大数据时代社会科学研究方法的拓展
——基于词嵌入技术的文本分析的应用

○ 冉雅璇 李志强 刘佳妮 张逸石

摘要 在大数据时代的背景下，基于大数据的分析处理技术为“数据驱动”的社会科学研究创造了新的发展契机。其中，词嵌入技术借势大数据浪潮，以其高效的词表征能力和强大的迁移学习能力在文本分析领域受到越来越多的关注。不同于传统的文本分析路径，词嵌入技术不仅实现了对非结构化文本数据的表征，还保留了丰富的语义信息，可以实现对跨时间、跨文化文本中深层次文化信息的挖掘，极大丰富了传统社会科学实证的研究方法。文章总结了词嵌入技术的基本原理及特点，系统地梳理了词嵌入技术的五大应用主题：社会偏见、概念联想、语义演变、组织关系和个体判断机制。随后，文章归纳了词嵌入技术的基本应用流程及结论有效性与稳健性的评估方法。最后，文章归纳了词嵌入技术在文本语料的选择、文本的分词处理、单词语义信息的表征层次三方面所面临的挑战，随即总结了相应的应对思路与方法。基于词嵌入技术的强大适应能力，未来研究可以进一步关注该技术在管理领域的应用前景，包括品牌管理、组织内部管理、中国传统智慧与管理问题三个方面。

关键词 词嵌入；自然语言处理；文本分析；社会科学；管理领域应用

引言

作为人类开展文化交流和情感沟通的基本载体，语言承担了重要的信息交换功能。借助于各类语言表达形式，人们将诸如知觉、思维、态度和情感等复杂的心理活动转化成特定的语言。[1] 而作为语言的典型载体之一，文本既能够在个体层面上反映人们的内心活动，也能够在组织和社会层面上反映集体文化。[2,3] 因此，从文本内容挖掘个体深层次的心理活动和人类社会的文化沿革是社会科学的基本研究路径。[4]

长期以来，在社会科学尤其是管理学和心理学等领域，实证研究多以针对实验、问卷和结构化的二手数据的量化分析为主导，而对于非结构化的文本材料（如访谈记录）仍以质性分析为主。[5] 在大数据时代，“数据 + 行为 + 交叉学科”已成为社会科学发展的必然方向。而计算社会科学的兴起则为理解人类行为、探讨社会现象提供了新的研究素材、视角和手段。[6] 随着互联网技术的飞速发展，人们在网络上发表大量包含思维、情感、观点的文本信息，这些井喷式爆发的文本为“以数据驱动”的社会科学研究提供了可及的信息来源。若能加以利用，无疑将拓宽社会科学研究的方法。[7–9] 然而，社会科学领域的传统文本研究方法以人工编码为主，在时间投入、成本和结论客观性等方面的不足极大地限制了文本数据在实证研究中的应用。所幸的是，以自然语言处理（Natural Language Processing, NLP）为核心的计算机文本分析技术（Computerized Text Analysis）的发展为大数据文本在社会科学领域中的应用带来了契机。[8,10]

“词”作为文本的最小语义单元，是计算机进行文

作者简介 冉雅璇，中南财经政法大学工商管理学院副教授、博士，研究方向为消费者行为与大数据营销；李志强（通讯作者），中南财经政法大学工商管理学院硕士研究生，研究方向为营销智能与文本分析；刘佳妮，中南财经政法大学工商管理学院硕士研究生，研究方向为营销模型与因果推断；张逸石，武汉理工大学管理学院教授、博士，研究方向为大数据营销与营销模型

基金资助 本文受国家自然科学基金项目（71802192、71702066）、教育部人文社会科学项目（18YJC630137）资助

本分析的基础。在自然语言处理领域，“词”主要以向量的形式表示。而词嵌入（Word Embeddings）技术，是一种可以把高维词向量映射进低维向量空间，以此来实现词表征的计算机文本分析技术。相较于其他自然语言技术，词嵌入技术不仅展现出了高效的学习能力，而且允许计算机从更高的意义单元（即目标词的上下文）出发理解词义，刻画“词”之间的相对关系，在管理学、心理学等社会科学领域取得了丰富的研究进展。相比于传统以人工编码和词频统计为主导的文本分析方式，词嵌入的优势在于：第一，借助计算机分析技术，可以在短时间内以较低成本实现对大规模文本数据的高效处理；第二，在挖掘文本特征和理解文本内容时，更多地依赖文本自身的分布规律，遵循“数据驱动”的分析逻辑；第三，面对跨时间、跨文化比较的研究话题，在挖掘社会学、行为学变量及变量关系等领域有广阔的应用前景。

