数字文献资源聚合之高维表示模型与评价①

牛奉高[1]，邱均平[2]

（1. 山西大学 管理与决策研究所、数学科学学院，太原 030006；
2. 武汉大学 信息管理学院、中国科学评价研究中心，武汉 430072）

摘 要：为了提高数字文献资源主题聚合的水平，本文在向量空间模型（VSM）的基础上，通过补充文献特征词间的潜在语义相关性，提出了基于共现潜在语义的高维向量表示模型（CLSVSM），并采用人大复印资料 G9《图书馆学情报学》的文献为样本进行实验，较全面地评价了新模型及其增强模型相对于 VSM 的效果。

关键词：数字文献资源聚合；高维向量；VSM；CLSVSM

中图分类号：G350；O213 **文献标识码**：A **文章编号**：（2014）01-0072-10

0 引言

在海量信息背景下，资源聚合是解决信息过载的重要手段。数字文献资源的聚合是多层面和多角度的[1]。根据数字文献聚合的不同出发点，可以将其简单分为基于属性关联的聚合[2-4]和基于语义关联的聚合[5-6]。属性关联包括引用、耦合和合作关系等，而语义关联主要是基于本体或概念关系等。某些属性关联也蕴含着丰富的主题语义信息，因此两者的界分不是绝对的。本文更关注数字文献的主题聚合。主题聚合与文献主题检索、文献主题分类等实践工作密切相关。

文献的主题聚合可以通过文献特征向量的聚类来实现，其基本思想和文本分类、文本聚类相似。然而一直以来，相关研究或者关注聚类的方法，或者关注语义信息的提取，而鲜有对文献表示向量进行语义信息充实的研究。事实上，文献主题聚合结果的好坏，不仅与聚类方法和所借助的背景语义信息有关，更与文献语义信息是否充分提取、合理表示和语义相关性的计算是否科学等密切相关。正所谓“巧妇难为无米之炊”，如果没有很好地表示向量，再好的聚类算法都不能发挥到极致；如果没有语义充分表示的向量，再强大的语义信息背景面对很多零值的稀疏表示向量也无能为力。因此，算法固然重要，向量表示更重要。

1 相关研究综述与研究思路

1.1 文献高维向量表示的方法与困境

文献的表示方法源于文本的表示研究。文本的表示是文献实体的一种抽象代表，以便于文献间关系的建立、主题的比较和其他量化研究。常见的文本表示方法有：集合论方法、代数方法、概率统计方法、图论方法、混合方法[7]。其中，基于代数方法的空间向量表示最常用。文本的向量表示相当于一个映射，即

$$\phi: d \mapsto \phi(d) = (a_{1,d}, a_{2,d}, \cdots, a_{n,d}) \in \Re^{\mathfrak{D}}.$$

该映射将文本转化为高维空间中的向量，各特征词代表空间的坐标轴，文献集表示为矩阵。其代表模型就是哈佛大学的 Gerard Salton 等于 20 世纪 60 年代末提出的向量空间模型（VSM）。VSM 的基本思想是通过

① 基金项目：国家社会科学基金重大项目（11& ZD152）。
作者简介：牛奉高（1980-），男，山西沁水人，博士，管理与决策研究所、山西大学数学科学学院讲师，研究方向：信息计量与科学评价、应用统计，E-mail：nfgao@sxu.edu.cn；邱均平（1947-），男，湖南涟源人，武汉大学信息管理学院、中国科学评价研究中心教授，博士生导师，研究方向：信息计量与科学评价、知识管理与竞争情报，E-mail：jpqiu@whu.edu.cn。

对文本的处理，将文本内容表示为文档空间中的向量，采用空间相似度表征语义的相似度，当文本被表示为文本空间的向量后，利用向量之间的相似性度量文献间的相似性。VSM 模型在提出之后的 30 多年中，一直是信息检索的标准模型，Gerard Salton 也因此被称为现代信息检索的奠基人[8]。1985 年 Wong 等[9]针对 VSM 中关于词向量非正交的问题，提出了广义向量空间模型（generalized VSM，GVSM），提高了文本相似性计算的准确性。VSM 具有良好的可扩展性，此后的很多研究都是在它的基础上进行修正，比如增强语义[10-12]。

相对于文本的向量表示，文献的向量表示有很多难处，比如：①文献关键词少，表示向量具有严重的稀疏性。即使加入元数据中的其他主题词，比如对题目和摘要分词以后提取特征词作为文献特征词的补充，相对于文档空间（维数一般随着文献数量的增加而增加）来说，还是微不足道。这样的问题在文本表示中也存在，但一般的文本分析是基于全文提取特征词，可以选择适当数量的特征词，降低稀疏性。②权重的作用是有限的。频次或语义信息加权可以区别相同关键词的重要性，但对没有相同关键词的文献无法区别。因此，在计算文献两两相似度时，存在很多相似度为零，忽视了词间关系所表示的、可能在主题上相近的文献。③难以发挥语义增强方法的效果。通过语义关系来度量文献间的相似性，也是建立在表示词的显性关系上，对于文献与其不包括的关键词之间的潜在关系也少有挖掘和利用。

