分解的同时，对特征矩阵 W 或编码矩阵 H 进行稀疏度的控制。和基本的 NMF 相比，该算法能够更好地发现稳定直观的局部特征，可以自由地控制分解后矩阵的稀疏度，并且具有求解收敛速度快，特征矩阵和系数矩阵的相关性小等特点。

首先给出基于 L1 和 L2 范数的稀疏因子，具体定义如下：

$$\text{sparseness}(x)=\frac{\sqrt{n}-(\sum|x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n}-1} \tag{7}$$

其中，n 表示向量 x 的维数。当向量 x 只包含一个非零向量，则稀疏因子的值为 1；当所有向量元素相等时，稀疏因子的值为 0。因此稀疏因子的取值范围在 0 和 1 之间。

接下来将带有稀疏约束的 NMF(NMF_{SC})定义如下：

假定一个大小为 m×n 的非负矩阵 V，将其分解为两个大小分别为 m×r 和 r×n 的非负矩阵 W 和非负矩阵 H 的乘积，也就是求

$$E(\mathbf{W},\mathbf{H})=\|\mathbf{V}-\mathbf{WH}\|^2 \tag{8}$$

最小，同时满足以下两个约束条件

$$\text{sparseness}(w_i)=s_W,\forall i \tag{9}$$

$$\text{sparseness}(h_i)=s_H,\forall i \tag{10}$$

其中，w_i 为 W 的第 i 列，h_i 为 H 的第 i 行。 r 为提取的主题数，S_W 和 S_H 则分别为矩阵 W 和 H 的稀疏因子，这 3 个参数都是由用户设定的。

这样，在 NMF_{SC} 的迭代过程中我们就可以将以上定义的稀疏因子加入到目标函数约束中，过程为：

1) 如果对 W 的列进行稀疏约束，则

i. $W:=W-s_w(WH-V)H^T$; (11)

ii.那么根据非负稀疏投影算法将 W 每一列的元素变为非负值，并保持其 L2 范数不变，但在 L1 范数上使 w_i 达到要求的稀疏度。

2) 如果没有对 W 的列进行稀疏约束，则

$$\mathbf{W}:=\mathbf{W}\otimes(\mathbf{VH}^T)\varnothing(\mathbf{WHH}^T) \tag{12}$$

3) 如果对 H 的行进行稀疏约束，则

i. $H:H-s_HW^T(WH-V)$; (13)

ii.那么根据非负稀疏投影算法将 H 的每一行的元素变为非负值，并将其 L2 范数单位化，但在 L1 范数上使 hi 达到要求的稀疏度。

4) 如果没有对 H 的行进行稀疏约束，则

$$\mathbf{H}:=\mathbf{H}\otimes(\mathbf{W}^T\mathbf{V})\varnothing(\mathbf{W}^T\mathbf{WH}) \tag{14}$$

迭代上述过程，直到收敛到满足稀疏度的要求为止。

4 基于NMF_{SC}文本聚类新方法

4.1 基于 NMF_{SC} 的文本聚类方法

在文本聚类阶段，我们使用 NMFSC 对词-文本矩阵进行分解[8]，同时适当地控制稀疏因子，从而使 W 更加突出主题，H 聚类更加容易。然后用分解的得到的 H 矩阵来确定每个文本的聚类标签，如果

$$x=\arg\max_{\mu} H_{\mu j} \tag{15}$$

则将矩阵 V 中的第 j 列代表的文本划分到簇 x 中，以此类推，直至 n 个文本均被划分为止。

4.2 改进

由于基于 NMFSC 的文本聚类方法不能保证每次都能从任意数据对象集合中成功地分解出语义特征，因此本文引入了簇细化的概念[9]。簇细化指的是将簇细化成簇内各个对象高度相关的簇。其具体改进步骤如下：

1) 确定阈值

计算每个簇中不同文本之间平均相似度，这些平均值作为各自簇的阈值 t(k)。t(k)表示簇中文本之间的相似程度。

$$t(k)=\frac{\sum_{d_k\in C^k}\left(\sum_{d_j\in C^k-\{d_k\}} sim(d_k,d_j)\right)}{s_k} \tag{16}$$

$$sim(d_k,d_j)=\frac{d_k\times d_j}{|d_k|\times|d_j|}=\frac{\sum_{i=1}^{m} w_{ki}\times w_{ji}}{\sqrt{\sum_{i=1}^{m} w_{ki}^2}\times\sqrt{\sum_{i=1}^{m} w_{ji}^2}} \tag{17}$$