词嵌入技术已在社会科学领域得到了广泛的应用，主要包括社会偏见、[11,12] 概念联想、[13] 语义演变、[14] 关系网络 [15] 和判断机制 [16] 五大主题，大量研究见诸国际知名期刊。反观国内的社会科学领域，词嵌入方法的应用价值还未得到足够重视和讨论。据此，本文通过介绍词嵌入技术的基本原理、梳理国外社会科学领域对词嵌入的应用情况，以期帮助国内社会科学研究者了解该技术独特的应用价值，推动大数据时代背景下我国的社会科学研究。

一、词嵌入技术的基本原理

不同于基于词频统计的文本分析方法，词嵌入技术的核心特征在于从文本的全局语义信息出发对“词”进行表征学习，[17,18] 即利用大规模文本中词的上下文信息，将文本词汇映射至高维向量空间以实现词的向量化表示，使得词向量之间既保留着词在语义层面的关联，又满足向量所适用的代数运算性质。在此基础上，通过度量词向量之间的几何关系（即距离）便能够刻画词在现实语义中的关系。[19] 进一步地，我们利用词与词之间这种可被量化的语义关系来探讨社会科学领域下概念之间的相似性或相关性，并由此刻画社会文化、心理变量与其他行为变量间的相关关系。因此，词嵌入技术的应用主要包含两大步骤，即首先利用词嵌入模型从文本数据中获得对词的向量表征，再计算词向量距离进行相关性分析。

1. 词的向量表征

纵观计算机文本分析的历史，词向量的表征方法主要经历了两个发展阶段。第一个阶段是从词典出发，基于词频统计规则对词的离散型表征。例如，热向量编码（One–hot Vector）通过建立基于目标文本（猫很可爱，狗也很可爱）的分词词典（{“猫”：0，“狗”：1，“也”：2，“很”：3，“可爱”：4}），将每个词表示为维度与词典长度相当的向量，且每个元素取值为 0 或 1（“猫”=（1, 0, 0, 0, 0），“狗”=（0, 1, 0, 0, 0），“也”=（0, 0, 1, 0, 0），“很”=（0, 0, 0, 1, 0），“可爱”=（0, 0, 0, 0, 1））。这一类词表征方法虽然简单直观，但在面对大规模文本时，词典长度的激增易造成“维度灾难”问题。①并且，该方法忽视了词的频率、上下文及词之间的关联，使得这一类词向量无法反映词的语义信息。为了提升词向量的表征质量，Deerwester 等主张从更高的文本意义单元理解文本词汇的含义。[20]

由此，分布式表征 [18] 成为了第二阶段的词表征方法。分布式假设是词嵌入技术背后的核心逻辑基础——即上下文相似的“词”拥有相似的或相关的语义，[21] 它既反映了人类的语言使用习惯，也符合人的现实认知逻辑。人们倾向于对具有相似或者相关特征的对象产生认知关联，体现在文本层面则是相近的语言表达或高度的共现频率，即相似的上下文语境。基于此，通过分析目标词与其上下文词汇之间的统计分布规律可以学习到目标词的众多文本语义信息。因此，分布式表征的思想被广泛应用于后续的语义学习中，成为了词嵌入技术的基本逻辑。

其中，较为出色且经典的是 Mikolov 等在 2013 年提出的 Word2Vec 模型，推动了词表征领域的变革。[18] Word2Vec 模型包含了两种神经网络结构——CBOW（Continues Bag–of–Words）和 SG（Skip–gram）。如图 1 所示，二者均由三个部分构成——输入层、输出层和隐藏层。CBOW 模型用于中心词 w_i 的预测任务，SG 模型则利用中心词 w_i 对其上下文词汇进行推断。通过隐藏层的特征学习“机器”，即参数矩阵 $w_{V\times N}$，将该中心词的上下文词向量 x_{ik}（由热向量编码表示）转化为低维实值向量（即词向量表征结果）。其中，参数矩阵 $w_{V\times N}$ 集合了词的 n 项关键特征维度，例如，“n=300”表示提取目标词在 300 个维度层面的特征信息。在此模型中，目标词的特征信息依赖于其与上下文词汇的共现模式，因而词的文本分布越相似，其对应的词向量具有更加相近的特征。因此，Word2Vec 模型既可以通过特征提取来实现词向量的降维，又可以反映语义信息。