1.2 本文研究思路

面向文献聚合而提出的共现潜在语义向量空间模型基于 VSM 框架。但为了提高向量表示中的主题语义信息，本文采用了关键词潜在语义相关性来增强语义信息，同时降低高维表示向量的稀疏性。而对潜在语义相关信息的提取是基于共现分析。共现潜在语义就是基于词语共现关系强度来衡量特征词之间、特征词和文献之间的语义相关性，进而建立文献之间的主题相似性。

我们所使用的语义可以分为显在语义和潜在语义，相应的方法被称为显在语义分析（explicit semantic analysis，ESA）和潜在语义分析（latent semantic analysis，LSA）。基于序词表、本体和语料库等专业的、大型的“背景知识”来获取词汇语义及语义相关性，进一步得到文献的相似性并用于文献聚类的方法是比较理想的。但是因为这些“背景知识”尚不完善，如语义信息丢失、增加计算复杂性等[13]，因此在此基础上的分析结果就有失偏颇，再加上本体的学科领域限制、语料库的建设成本较高等问题，使得显在语义分析方法不能广泛应用。因此本文主要是挖掘和利用了潜在语义信息来进行建模。

2 面向文献资源聚合的高维表示模型

2.1 相关记号

设既定文献集中有 n 篇不同的文献，记为 $\boldsymbol{D}=\{d_1, d_2, \cdots, d_n\}$，所有文献共有 m 个互异的特征词，它们构成文献集的特征词集，记为 $\boldsymbol{T}=\{t_1, t_2, \cdots, t_m\}$。每篇文献 d_i 在 n 维特征词空间中对应为一个行向量，仍记为 d_i，每个分量取 0 或 1。当该文献的特征词中包含 k_j 时 $a_{ij}=1$，反之则 $a_{ij}=0$。n 篇文献按列排列构成如下布尔值型矩阵 $\boldsymbol{A}$，称之为文献-特征词矩阵，简称篇词矩阵。其中 $a_{ij}=0$ 或 1，$\boldsymbol{A}$ 是 $n\times m$ 矩阵。

$$\begin{array}{c} \quad t_1 \quad t_2 \quad \cdots \quad t_j \quad \cdots \quad t_m \\ \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_i \\ \vdots \\ d_n \end{matrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{i1} & a_{i1} & \cdots & a_{ij} & \cdots & a_{im} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nm} \end{pmatrix} = \boldsymbol{A} \end{array}$$

进而，文献-特征词的共现矩阵 $\boldsymbol{C}=(c_{ij})_{m\times m}$ 可以通过矩阵运算的形式给出，即 $\boldsymbol{C}=\boldsymbol{A}'\boldsymbol{A}$，特征词共现强度矩阵 $\boldsymbol{B}$ 表示为

$$\boldsymbol{B}=(b_{ij})_{m\times m}=\mathrm{diag}(\sqrt{c_{11}},\sqrt{c_{22}},\cdots,\sqrt{c_{mm}})A'A\mathrm{diag}(\sqrt{c_{11}},\sqrt{c_{22}},\cdots,\sqrt{c_{mm}})$$

式中，$b_{ij}=\dfrac{c_{ij}}{\sqrt{c_{ii}}\sqrt{c_{jj}}}$ $(i,j=1,2,\cdots,m)$。

2.2 基于潜在语义的数字文献资源高维向量空间模型

模型构建的具体思路是根据特征词共现对主题语义的凸显作用，用共词强度量化文献在空间中与各维度的相关性，从而使文献的向量表示能呈现更多主题信息，也使得文献在相关的维度上有连续取值，而不是只取 0 和 1 的离散数值，同时还一定程度上缓减了稀疏性，使聚类的结果更加可信。

首先给出文献 d_i 与特征词 k_j 的语义相关性测度方法。给定文献集 $\boldsymbol{D}$，及其特征词集 $\mathbf{K}$，篇词矩阵 $\boldsymbol{A}=(a_{ij})_{n\times m}$，共词矩阵 $\boldsymbol{C}=(c_{ij})_{m\times m}$，共词强度矩阵 $\boldsymbol{B}=(b_{ij})_{m\times m}$。记所有 $a_{ij}=1$ 的 j 的指标集为 $\mathbf{I}_{i1}$，即 $\mathbf{I}_{i1}=\{j\mid a_{ij}=1\}$，称 $\max\limits_{t\in I_{i1}}\{b_{jt}\}$ 为文献 d_i 与特征词 k_j 的语义相似性，记为 q_{ij}，即 $q_{ij}=\max\limits_{t\in I_{i1}}\{b_{jt}\}=\max\limits_{t\in I_{i1}}\{b_{tj}\}$。当 $a_{ij}=1$ 时，易知 $q_{ij}=0$。当 $a_{ij}=0$ 时，$0\leqslant q_{ij}<1$。

