基于稀疏非负矩阵分解的自动多文摘方法[1]

蒋永锴[1*] 叶东毅[1]

1. 福州大学数学与计算机科学学院，福建省，福州市，邮编 350108

E-mail: *.jiangyk2007@gmail.com

摘 要： 自动多文摘是理解多文档信息的有效方法，是信息处理领域的重要课题。目前的大部分文摘方法不具有明显的潜在语义解释。本文提出一种基于稀疏非负矩阵分解（SNMF）的多文摘方法，通过控制稀疏度，并结合模型选择方法，提高分解得到的潜在语义信息，改进了文档集的话题划分，并能提取主题相关的语句用于文摘生成。相比 LSA 和普通的 NMF 文摘方法，SNMF 方法更具有灵活性，且能强化 NMF 学习局部特征的能力。实验结果表明该方法生成的文摘效果有明显提高。

关键词： 稀疏非负矩阵分解,多文本自动文摘

Multi-Document Summarization Based on Nonnegative Matrix Factorization

Jiang Yongkai, Ye Dongyi

College of Mathematic and Computer Science, Fuzhou University, Fuzhou, 350108, China

E-mail: *.jiangyk2007@gmail.com

Abstract: Multi-document summarization as an important research topic in Information Retrieval is an effective method to understand document corpus. Most of the recent methods do not have a clear semantic explanation. This paper proposes a generic multi-document summarization method based on sparse non-negative matrix factorization （SNMF）. By controlling the sparseness in the factors and combining modeling selection, the proposed method extracts more meaningful latent semantic features and improves topic identification, thus can be used in sentence extraction for summarization. Comparing with other methods such as LSA and NMF, SNMF has the capability to explicitly control sparseness and strengthen local properties of objects. Experiments show an improvement of the quality on multi-document summarization.

Keywords: Sparse NMF， Multi-Document Summarization

1 引言

随着人类步入信息社会，信息呈现出爆炸式增长态势，人们迫切需要一些技术和工具来快速、准确、甚至个性化地从海量信息中获取信息。自动文摘作为一种内容汇总和信息压缩的方法，有助于对海量信息进行信息管理与知识管理，目前已经得到广泛应用。基于句子抽取的摘要方法主要通过TF-ISF（Term Frequency-Inverse Sentence Frequent）、句子或者词的位置、文档关键词等文档特征计算句子的排名，从中抽取排名最好的句子按照一定的规则组织生成文摘[1]。

近年来，基于无监督学习的普通文本摘要方法得到了多个研究者的研究：Nomoto等[2]基于K-means的一种变形提出了文本句子聚类的算法，在句子不同主题类内应用句子权重模型选择文摘语句。Zha等人[3]提出基于主题聚类，并使用显著关键词与语句生成文摘。Gong等人[4]提出了使用Latent Semantic Analysis(LSA)技术从潜在语义上识别关键语句进行文本摘要；国内的林[5]等将LSA算法引入中文文摘领域。尽管LSA是目前很多学者研究的对象，但是因为它得到的特征值

1 福州大学空间数据挖掘与信息共享教育部重点实验室开放基金资助项目（编号：2008-06）

矩阵不是非负、稀疏的，不具有直观的语义解释。所以Park等人[6]将具有稀疏性和非负性的非负矩阵分解（NMF）算法应用于多文本文摘，通过对词-句子矩阵进行NMF分解，提出了基于语句普通相关性的句子提取算法。

然而，现有的NMF文摘方法在计算时采用经典的乘法更新法则，收敛速度慢，并且其因子的稀疏度无法就不同的需求进行不同的控制。随着近年来研究的逐渐完善，NMF算法在分解算法和附加约束条件上取得了很多进展。Li等人[7]提出了LNMF算法，通过在目标函数中加入局部化约束从而更多的保留局部信息，学习到更明显的局部特征信息。Hoyer[8]等人提出基于统计的稀疏度惩罚的分解算法（NNSC），Liu[9]对其进行了改进，两者都强化了NMF学习局部特征的能力。Hoyer提出了基于L1 和L2 范数的稀疏度限制的算法NMFsc[10]，可以灵活任意控制分解因子的稀疏度。Kim[11]等人利用 1-范式对NMF分解的H矩阵添加稀疏度约束，提出了具有较优收敛性的交替最小平方法的稀疏NMF(Sparse NMF)分解算法，且具有模型选择的能力。