除了 Word2Vec 词嵌入模型外，Pennington 等提出了同样具有高效学习能力的 GloVe（Global Vectors for Word Representation）框架，[23] 通过对词共现矩阵的矩

阵分解，实现对“词”的表征。后续学者提出了诸多对于词的分布式表征的改进算法，包括 fastText 算法、[24] 谷歌的 ELMo（Embedding from Language Models）语言模型 [25] 和 BERT（Bidirectional Encoder Representation from Transform）语言模型 [26] 等。

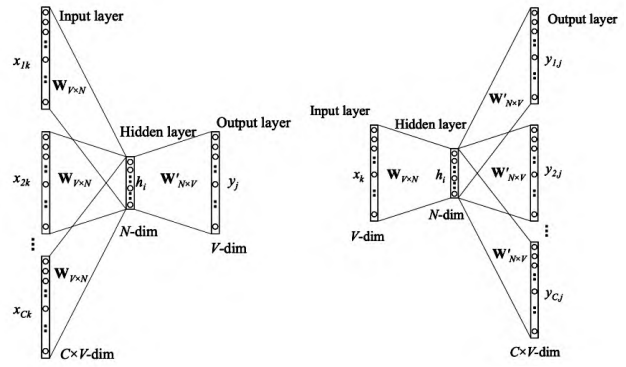

图1 Word2Vec模型中CBOW和SG原理示意 [22]

词嵌入技术下的向量具有两项重要的几何性质——“聚类”和“并行”。“聚类”性质是指现实语义相近的词在向量空间中的位置也相近。例如，“挪威”与“瑞典”的词向量更接近，而“意大利”和“德国”的词向量更接近。[27] 而“并行”性质是指向量空间中的词向量之间满足基本的代数运算性质，且这种运算逻辑基本符合词的现实语义逻辑。[19] 例如，从语义逻辑来看，“国王”和“王后”的区别平行于“男人”和“女人”的区别，反映到对应词向量上即是“$\overrightarrow{\text{King}}-\overrightarrow{\text{Man}}+\overrightarrow{\text{Woman}}=\overrightarrow{\text{Queen}}$”的代数形式。[28,29] 综合以上内容可知，词嵌入虽然聚焦在词这一最小的文本单位上，但是看到的是全局文本语义信息在词上的投射和体现。这不仅与传统的、基于词频的文本分析方法有着本质区别，还能够为文本分析提供更深刻、更生动的洞察。

2. 词向量的距离计算

词嵌入技术将文本中的词映射为 N 维向量空间中的点，而点与点的空间距离能够度量词与词之间、概念与概念之间乃至文档与文档之间的相关性。

（1）词与词之间的距离

设在 n 维语义空间中，单词 A 和 B 分别对应词向量 $v_A=(v_{A1},\ldots,v_{An})$ 和 $v_B=(v_{B1},\ldots,v_{Bn})$，$v_A$ 与 v_B 之间的距离计算方式主要有以下两种：

①余弦相似度（Cosine Similarity）：

$$sim(v_A,v_B)=(v_A\times v_B)/(\|v_A\|\times\|v_B\|)$$
$$=(\sum_{i=1}^{n}v_{Ai}\times v_{Bi})/(\sqrt{\sum_{i=1}^{n}(v_{Ai})^2}\times\sqrt{\sum_{i=1}^{n}(v_{Bi})^2})$$

余弦相似度衡量词向量 v_A 和 v_B 之间的向量夹角的余弦值，其取值范围为 [–1,1]。余弦相似度取值为 0，则代表单词 A 和 B 之间不存在语义关系；而取值的绝对值越靠近 1，表明单词 A 和 B 之间的相关性越强。