文献与特征词的语义相似性主要衡量该文献与其不包括的特征词的主题接近性，并作为文献在文档空间中该维度上的坐标分量，与其所含特征词共同表征文献的语义信息。在以上定义中选取与文献每个特征词共现强度的最大值作为文献与该特征词的语义相关性，即文献表示向量中相应该特征维上的坐标，是为了最大限度地表现主题相似性。因此，我们在原先只用 0 和 1 表示的文献特征向量基础上，增加更多语义信息即得到以下文献表示模型：

$$\varphi(d_i)=\tilde{d}_i=(q_{i1},q_{i2},\cdots,q_{im})\in\Re^{\mathrm{D}}$$

其中，$q_{ij}=\begin{cases}1, & a_{ij}=1\\ \max\limits_{t\in I_{i1}}\{b_{jt}\}, & a_{ij}=0,\ \max\limits_{t\in I_{i1}}\{b_{jt}\}\neq 0\\ 0, & \text{其他}.\end{cases}$ 且 $0\leqslant q_{ij}\leqslant 1$。当 $q_{ij}=1$ 表示特征词 k_j 是文献 d_i 的特征词，$q_{ij}=0$ 表示文献 d_i 不包含特征词 k_j，且 d_i 的其余特征词也与该特征词没有共现关系，即在文献集内不存在潜在语义关系。

新的向量表示也可以视为是布尔型表示向量与语义增强向量的叠加，该向量不仅保留了原有特征词（取值为 1），而且还在不包含的特征词分量上增加了通过共现潜在语义分析所提取的信息（取值小于 1），因此称之为共现潜在语义向量空间模型（co-occurrence latent semantic VSM，CLSVSM）。

2.3 模型的讨论

CLSVSM 模型的提出是面对只有简单元数据（主要是关键词）高维表示的文献资源的聚合，是适应“大数据”文献集基于元数据检索或聚类的低成本方法。对于小数据集的文献聚类可以采用较大数据集的共现关系作为潜在语义信息提取的知识背景，比如通过学科领域特征词的共现关系增强文献表示中的语义信息，或者采用题目词加强共现关系，如引入标题和摘要信息；对于有关键词频次信息的文献全文检索和文本聚类也适用，只需简单扩展即可；对于文献数量过多造成特征词维度过高的情况，可以使用频次较高的前 K% 关键词构建简约的 Top-K CLSVSM 模型。

3 模型的评价实验

3.1 评价指标和工具

为了不引起混淆，本文称根据已有标签的各个划分为类（class），称聚类以后的各个划分为簇（cluster）。在本实验中，假设簇数目与类数目相等，用聚类结果和原有分类作比较，采用熵值（entropy），纯度（purity）来判断聚类结果的优劣。

假设用于实验的文献按主题分为 k 个类，记为 L_j $(1\leqslant j\leqslant k)$，那么文献集采用某种聚类算法得到的聚类结果也应是 k 个簇，记为 S_r $(1\leqslant r\leqslant k)$。再设簇 S_r 和类 L_j 分别包含了 n_r 和 n_j 篇文献，其中有 n_{rj} 篇是相同的，而文献集中共有 n 篇文献，那么纯度和熵值定义为

$$Purity = \sum_{r=1}^{k} \frac{1}{n} \max_{1 \leqslant j \leqslant k} n_{rj} = \frac{1}{n} \sum_{r=1}^{k} \max_{1 \leqslant j \leqslant k} n_{rj}$$

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} \left(-\frac{1}{\log k} \sum_{j=1}^{k} \frac{n_{rj}}{n_r} \log \frac{n_{rj}}{n_r} \right)$$

最理想的聚类结果是簇和类的划分完全一致，此时纯度等于 1 或熵值等于 0。因此，实验中，纯度越大或熵值越小，表示聚类效果越好。

在数据处理阶段主要采用 Matlab 运行高阶矩阵运算，用 VBA 实现文献表示向量中共现潜在语义信息的补充。高维向量聚类采用高维可视化聚类软件图形聚类工具包（graphical clustering toolkit，gCLUTO），这是一个跨平台的高维和低维数据聚类且可以分析簇类特征的图形应用程序，可通过树状结构（tree）、矩阵（matrix）和基于 OpenGL（开放图形接口）的山丘型曲面（mountain）将聚类结果可视化，可从 Karypis 实验室[14]下载。