本文使用句子抽取的方法，提出了基于稀疏非负矩阵分解（SNMF）的多文本自动文摘方法，对词-句子矩阵进行稀疏非负矩阵分解，通过稀疏度控制，学习更明确的文档集潜在语义主题，并将句子通过聚类划分到各个主题，最后计算句子的排名实现自动文摘提取。该算法还使用模型选择方法自动确定SNMF分解的基矩阵维度。实验证明，它比基于LSA[4]和NMF[6]的文摘方法具有更好的摘要效果。

2. SNMF 文摘方法相关技术

2.1 文本集表示

文摘系统提取摘要的过程可以分为三个阶段：文档分析，主要对原文档进行各种特征提取；转化阶段，通过文摘方法把分析阶段得到的特征进行加权合并，得到文中句子的最终权值；摘要合成，将摘要表示经过句子平滑、加工之后以某种顺序输出，形成具有可读性、较准确、简洁、概括的文摘。在自然语言处理领域，向量空间模型（VSM）是一种常用的文本特征表示模型，可用于对文本建模产生基于词频统计表示的文本矩阵 A，A 一般为非负稀疏矩阵。对文档运用停词表、词干提取等预处理技术后，文档集可以用词-句子矩阵 $A \in R^{m \times n}$ 表示，其中，m 是整个文档集字典 $w = \{f_1, f_2, \cdots, f_m\}$ 的单词总数,n 是所有文档的句子总数，A 可以用(1)向量化表示，其中，t_{ji}、isf_j、n 分别表示字典中第 j 个单词 $f_i \in W$ 在句子 $a_j \in A$ 中的词频统计信息，包含第 j 个单词的句子总数和文档集中句子总数，式子(1)中 a_{ji} 使用经典的 TF-ISF 进行加权。

$$\begin{cases} A = \{a_1, a_2, \cdots, a_n\} \\ a_i = [a_{1i}, a_{2i}, \cdots, a_{mi}]^T \\ a_{ji} = t_{ji} \times \log\left(n / idf_j\right) \end{cases} \tag{1}$$

$$A \approx WH$$

$$\text{其中，} \begin{cases} A \in R^{m \times n}, A_{ij} \geq 0 \\ W \in R^{m \times r}, W_{ia} \geq 0 \\ H \in R^{r \times n}, H_{bj} \geq 0 \end{cases} \tag{2}$$

2.2 SNMF 分解

Lee等研究人员提出了一种非负矩阵分解算法（NMF）[12]，将$m \times n$维的非负数据矩阵A分解为具有潜在语义的基向量空间矩阵W和对应的编码矩阵H的乘积，见公式(2)，其中$r \ll \min(m,n)$是矩阵分解基维度，通常是NMF算法的输入参数。

NMF 分解因子的稀疏性性质只是其分解算法的自然结果，因此[11]基于 1-范式归一化对NMF 分解的 H 矩阵添加稀疏度约束，提出 SNMF 分解算法，SNMF 目标函数为：

$$\min_{W,H} \frac{1}{2}\left[\|A - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \sum_{j=1}^{n} \|H^T(j,:)\|_1^2 \right], \quad s.t. \quad W, H \geq 0 \tag{3}$$

其中，参数$\alpha > 0$控制了 W 元素的大小，稀疏度参数$\beta > 0$，用于控制 SNMF 分解的近似误差和 H 矩阵的稀疏度之间的平衡，稀疏度β值越大H的稀疏度越强，β值越小则 SNMF 分解的误差越小，α、β的选择和实验数据集相关，对于高维矩阵和稀疏文本矩阵，β一般取较小值，如 0.1，本文默认取β=0.2。SNMF 问题可以通过交替最小平方法解决。