②欧式距离（Euclidean Distance）：

$$dis(v_A,v_B)=\sqrt[2]{\sum_{i=1}^{n}(v_{Ai}-v_{Bi})^2}$$

欧式距离越小表明单词 A 和单词 B 之间的语义关系越强，反之则越弱。

（2）概念与概念之间的距离

在词嵌入分析领域，一个概念由一系列“相关词”组合而成，如“女性”概念可以通过“女人”“女生”“母亲”等名词来表达。Garg 等、Caliskan 等分别构建了以下两种相对距离的计算方法，[12,17] 以此对比不同属性概念（如“女性”vs.“男性”与“智慧”）之间的相关性：

① 相对范数差函数

$$\sum_{v_m\in M}\|v_m-v_A\|-\|v_m-v_B\|$$

该函数用于衡量两项目标词概念与某一项特征概念的相对距离。其中，M 代表特征概念（如“智慧”），v_m 为所属概念的相关词向量（如“智慧 / 聪明”）；v_A 和 v_B 分别代表两类目标词向量（如“男性”vs.“女性”）。该函数的含义为，在“男性”和“女性”两类群体中，哪一类群体与“智慧”这一概念更相关。若函数值为正，则代表“女性”与“智慧”更相近；负值则代表“男性”与“智慧”更相近；若函数值靠近 0，则表明“智慧”不存在明显的性别偏向。

②词嵌入相关性检验（WEAT）

$$s(X,Y,A,B)=\sum_{v_x\in X}s(v_x,A,B)-\sum_{v_y\in Y}s(v_y,A,B)$$
$$s(v_w,A,B)=mean_{v_a\in A}\cos(v_w,v_a)-mean_{v_b\in B}\cos(v_w,v_b)$$

该框架用于衡量两组目标词 X,Y（如“男性”vs.“女性”）与两组属性词 A,B（如“事业”vs.“家庭”）在语义上的相对距离差异，其中 v_w 为所属概念的相关词向量（如用“男生”“父亲”“男人”等词语描述“男性”概念）。$s(v_w,A,B)$ 表示单词 v_w 与两类属性词 A 和 B 的相对距离，正值代表其与 A 属性更相关，反之则相关性弱；而 s(X,Y,A,B) 则衡量了两项目标词 X,Y 和两项属性词 A,B 相对距离的差异，即在“男性”“女性”两类群体中，哪一类群体与“事业”或“家庭”的文化概念更相关。若 s(X,Y,A,B) 为正值，则表明相比于“女性”，“男性”与“事业”的语义相关性更高，反之则表明“女性”与“事业”的语义相关性更高。此外，WEAT 框架还提供了相应结果的显著性检验方式及效应量指标。

（3）文档与文档之间的距离——词移距离

$$\min_{T\geq 0}\sum_{i,j=1}^{n}T_{ij}\times c(i,j),\ s.t.\ \sum_{j}^{n}T_{ij}=d_i,\forall i\in\{I,\ldots,n\}$$

除了概念间的相关性分析，我们可以通过文档间的相似性来探讨如文本主题、个体及组织之间的关系问

题。Kusner 等提出“词移距离”的计算方法，[30] 即通过对两个向量语义空间中所有词向量间的欧式距离进行加权求和，以此衡量两个文本间的相似性，如上式所示。其中，c(i,j) 为词向量间的欧式距离；T_{ij} 为词向量间距离的权重（由 TF–IDF②计算加权值）。函数值越大代表两个文本的相关程度越低，反之越高。

二、词嵌入技术的优势

传统的社会科学研究通常需要借助科学实验、社会调查和人工编码等方法，依赖于专家学者的领域知识和实践直觉，存在主观性较强、耗时、耗资源的缺点。另外，传统的社会科学研究局限于小样本数据和历史数据的不足，通常关注当下的、有限范围的社会情景，难以进行跨时间、跨文化的分析。[12] 反观以词嵌入为代表的计算机文本分析，可高效地处理大文本数据；能够利用现有数据和先验知识改进算法，可拓展性和重复性强；依据文本内在的分布规律学习和提取信息，结果更加客观；从大规模文本中挖掘代表整体社会的认知，尤其擅长跨时间、跨文化的文本研究，结论不仅具有广泛的代表性，而且可以展示相关文化概念、思想观念的历时演化。这些优点极大地丰富了社会科学的研究方法，拓展了社会科学的研究视野。两类方法的具体区别见表 1。