3.2 数据收集与整理

中国人民大学书报资料中心主办的《人大报刊资料复印资料》[15]是我国社会科学领域内重要的二次文献刊物，也曾是我国三大社会科学文摘刊物之一[16]。其中 G9《图书馆学情报学》设了 7 个稳定的栏目，分别是理论探索、实践研究、信息资源管理、信息技术与方法、信息服务、信息政策与法规和国际信息。7 个栏目是在同一学科下不同研究方向的划分，可以认为是目前较好的分类参考。个别栏目不是在主题层次上划分的，根据需要可以剔除。其中的文章都是编委从国内相关核心期刊中精选出来的，文献的质量、创新性、主题新颖性是有保证的[17]。这些文献是经过编辑分配到相应栏目的，因此文章的主题和栏目主题是非常接近的。如果专家是根据栏目来精选相应文献，就更贴合本实验的需要。

本文采集了 G9《图书馆学情报学》近 4 年（2010-2013 年）的所有文献。由于中国人民大学书报资料中心不提供文章题录格式的下载，因此数据的采集基本是纯手工的。整理阶段选择 refwork 题录格式，并对每篇文章进行编号。表 1 为栏目编号及各栏目文章数量分布。

表 1 G9 文献栏目-年数量分布

Table 1 Distribution of G9 literatures in sections-years

栏目编号	年份				汇总	
	2010	2011	2012	2013	总计	平均数
1 理论探索	47	42	48	40	177	44
2 实践研究	49	44	41	50	184	46
3 信息资源管理	48	57	53	55	213	53
4 信息技术与方法	42	40	33	30	145	36
5 信息服务	49	38	35	38	160	40
6 信息政策与法规	27	25	21	16	89	22
7 国际信息	25	30	25	25	105	26
8 信息法学	3				3	
总计	290	276	256	254	1076	1

数据显示，2010 年第 7 期出现了“理论研究”栏目，而少了“理论探讨”，数据整理中将它们视为一致的栏目，将这 4 篇“理论研究”的文章并入“理论探讨”栏目。此外，同年第 4 期出现了“信息法学”栏目，而该栏目在前后各期都没有，数据整理中保留了其中的 3 篇文章，归入栏目 6，全部数据仍分为 7 个栏目类。人大复印资料 G9《图书馆学情报学》2013 年的转载文献组成文献集Ⅰ，2010-2013 年 4 年的转载文献组成文献集Ⅱ，都包括栏目分类信息。表 2 是文献集的基本统计描述和文档空间的特征。其中关键词词频极差是最大词频与最小词频的差，空间维度等于不同关键词的总个数，稀疏程度用文档-关键词矩阵中非零值占总元素的比例表示，数值越小越稀疏。

表 2　实验文献集基本指标

Table 2　Basic indicators of experimental literature sets

文献集记号	说明	文献数	总词频	平均词频	词频极差	空间维度	稀疏程度
Ⅰ	人大复印资料 G9，2013 年	254	1048	1.3	15	791	1.7‰
Ⅱ	人大复印资料 G9，2010-2013 年	1075	4392	1.7	75	2586	0.7‰

3.3　实验设计

主要实验内容是比较基于 VSM 和基于新模型 CLSVSM 的聚类效果，以及为了提高精确度而进行的语义信息增强处理。实验代号见表 3。

表 3　文献聚类实验编号与说明

Table 3　Code and description of the literature clustering experiments

编号	模型	数据
V1	VSM	Ⅰ，关键词
VD1		减少分类的Ⅰ，关键词
V2		Ⅱ，关键词
VD2		减少分类的Ⅱ，关键词
C1	CLSVSM	Ⅰ，关键词及其共现语义信息
C12		Ⅰ，关键词+Ⅱ关键词共现语义背景
CD1		减少分类的Ⅰ，关键词及其共现语义信息
C2		Ⅱ，关键词及其共现语义信息
CT2	Top CLSVSM	Ⅱ，关键词+标题词及其共现语义信息
CTD2		减少分类的Ⅱ，关键词+标题词及其共现语义信息

以上实验分为三轮对比，具体是：第一轮对比 V1 和 C1，V2 和 C2，考察新模型对语义信息的提取情况，比较基于 CLSVSM 模型与基于传统 VSM 模型的聚类效果，以及与不同算法和准则搭配聚类的效果；第二轮对比 C1 和 C12，C2 和 CT2，考察不同共现潜在语义知识背景对聚类的影响，以及语义信息增强前后的效果；第三轮对比 VD1 和 CD1，VD2 和 CD2，以考察分类不准对结果的影响。

根据 Karypis 等研究中表现良好的算法[18]，笔者选择了 Direct K-means，Repeated Bisecting K-means。参考 Nasir 等的做法[19]，还选择了 Biased Agglomerative（Bagglo）算法。这些算法均可由 CLUTO Toolkit[20] 实现。实验中所采用的算法与准则函数的组合方案，以及记号等见表 4。