其中，$sim(d_k,d_j)$表示簇 C^k 中文本 d_k 与文本 d_j 之间的相似度，s_k 为簇 C^k 中的文本总数，w*i 表示特征项 ti 在文本 d*上的权重。

2) 选取候选集

计算文本 d_k 与所在簇 C^k 中文本之间的平均相似度 $csim(d_k,C^k)$,选取平均相似度 $csim(d_k,C^k)$小于阈值 t(k)的文本对象生成候选集 E。

$$csim(d_k,C^k)=\frac{\sum_{d_j\in C^k-\{d_k\}} sim(d_k,d_j)}{s_k},\quad d_k\in C^k \tag{18}$$

其中，$sim(d_k,d_j)$表示簇 C^k 中文本 d_k 与文本 d_j 之间的相似度，s_k 为簇 C^k 中的文本总数。

3）重新分配文本对象

从候选集 E 中依次选出文本,计算该文本与其他簇文本的平均相似度 $csim(d_l,C^k)$，将文本划分到 $csim(d_l,C^k)$的值最大的簇中。

$$csim(d_l,C^k)=\frac{\sum_{d_j\in C^k} sim(d_l,d_j)}{s_k},\quad d_l\notin C^k$$

4）重复步骤 2）--3），直至所有簇中的文本重新分配完毕。

4 实验及结果分析

在实验中，采用搜狗实验室的文本分类语料关系库 SogouC。该语料关系库具有 mini 版、精简版和完整版三个版本。本文从完整版中的汽车、财经、IT、健康、体育、军事 6 类中各选取出 20 个文本组成测试文本集。

首先对测试文本集进行预处理，将所有文本标上主题类别标签，并对文本进行分词，根据停用词(stop word)表去掉无意义停用词。统计文本集中每个词出现的词频、文本频率和总文档数，最后选出词频最高的 200 个词作为关键词,根据余弦正规化的 TFIDF 权重公式，计算每个关键词在相应文本中的权重,建立文本集的词-文本矩阵 V。

然后选取适当的稀疏因子，分别让稀疏因子 S_W=0.4，S_H=0.5 和 S_W=0.8，S_H =0.6 的 NMFsc 算法作用于原始数据矩阵 V，将得到的矩阵 W 的每列按从大到小排列，取前 5 个元素的值，对照文本空间的基得到提取的主题。当稀疏因子 S_W=0.4，S_H=0.5 时，准确度为 0.645，时间为 2640s，当稀疏因子 S_W=0.8，S_H =0.6 时，准确度为 0.803，时间为 4075s，由此可以看出，随着稀疏因子的提高，主题矩阵 W 中提取的主题更准确，虽然所需的时间多一些。因此实验使用稀疏因子为 S_W=0.8，S_H =0.6。

最后使用归一化互信息度量(NMI)分析聚类结果[10]，此方法优于纯度(purity)和熵(entropy)的方法。它可以消除聚类结果簇的个数对评价聚类结果的影响，NMI 的值越接近 1，表示聚类效果越好。我们将提出的方法(简称为 RNMFSC 方法)与基于 k-means 的文本聚类方法(简称为 KM 方法)和基于 NMF 的文本聚类方法(简称为 NMF 方法)进行比较。我们是在不同的聚簇数目 k∈[2,6]上进行实验的，在每个聚簇数目上分别进行 50 次实验，用 NMI 的平均值获得最终的性能值。具体评估结果见图 1。

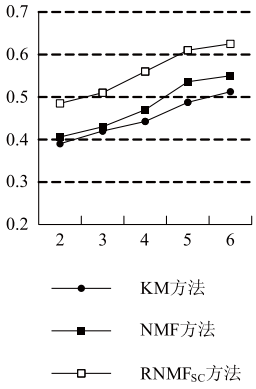

图 1 不同方法的性能比较

在图 1 中，RNMFsc 方法的平均归一化度量的评价结果比 KM 方法提高了 23.76%，比 NMF 方法提高了 16.58%，因此 RNMFsc 显示了最佳的性能。