非负矩阵分解可以学习物体的局部特征，数据经过NMF分解后，原始数据可以看成是稀疏基矩阵的非负线性叠加的结果，符合人类由局部认识整体的认知方式[12]。使用SNMF对词一句子矩阵进行分解，得到的基矩阵可以看成是文档集主题特征向量的组合，其张成的空间是文本的潜在语义空间；编码矩阵则是文档集中句子在该潜在语义空间的投影。基矩阵向量不再像LSA分解结果那样具有正交约束，使得主题句之间可以有语义相关性。SNMF增加了对编码矩阵的稀疏度控制，有助于得到更稀疏、准确的语义编码。对词一句子矩阵经过SNMF进行分解后，编码矩阵的元素可作为每个句子与各个文档主题的相关度，其值越大则说明相应句子越能表达文档集主题思想。用于提取多文档文摘时，通过计算语句相关度(Generic Relevance of a Sentence, GRS)对句子进行排名,选取排名最优的句子作为候选文摘句[6]。

2.3 模型选择

在将NMF应用于文摘之时，需要手动选择一个适合的基矩阵维度作为参数，因此不利于自动文摘生成。加入稀疏度控制的SNMF算法可以看作一种与K-Means等效的聚类算法，可以用模型选择(Model Selection)的方法自动确定数据集中包含的类数目[11]，通过分散系数(Dispersion Coefficient) ρ_k (k为类数)衡量聚类稳定性：

$$\rho_k = \frac{1}{n^2}\sum_{i}^{n}\sum_{j}^{n} 4\left(\widehat{C_k}(i,j) - \frac{1}{2}\right)^2 \tag{4}$$

ρ_k的取值范围为[0,1]，$\rho_k = 1$表示聚类算法的类分配具有完全的一致性。其中，$\hat{C}_k$是聚类的平均关联矩阵，经过多次实验取关联矩阵$C_k \in R^{n \times n}$的平均值。SNMF算法通过计算不同的聚类数目k对应的ρ_k，可以选择ρ_k值下降时对应的那个k值确定数据集中语义分类的数目[11]。

3. 基于 SNMF 的多文摘方法

本小节介绍使用 SNMF 方法抽取多文本摘要句子的方法。在预处理阶段，对文档集进行预处理，得到基于词频统计的词-句子矩阵 A 。进行摘要提取前，首先利用 SNMF 算法对 A 进行

多次分解计算分散系数 ρ_k，选择 ρ_k 下降时的最优值对应的 k 作为 SNMF 分解的基空间维度 r；接着根据设定的参数进行 SNMF 非负矩阵分解。在摘要提取阶段，通过计算句子相关度，选择相关度最大的句子作为候选文摘句子。具体流程如下，见图 1：

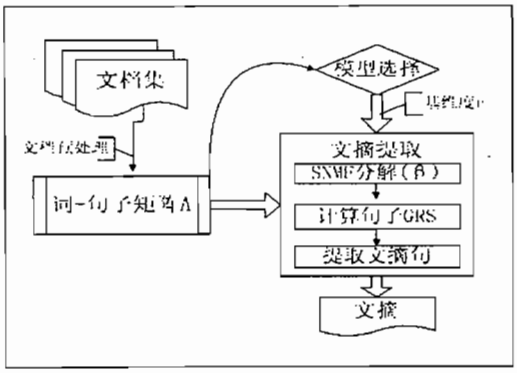

图 1　基于 SNMF 的自动文摘流程图

算法 1: 基于 SNMF 的多文摘算法

输入：TF-ISF 加权的词-句子矩阵 A ,文摘长度 L

输出：包含最大相关度的句子的多文本文摘.