表 4　实验方案记号与说明

Table 4　description about Experimental plan

方案记号	聚类算法	准则函数
D-I_2，D-h_2	Direct K-means	$\mathcal{I}_2$，$\mathcal{H}_2$
RB-I_2，RB-h_2	Repeated Bisecting K-means	$\mathcal{I}_2$，$\mathcal{H}_2$
B-h_2	Biased Agglomerative（Bagglo）	$\mathcal{H}_2$

4　实验结果分析

4.1　CLSVSM 模型的语义信息增强效果分析

为了量化说明 CLSVSM 相对于 VSM 的语义信息增强情况，先给出如下定义：

1）共现语义信息：特征词共现关系所蕴含的语义信息；

2）共现语义信息增益率：通过 CLSVSM 方法有效植入文献表示向量中的共现语义信息与原来向量表

示中非零信息总数（总词频）的比例。

文献集Ⅰ有254篇文献，所有不同关键词791个，总频次1048，有663个关键词频次为1，约占总关键词的84%。通过共现分析揭示共现语义信息对3630个（考虑到共现的对称性，实际共现对有1815个不同的共现关系对，下同），其中共现强度为1的占46%。通过本文提出的共现潜在语义信息补充方法，选取利用共现强度非1的1946个有效共现信息中的1152个，为254篇文献的代表向量补充信息3775条。文献集Ⅱ有1075篇文献，不同关键词共2586个，总词频4392，只出现过一次的关键词占总关键词的78%。通过共现分析揭示共现语义信息对14268，其中共现强度为1的3982个，共现强度非1的10286个有效共现语义信息中被新模型采用6906个，共为1075篇文献补充信息66062个。对比结果见表5。

表5 两数据集基于CLSVSM的信息增益情况

Table 5 Information gain of sets based on CLSVSM

对比指标	文献集Ⅰ	文献集Ⅳ
总词频	1048	4392
共现语义信息对	3630	14 268
共现语义信息利用率	1152/3630≈32%	6906/14 268≈48%
共现语义信息增益率	3775/1048≈3.6	66 062/4392≈15.0
原篇词矩阵的稀疏度	1.7‰	0.7‰
新篇词矩阵的稀疏度	6.1‰	1.1%

由表5可见，文献集越大共现关系越多，所能利用的就多，这也是后面实验强调以更大共现矩阵为语义背景挖掘信息的原因。如果文献集Ⅰ以文献集Ⅱ的共现语义信息为背景，则利用的有效共现信息对达1796，大于只用其自有文献所提供的1152，而且补充信息对也达到6388，信息增益几乎是原来的2倍。

4.2 CLSVSM模型与VSM模型的对比

第一轮实验对比V1和C1，V2和C2，采用cos相似度和D-I2方案，因为这种组合经证实最适合VSM表示下的文献聚类，而且别的方案的聚类结果区别也不明显。结果见表6，该表中列出了两种对比5次实验的熵值和纯度以及各自5次的平均值，其中↓表示熵值越小越好，↑表示纯度越大越好，●表示同一模型或方案下最好，▲表示各组对比之下最好，▽表示各组对比之下最差（下文同）。

从表6可见，2013年数据的实验对比显示两模型的聚类效果相当，个别实验中新模型还略逊于旧模型，虽然熵值最小值出现在CLSVSM模型中，但最差的也在其中。而4年数据的对比结果中，新模型占了优势。5次实验最好的结果出现在基于CLSVSM模型的S1中，而最差的出现在基于VSM模型的S2中。而且5次实验的均值也显示基于新模型的结果要好，熵值最小为0.9026，纯度最大为0.2994。此外，同一模型下比较发现，数据量越大VSM模型的结果越差，而基于CLSVSM模型的结果受数据量影响不大。

表6 第一轮实验结果之熵值和纯度对比

Table 6 The compare results for the entropy and purity in the first round

评价指标	聚类编号	V1	C1	V2	C2
Entropy↓	S1	0.902	0.893	0.911●	0.893▲
	S2	0.887	0.873	0.927▽	0.901
	S3	0.885	0.910▽	0.926	0.897
	S4	0.879●	0.890	0.917	0.923
	S5	0.883	0.871▲	0.912	0.899
	平均值	0.8872	0.8874	0.9186	0.9026

续表

评价指标	聚类编号	V1	C1	V2	C2
Purity ↑	S1	0. 287	0. 276 ▽	0. 287	0. 324 ▲
	S2	0. 315 ▲	0. 311 ●	0. 274 ▽	0. 298
	S3	0. 291	0. 283	0. 279	0. 296
	S4	0. 299	0. 307	0. 287	0. 276
	S5	0. 291	0. 295	0. 301 ●	0. 303
	平均值	0. 2966	0. 2944	0. 2856	0. 2994