表1　传统社会科学研究与词嵌入技术社会科学研究的区别

对比维度	传统的社会科学研究路径	基于词嵌入技术的社会科学研究路径
研究工具	问卷、访谈、实验、案例分析等	Word2Vec、GloVe 等词嵌入模型，及词向量、概念及文本的相关性计算
方法依据	基于实践经验和严格的理论推断，依赖专家学者的领域知识和实践直觉，是以人为中心的研究方法；围绕研究假设进行数据检验的分析路径	基于语言文本来理解文化概念和思想观念，综合利用社会科学理论、计算机科学等探讨社会、心理和行为层面的问题，是人智与计算机相结合的分析方法；不依赖严格的假设，利用数据挖掘展开探索性的研究
检验标准	大部分研究结论缺少严格客观的评断标准，主观性较强；结论的可复现性较弱	有多项较为成熟的指标及评价流程，具体包括检验词嵌入模型的训练效度（模型在特定的测试任务集上的表现）和检验研究结论的外部效度（将结论与其他社会调查数据、其他研究方法的结果对比）；结论的可复现性较强
数据来源	调研记录、实验数据、文献等；受限于成本投入，数据来源较为单一且体量较小	数据来源广泛，能够熟练处理包括会议记录、网络文本、新闻书籍等非结构化的文本数据；在处理大规模、跨文化、跨时间的文本数据时有优势
信息层次	以基于自我报告的外显认知为主，在获取被访者内隐认知时依赖于间接的方法设计；研究较大地依赖于样本选择，结论在跨时间、跨文化上的代表性有限	允许研究者直接挖掘文本所反映的内隐社会认知；研究较多从社会、集体层面的文本范围着手，结论具有较强的代表性和普适性

三、词嵌入技术在社会科学领域的应用话题

1. 社会偏见 / 刻板印象

文本语言能够反映人类对世界的认知和态度，基于词嵌入的文本分析方法可以有效地挖掘社会偏见和刻板印象。Garg 等采用词嵌入技术分析了来自纽约时报、谷歌新闻、谷歌图书及美国历史文本库，③揭示了 1900–1990 年美国社会在性别和种族两大议题上的刻板印象及其历史变化。[12]Garg 等首先以 10 年为单位将文本分为 9 份，并针对每一时段使用词嵌入技术以获得词的向量表征。[12] 进一步，借助相对范数差函数计算了一组词向量（如“男性”vs.“女性”）与目标词（如“专业工作”）的相对距离，以此度量社会刻板印象的程度。他们发现近百年间在美国社会中始终存在着较为明显的性别偏见和种族偏见。例如，“女性”和护士、保洁、舞者、秘书等职业联系更为紧密，而与工程师、木匠、技术人员等职业更为疏远。此外，亚裔姓名和教授、科学家、化学家和工程师等学术职位联系更紧密，白人姓名往往与警察、统计学家、摄影家等职位联系更紧密。通过分析概念间距离随时间推移的动态演化发现，这两类刻板印象有减弱势态，部分时段的变化也反映了美国民权运动的深刻影响。

作为文化的产物之一，歌曲也能反映社会认知中的偏见。Boghrati 等利用 Word2Vec 词嵌入技术，挖掘了自 1965–2018 年美国公告榜上流行歌曲歌词中隐含的性别偏见。[11] 一方面，在流行音乐中发现了“厌女症”的刻板印象（如图 2），即相比“男性”词汇，创作者们更少将“女性”词汇和“能力 / 成功 / 热情”等具有积极属性的词汇联系起来。但另一方面，歌曲中所反映的性别偏见有减弱趋势。作者们通过控制创作者的性别，发现男性作词人是影响歌曲“厌女症”现象变化的关键因素。

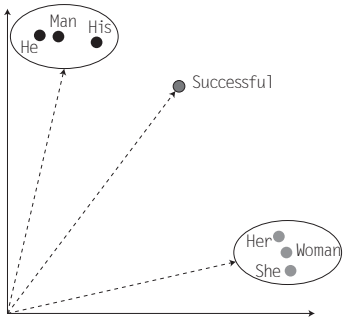

图2　流行音乐歌词中“成功”与“男性”的距离相较于与“女性”的距离更近[11]