第一步：多文档预处理，利用 VSM 模型构建词-句子矩阵 A，同时统计整个文档集的单词数 D 和句子总数 S。设定文摘长度 L，本文取 L=250;

第二步：应用模型选择确定分解维度 r

使用 SNMF 算法，针对不同维度参数 r 的值 k，利用公式（4）计算分散系数 ρ_k，并选择 k 作为文档的潜在语义空间维度 r;

第三步：对词-句子矩阵 A 进行 SNMF 分解

设定参数，通过对 A 进行 SNMF 分解，得到非负语义变量矩阵 H；

第四步：计算 GRS 相关度并提取文摘

根据句子平均长度确定文摘句数 $M = L/\frac{D}{S}$，从文档集中提取相关度最大的 M 个句子作为文摘句输出。

4　实验与分析

本实验采用DUC测评会议提供的DUC2005 数据集[13]作为测试文档。DUC2005 包含 50 个话题的文档数据集，每个话题包含 25 到 50 个文档。我们从 50 个话题的文档抽取了 45 个进行了多文本文摘实验，并使用ROUGE工具包对文摘质量进行评价，采用Rouge-1 的召回率(Rouge-1 R)、准确率(Rouge-1 P)、F-测度(Rouge-1 F)比较多种文摘方法。在数据预处理阶段，我们采用TMG Matlab软件包[14]，该软件包提供了基于VSM模型的文本预处理功能，并可以对文档集进行停词处理、提取词干等操作。

文献[6]比较了基于LSA[4]和NMF[6]算法的文摘质量，因此，我们仅对基于NMF[6]和本文提出的基于SNMF的方法进行比较，NMF和SNMF算法使用模型选择得到的值作为基维度参数r。生成的文摘长度限定为 250 个词，不同文摘算法根据文档集句子的平均长度确定提取的文摘句数目，并按照句子的排名输出文摘。Wang[1]等人采用基于句子语义分析的对称NMF分解算法进行文摘提取，与本文采用的基于词频统计的文本表示法不同，因此不进行比较。

4.1 实验 1 根据模型选择确定基维度

针对 DUC2005 数据集，我们设定 SNMF 分解算法的稀疏度控制为 β =0.2，并使用 SNMF 算法计算不同基维度下的分散系数 ρ_k 的值，选择 ρ_k 值下降时对应的 k 值作为 NMF/SNMF 算法的维度参数。对 DUC2005 部分数据集（共 45 个子集）进行模型选择实验，并通过比较文摘质量说明模型选择的有效性。从模型选择结果表 1 中可知，在所有测试子集中，有 9 个文档集通过模型选择得到了最优的文摘质量，另有 14 个取得最优文摘质量的文档集在基维度取模型选择确定的次优值情况下得到，其他情况得到的文摘质量结果也较优。因此应用模型选择可以为 SNMF 分解算法基维度参数 r 的选择提供一定的依据，为文摘提取算法提供有效的指导作用。

基维度	文摘质量最优数	占测试数据集比	
模型选择最优值	9	20%	51.11%
模型选择次优值	14	31.11%	
其他值	22	48.89%	48.89%

表1 模型选择统计信息

4.2 实验 2 不同稀疏度对文摘质量影响

在实验1，我们由模型选择得到每个话题文档集对应的话题个数值，根据这个值，通过不同的稀疏度控制，我们得到了一组文摘，对其质量进行评价可以观察到不同的稀疏度对于SNMF算法学习文本潜在语义能力的影响，如图2所示，$\beta = 0.2$ 时SNMF文摘算法取得文摘的召回率最高，正确率也相应提高。当稀疏度比较高的时候 $\beta = 0.8$，文摘Rouge-1评价的正确率也会提高，说明此时句子的话题表示比较准确。