为深入考察模型的区别，接下来比较簇内的相似度。选择最好的实验结果：V1 的 S2 和 C1 的 S5，V2 的 S5 和 C2 的 S1，簇容量（即簇内文献数 Size）和簇内相似度比较见表 7。

表 7　第一轮实验结果之簇相似度对比

Table 7　The compare results for the cluster similarity in the first round

簇	V1-S2		C1-S5		V2-S5		C2-S2	
	Size	ISim	Size	ISim	Size	ISim	Size	ISim
0	7	0. 244	17	0. 513	95	0. 184	110	0. 26
1	21	0. 178	25	0. 345	88	0. 166	107	0. 238
2	30	0. 085	24	0. 26	144	0. 039	109	0. 119
3	46	0. 078	26	0. 238	106	0. 034	171	0. 067
4	31	0. 076	33	0. 169	157	0. 031	185	0. 063
5	41	0. 06	38	0. 15	179	0. 02	180	0. 05
6	78	0. 014	91	0. 019	306	0. 009	213	0. 037
簇平均		0. 105		0. 242		0. 069		0. 119

从表 7 可见，各簇基于 CLSVSM 的簇内相似度比基于 VSM 模型的簇内相似度要有很大的提高，因此平均相似度也就高了很多，一定程度上表明新模型对相关文献的识别度高了。此外，基于 VSM 的结果中也可以看到，簇内相似度高的都是簇容量较小的聚类。簇容量过小明显与原有分类不符，这一点也暴露了基于 VSM 模型聚类时稳定性较差。

分析 2013 年数据聚类结果没有优势的原因，可能是数据量少导致共现信息不足。2013 年共 254 篇文献，表示向量却高达 791 维，而 4 年的数据有 1075 篇文献，虽然维度达到了 2586，但共现关系加强了，发挥出了新模型的作用。对于所有实验整体表现的熵值偏高和纯度偏低的问题，可能是原有分类不准确造成的。因为熵值和纯度都是在聚类和原有分类对比之下计算的结果，如果原有分类有误，结果不好是必然的。好在以上结果对比表明新模型在较大文献集中略胜一筹，而且信息增益提高了相关文献的相似度。

4. 3　增强型 CLSVSM 模型与 VSM 模型的对比

第二轮对比实验是为了考察文献数量少所带来的共现信息不足对新模型聚类效果的影响。其中一组是对 2013 年数据采用 2010—2013 年的关键词共现信息进行语义信息补充，称之为“知识背景增强”实验（C12）；另外一组是对 4 年数据增加题目信息，即将题目分词以后，选择与主题相关的词和关键词一起表示文献向量，称之为“语义信息再增强”实验（CT2）。然后分别与第一轮实验结果比较。需要说明的是在 CT2 实验中，数据准备工作主要由手工完成，包括题目分词、选词与重组，因为分词系统准确性差；对题目词和关键词“一视同仁”，未进行加权处理；对合并后的关键词进行了整理，并采用 Top-K 模型。结果见表 8。

表 8 第二轮实验结果之熵值和纯度对比

Table 8 The compare results for the entropy and purity in the second round

评价指标	方案	V1	C1	C12	V2	C2	CT2
Entropy ↓	S1	0.902	0.893	0.858	0.911●	0.893	0.857
	S2	0.887	0.873	0.862	0.927 ▽	0.901	0.850▲
	S3	0.885	0.910 ▽	0.865	0.926	0.897	0.894
	S4	0.879●	0.890	0.851▲	0.917	0.923	0.865
	S5	0.883	0.871●	0.876	0.912	0.899	0.869
	平均值	0.887	0.887	0.862	0.919	0.903	0.867
Purity ↑	S1	0.287	0.276 ▽	0.311	0.287	0.324	0.342
	S2	0.315●	0.311●	0.315	0.274 ▽	0.298	0.348▲
	S3	0.291	0.283	0.331	0.279	0.296	0.326
	S4	0.299	0.307	0.339▲	0.287	0.276	0.345
	S5	0.291	0.295	0.303	0.301●	0.303	0.318
	平均值	0.297	0.294	0.320	0.286	0.299	0.336

从表 8 中看到了幸运的结果，上三角形均转移到了新的实验。对 2013 年数据而言，以 4 年关键词共现语义为背景知识，明显提高了同主题文献的相似度；最小熵值比第一轮实验的最小值降低了 2.3%，虽然较小，但是稳定降低，平均值也降为 0.862，比 VSM 的相应结果降低 2.8%；纯度也整体上升，都在 0.3 以上，平均值 0.320，增幅 7.8%。4 年关键词加题目词应用 CLSVSM 模型聚类的结果相比只用关键词的聚类结果提高更明显，熵值降幅 3.9%，纯度增幅 12.2%。但是这些提高并不能体现模型的真正优势，因为我们对文献集原有分类是有怀疑的，可能存在部分栏目的主题分散，有的栏目主题范围不明确。但多数栏目有主题意义，这也是以上实验没有更换数据的原因。