为了检测词嵌入技术是否能够有效挖掘文本中的社会偏见，Caliskan 等对比了词嵌入模型和内隐联想测试（IAT）④的差异。[17] 虽然 IAT 是社会科学领域最常用于测量内隐认知的方法之一，但该方法需要严格的实验环境、耗时较长且测量样本受到时间和空间的局限。Caliskan 等基于 GloVe 词嵌入模型构建了 WEAT 分析框

架，并利用这一框架研究了人类社会的 8 项内隐认知，如“科学—男性”和“艺术—女性”。[17] 他们的研究表明，基于词嵌入技术的结论与基于 IAT 的结论具有高度一致性。在未来的内隐态度研究中，词嵌入技术能够作为 IAT 研究方法的替代性方案。

除了探讨社会偏见与刻板印象的跨时特征之外，词嵌入技术还适用于跨文化的对比分析。Defranza 等利用词嵌入技术探讨了不同地域的性别偏见差异。[31] 他们利用 fastText 模型和 WEAT 分析框架，从 49 类不同语种的文本中分别挖掘不同社会文化下的性别偏见。结果显示，当一个地区的性别语言，即语言中更加区分词汇的性别属性和使用者的性别身份（如泰语、芬兰语）更强时，该地区的性别偏见更加明显（图 3），具体表现为男性与积极属性词汇的关联性更强。这一结果在一定程度上验证了萨皮尔—沃夫假说，即语言能够塑造人的思维和认知。

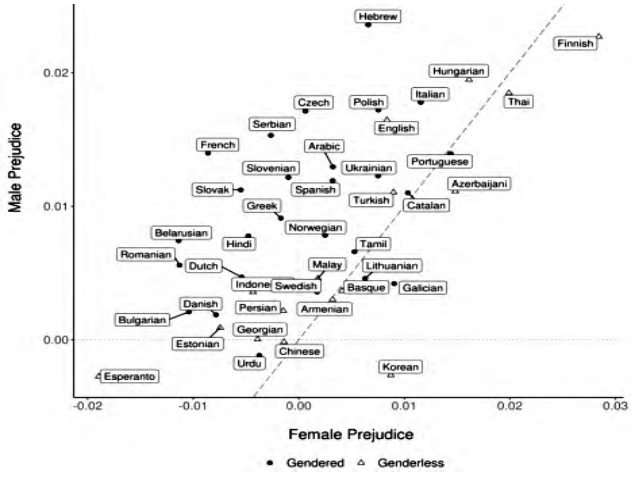

图3 语言文化与性别偏见的相关性[31]

2. 文化认知

历史无法复刻，但承载历史痕迹的文本资料能够帮助人们窥探特定时代背景下的文化内涵。Kozlowski 等利用词嵌入技术分析了 1900–1999 年百余年间公开发表的书籍，[28] 探讨了 20 世纪美国社会对七大等级文化（财富、道德、职业、性别、教育、品味、身份地位）的共识和演变规律。作者利用 Word2Vec 模型构建了一系列标度不同等级文化维度（如“经济水平”“道德”“性别”）的词向量空间（如“性别—财富”“职业—道德”“品味—职业”“教育—地位”）。进一步，作者将一系列目标词分别映射进相应的等级维度空间，以此探索这些词的多元等级属性。例如，在被映射进如图 4 所展示的“性别—财富”等级维度空间后，“Volleyball（排球运动）”一词表现出更靠近“Feminine（女性气质）”和“Rich（富有）”的特征。此外，标度不同等级维度的向量之间的夹角也具有社会文化含义。例如，“教育”与“道德”和“品味”的相关性较强且历时稳定，但与“职业”这一维度的相关性较弱，这说明提升教育水平与提升人的修养和品味相关，但与职位状态的关联较小。Kozlowski 等的研究表明，[28] 词向量可以用于探讨多元文化维度之间的语义结构和社会文化共识。