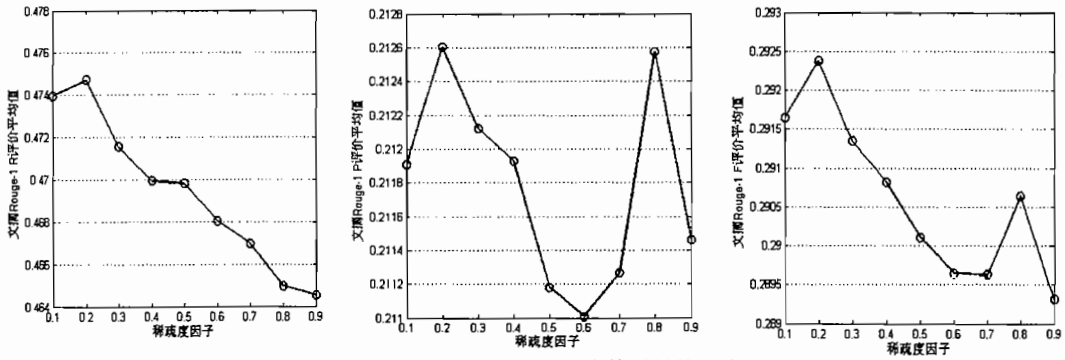

图2 不同稀疏度对文摘质量的影响

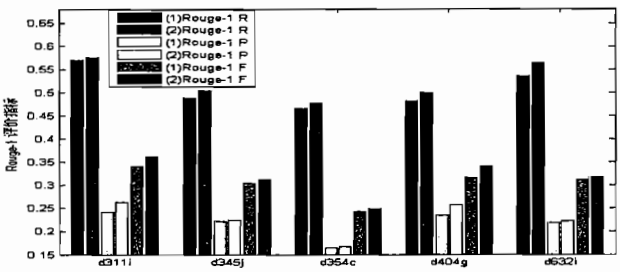

图 3 不同文摘算法比较

4.3 实验 3 不同文摘算法比较

本实验选取了实验 1 中经过模型选择正确确定基维度的 5 个 DUC2005 子文档集进行实验，比较了 NMF 算法和 SNMF 算法获得的文摘质量，如图 3 所示。由本文算法得到的文摘质量在召回率、正确率和 F-测度上都有所提高。不足之处在于，本文算法和文档集的原始属性有关，对于文摘质量改善不是非常明显。

5 总结与展望

本文基于稀疏非负矩阵分解提出了一种多文本摘要方法，通过稀疏度控制，比较了使用 1-范式稀疏度约束下不同稀疏度对自动文摘的影响；运用模型选择的方法自动学习文档集的潜在语义空间维度，较之 NMF 方法能够自动确定话题分类，并改进了文摘质量。后续研究可以比较不同约束控制对 NMF 分解算法的影响与意义，提出新的稀疏非负矩阵分解算法，改进本文提出的文摘算法。文档的不同的特征描述对挖掘算法有很大影响，因此，下一阶段的研究还可以针对文档的语义表示等特征描述方法、文本表示的加权算法、句子在潜在语义空间的权重计算及文摘句的选择、平滑、输出等问题研究。

参 考 文 献

[1] Dingding Wang, Tao Li, et al. Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization[C]. Proc. SIGIR'08, July 20-24, 2008

[2] Nomoto, Yuji, a new approach to unsupervised text summarization[C]. SIGIR'01, pp. 26-34

[3] Zha H. Generic Summarization and Keyphrase Extraction using Mutual reinforcement and sentence clustering[C]. Proc. SIGIR'02, pp. 113-120.

[4] Yihong Gong, et al. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis[C], SIGIR'01, 19-25.

[5] 林鸿飞，高仁璟，基于潜在语义索引的文本摘要方法[J]，大连理工大学学报，2001，Vol 41, No. 6。

[6] Lee, J. H., Park,S: 2007, Automatic Generic Document Summarization based on NMF[C], In Proceeding of BIS 2007

[7] Stan Z. Li, Xinwen Hou, et al. Local Spatially Localized, Parts-Based Representation[C], Proc. CVPR'01, 2001

[8] Patrik O. Hoyer, Non-Negative Sparse Coding[C], Proc. IEEE Workshop Neural Networks for Signal Processing, 2002

[9] Weixiang Liu, et al. Non-negative matrix factorization for visual coding, Proc. ICASSP 2003

[10] P. O. Hoyer, Non-negative Matrix Factorization with sparseness constraints[J]. The Journal of Machine Learning Research, 5:1457-1469, 2004

[11] Jingu Kim and Haesun Park, Sparse Nonnegative Matrix Factorization for Clustering[R], Tech Report, Georgia Institute of Technology, 2008.

[12] D. Lee and H. Seung, Learning the parts of object by non-negative matrix factorization[J]. Nature 401(6755):788-791, 1999

[13] http://duc.nist.gov/duc2005/, 2008

[14] TMG, URL http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/, 2009