4.4 删减栏目后 CLSVSM 模型与 VSM 模型的对比

鉴于以上结果，我们需要重新考虑文献的主题分类问题。虽然栏目 1 实践探索和 2 实践研究文献很多，但主题并不能凸显主题，因此在以下的实验中我们删掉这两类，保留其余 5 类文献进行实验。删除前两个栏目后，数据集Ⅰ剩下 164 篇文献，数据量更少了；数据集Ⅱ剩下 715 篇文献。在删除两栏目相应的文献后，也对相应向量进行了降维，因为有一些维度全为零值。对新表示模型先补充语义信息后降维。重新整理以后数据集情况见表 9。

表 9 G9 文献集栏目删减后的数量和维度

Table 9 Number and dimension of the G9 literatures columns after cutting two classes

数据集	文献数	空间维度	数据集	文献数	空间维度
2013	164	502	4 年	715	1804
2013 共现		628	4 年共现		2363
2013 4 年共现		663	4 年+主题词共现		1549

4 年关键词和主题词的共现信息作为补充后的数据集进行了关键词合并，因此删除两个栏目后维数降得比较大。实验结果见表 10。

采用 5 次实验的平均值加标准差表示，同时对比了 D-I2 和 RB-I2 聚类方案的结果。首先比较分类减少前后的效果，5 类的熵值最小为 0.846，小于 7 类的 0.862；5 类的纯度最大是 0.394，大于 7 类的 0.336，因此可以说明原数据的分类影响了聚类的结果。其次是对方法进行比较，除了 2013 年数据下两种方案出现分歧外，基于新模型和其再增强形式的结果都好于 VSM 模型。

表 10　原数据集中保留 5 类后的聚类比较

Table 10　The compare results of five clustering after cutting the original data set

评价指标	Entropy ↓		Purity ↑	
	D-I2	RB-I2	D-I2	RB-I2
VD1	0.855±0.013	0.856±0.018	0.389±0.018	0.371±0.019
CD1	0.889±0.007	0.901±0.016	0.356±0.012	0.367±0.014
CD12	0.853±0.027	0.846±0.022▲	0.383±0.013	0.393±0.028
VD2	0.896±0.015	0.884±0.015	0.362±0.009	0.355±0.011
CD2	0.892±0.012	0.886±0.006	0.366±0.024	0.375±0.019
CTD2	0.849±0.025	0.856±0.013	0.394±0.029▲	0.381±0.019

5　结语

本文通过在 VSM 基础上植入潜在语义相关信息，提出了共现潜在语义向量空间模型（CLSVSM），并采用 G9《图书馆学情报学》文献集为数据，与 VSM 进行了对比实验。通过三轮对比，全面比较了新模型和 VSM 模型对文献主题信息的提取情况，模型聚类效果，以及受数据量的影响和共现语义信息增量的影响，整体结果显示 CLSVSM 是有优势的。

由于实验数据的匮乏，本文选择 G9 文献进行实验可能存在数据源单一不能完全展示新模型潜力的问题。虽然结果显示新模型的优势但是并不明显，原因主要是文献原有分类不能完全反映主题分类。在以后的研究中将寻求主题分类更加准确的实验文献集，继续讨论该模型的聚合能力。

参考文献：

[1] 贺德方，曾建勋．基于语义的馆藏资源深度聚合研究［J］．中国图书馆学报，2012，(4)：79-87.
He，D.，J. Zeng. Study on in-depth integration of library collections based on semantics［J］. *Journal of Library Science in China*，2012，(4)：79-87.（in Chinese）

[2] 杜晖．基于耦合关系的学术信息资源深度聚合研究［D］．武汉大学，2013.
Du，. H. Research on in-depth aggregation of academic information resource on the basis of coupling relationships［D］. Wuhan University，2013.

[3] 邱均平，董克．引文网络中文献深度聚合方法与实证研究——以 WOS 数据库中 XML 研究论文为例［J］．中国图书馆学报，2013，(2)：111-120.
Qiu，J.，K. Dong. Methods and empirical research on deep integration of literature in citation network：case study on xml research literature from WOS［J］. *Journal of Library Science in China*，2013，(2)：111-120.（in Chinese）

[4] 邱均平，王菲菲．基于共现与耦合的馆藏文献资源深度聚合研究探析［J］．中国图书馆学报，2013，(3)：25-33.
Qiu，J.，F. Wang. An Exploration of In-depth Aggregation of Library Document Re-sources Based on Co-occurrence and Coupling［J］. *Journal of Library Science in China*，2013，(3)：25-33.（in Chinese）

[5] 何超，张玉峰．基于本体的馆藏数字资源语义聚合与可视化研究［J］．情报理论与实践，2013，36（10）：73-76.
He，Ch.，Y. Zhang. A semantic integration and visualization model for library digital resources based on ontology［J］. *Information studies：Theory & Application*. 2013，36（10）：73-76.（in Chinese）

[6] Galar，M.，A. Fernández，E. Barrenechea，F. Herrera. Empowering difficult classes with a Similarity-based aggregation in multi-class classification problems［J］. *Information Sciences*，2014，264：135-157.