5 结语

本文提出了一种基于NMFSC的文本聚类新方法。该方法通过引入细化簇的概念来更好地组织涵盖重要主题的文本，并且可以滤除噪声特征项。实验表明，本文提出的文本聚类方法具有较高的 NMI 值，从而与基于k-means的文本聚类方法和基于NMF的文本聚类方法相比，其方法可以提高聚类的性能。

参考文献

1 Han JW.范明译.数据挖掘概念与技术.第 2 版.北京:机械出版社,2007.

2 Hu T, Xiong H, Zhou W, et al. Hypergraph partitioning for

(下转第 156 页)

定”按钮，即可定位到相应的坐标点。

（4）状态栏可显示三部分内容：地图视野、跟踪鼠标位置、系统时间。

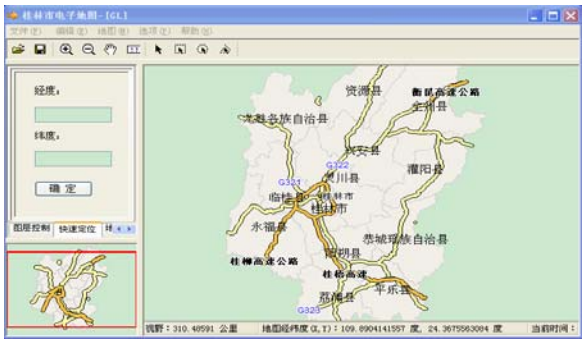

图 2　桂林市电子地图运行主界面

2.3.2 功能界面

（1）点击工具栏中的测距按钮“”，即可打开测量距离窗口，单击主图获取距离测量的起始点，移动鼠标在另一处停下时获取测量距离的终止点，运行结果如图 3 所示：

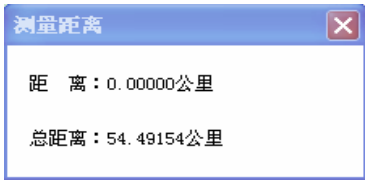

图 3　测量距离窗口

（2）“地图”、“根据坐标画线”，即可出现输入两点经纬度窗口。

3　结语

本文主要根据桂林市规划地图，结合科研项目的需求，设计并实现了桂林市电子地图的雏形。GIS 的许多技术仍然在发展之中，所以不可能对所有相关问题都进行探讨。但是，本课题所研究的桂林市电子地图，能有助于将来的进一步研究，在以后工作中还有很多功能需要实现。相信在将来，通过更加深入地研究后，该电子地图会更加完善与实用。

参考文献

1 吴秀琳,刘永革,王利军.MapInfo9.5 中文版标准教程.北京:清华大学出版社,2009.12–182.

2 齐锐,屈韶琳,阳琳赟.用 MapX 开发地理信息系统.北京:清华大学出版社,2003.5–30.

3 姜拓.基于 C#的 GIS 校园电子地图实现.电脑编程技巧与维护,2009:103–105.

4 陈红斌,杨福兴.基于 C#.NET 的 MapX 开发 GIS 系统.计算机系统应用,2006,15(4):21–55.

5 李佳,曹飞凤,杜光潮.基于 GIS 的钱塘江水质预警预报系统研究.浙江水利科技,2008,(4):65–68.

6 王志宜.大连市排水地理信息系统的设计与开发[硕士学位论文].大连:辽宁师范大学,2008.

(上接第 81 页)

document clustering:a unified clique perspective. Proc. of the 31st Annual International ACM SIGIR Conference. USA: ACM, 2008:871–872.

3 Ji X, Xu W. Document clustering with prior knowledge. Proc. of the 29th annual international ACM SIGIR conference. USA:ACM, 2006:405–412.

4 黄钢石,陆建江,张亚非.基于 NMF 的文本聚类方法.计算机工程,2004,30(11):113–114,176.

5 Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature, 1999,401(6755):788–791.

6 Lee D, Seung H. Algorithms for nonnegative matrix factorization. In Advances in Neural Information Processing, 2001,13:556–562.

7 Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research, 2004,5:1457–1469.

8 杨成福.非负稀疏信号分析理论及在文本聚类中的应用[硕士学位论文].成都:电子科技大学,2006.

9 张猛,王大玲,于戈.一种基于自动阈值发现的文本聚类方法.计算机研究与发展,2004,41(10):1748–1753.

10 Xu W, Gong Y. Document clustering by concept factorization. Proc. of the 27th annual international ACM SIGIR conference. USA:ACM, 2004:202–209.