[7] 宋胜利，陈平，王少龙．面向文本分类的中文文本语义表示方法［J］．西安电子科技大学学报，2013，（2）：109-119.
Song，Sh.，P. Chen，S. Wang. Chinese test semantic representation for text classification［J］. *Journal of Xidian University*，2013，(2)：109-119.（in Chinese）

[8] 贺德方等．数字时代情报学理论与实践——从信息服务走向知识服务［M］．北京：科学技术文献出版社，2006.
He，D.，et al. Information science：theory and practice in digital age［M］. *Scientific and Technical Documentation Press*，

Beijing, 2006. (in Chinese)

[9] Wong, S., W. Ziarko, P. Wong. Generalized vector spaces model in information retrieval [C]. *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1985: 18-25.

[10] Liu, G. Semantic vector space model: Implementation and evaluation [J]. *Journal of the American Society for Information Science*, 1997, 48 (5): 395-417.

[11] 耿焕同，陈少军．一种基于传统 VSM 和词共现概念的中文文本聚类的研究 [J]．安徽师范大学学报（自然科学版），2005，(1)：27-30.

Geng. H., Sh. Chen. Research of chinese text clustering based on traditional Vsm and term co-occurrence [J]. *Journal of Anhui Normal University* (*Natural Science*), 2005, (1): 27-30. (in Chinese)

[12] 张彰，樊孝忠．一种改进的基于 VSM 的文本分类算法 [J]．计算机工程与设计，2006，27 (21)：4078-4080.

Zhang, Zh., X. Fan. Improved VSM based on Chinese text categorization [J]. *Computer Engineering and Design*, 2006, 27 (21): 4078-4080. (in Chinese)

[13] J. Nasir, I. Varlamis, A. Karim, G. Tsatsaronis. Semantic smoothing for text clustering [J]. *Knowledge-Based Systems*, 2013, 54: 216-229.

[14] Karypis Lab. http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/download. gCLUTO [EB/OL]. [2014-01-20].

[15] 中国人民大学书报资料中心．http://ipub.zlzx.org/. [EB/OL]. [2014-01-20].

The Information Center for Social Sciences of Renmin University of China (ICSS). http://ipub.zlzx.org/. [EB/OL]. [2014-01-20]. (in Chinese)

[16] 毛冀云．《中国人民大学报刊复印资料》信息传播中介作用的考察 [J]．内蒙古师范大学学报（教育科学版），2005，18 (11)：147-148.

Mao, J. A study about the intermediary role of information dissemination based on the Press Copy of Renmin University of China [J]. *Journal of Inner Mongolia College of Education* (*Educational Science*), 2005, 18 (11): 147-148. (in Chinese)

[17]《图书馆学情报学》的博客 [EB/OL]. http://blog.sina.com.cn/u/2725546651. [2014-01-20].

The blog of Library and Information Science. http://blog.sina.com.cn/u/2725546651. [2014-01-20]. (in Chinese)

[18] Steinbach, M., G. Karypis, V. Kumar. A comparison of document clustering techniques [C]. *KDD Workshop on Text Mining*, 2000, 400 (1): 525-526.

[19] Nasir, J., I. Varlamis, A. Karim, G. Tsatsaronis. Semantic smoothing for text clustering [J]. *Knowledge-Based Systems*, 2013, 54: 216-229.

[20] Karypis Lab: gCLUTO [EB/OL]. http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/download. [2014-01-20].

The Evaluation of High-dimensional Representation Model for Digital Literature Resource Aggregation

Niu Fenggao[1], *Qiu Junping*[2]

1. Institute of Management and Decision, School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China;
2. School of Information Management, Research Center for China Science Evaluation, Wuhan University, Wuhan 430072, China

Abstract: In order to improve the effect of digital literatures resources aggregation by their themes, the paper introduced CLSVSM based on the VSM, embedding the co-occurrence latent semantic correlation between the literatures' keywords to their representative vector. So the CLSVSM is a high-dimensional vector space model. Afterwards many experiments were done to test comprehensively the results of new model and its enhancement model relative to the VSM, adopt the People's University of China reprinted "Library and information science" (G9).

Key words: Digital literatures resources aggregation; High-dimensional vector; VSM; CLSVSM