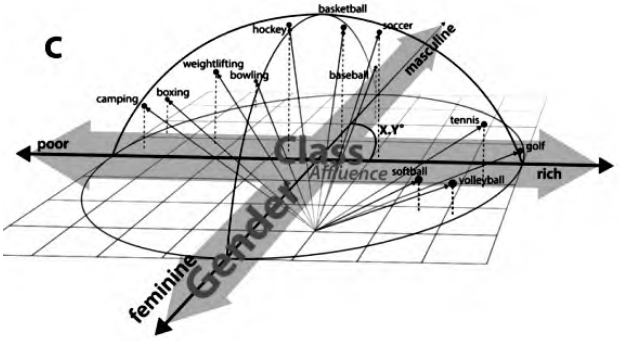

图4 目标词在“性别”与“财富”两重维度上的投影[28]

作为文化概念的关键形式，社会认知是人们社会动机系统和社会情感系统形成变化的基础。社会认知包括社会信念和社会态度两部分。而根据内隐—外显双系统理论，社会认知可进一步分为外显社会认知和内隐社会认知两类，前者强调个体可以通过自省来报告的社会认知，后者则描述个体无法内省的、潜意识层面的社会认知。[32] 然而，有关内隐信念和内隐态度的关系，已有研究要么将其混为一谈，要么将其作为互不干涉的独立构念。为厘清该问题，Kurdi 等利用预训练的 fastText 词嵌入技术，[13] 分析了内隐态度和内隐信念的关联，并对比其与外显态度和外显信念的差异。具体而言，基于被试自我报告的结果，个体的外显态度与外显信念存在方向上的不一致性，例如，亚裔群体常被白人给予负面的评价（外显态度），但在智商、能力方面被认为有突出优势（外显信念）；而词嵌入的分析结果则表明内隐态度与内隐信念具有一致性，内隐态度驱使内隐信念的产生。例如，白人群体有较高的自我评价（内隐态度），也认为本群体的智商高于亚裔群体（内隐信念）。可知，词嵌入技术可以作为挖掘社会认知的有效工具。

3. 语义内涵演变

语言的涵义会随着时代发生改变，而词嵌入技术的一大突出优势即表现为处理跨时段的文本数据。文本语言的运用具有系统规律性，通过针对来自不同历史时段的文本训练词嵌入模型，有助于学者在时间维度上分析

词义的演变。Hamilton 等通过测量目标词向量的时间位移，来描述历史文本中高频词汇与多义词汇两类词汇的变化。[14] 如图 5 所示，Gay 作为一个多义词，在 20 世纪 10 年代的文本中和 Cheerful（开心）和 Frolicsome（玩闹）词义更接近，而到 20 世纪 90 年代则与 Homosexual（同性恋）、Lesbian（女性同性恋）等更接近。据此，Hamilton 等提出了两条语义演化法则：一致性，高频词汇会保持词义相对一致的历时演化规律；新颖性，多义词汇的语义演化会更加快速。[14]

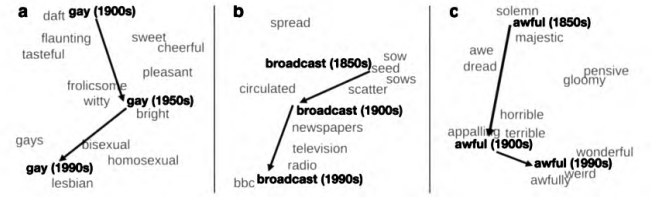

图5 目标词汇含义的历时演变[14]

除了对普遍意义上的词汇含义的演变规律进行探讨，某些具体词汇的演化同样值得关注，尤其是那些能反映特定时代背景的词汇。Rodman 等基于 1855–2016 年的纽约时报、路透社报道、美联社报道三大新闻文本集，[10] 挖掘并追踪了美国一个多世纪以来围绕“平等”一词的词义演变，其中包含了使用环境、指代对象等。在 20 世纪 50 年代之前，即美国民权运动前，“平等”的词义与“社会”话题相关的词汇的关联度整体较高，但随后呈现减弱趋势，这一结果与美国民权运动前对社会公平的激烈讨论现象符合（如种族歧视）。而自 20 世纪 70 年代，“社会”与“经济”、“教育”等概念下的词汇的关联程度不断增强，反映了 21 世纪以来美国社会对“教育公平”“经济公平”等热门话题的高度关注。可见，词嵌入分析方法能够敏锐地捕捉到社会文化的演变线索，进而为社会、文化等领域的运动转型提供预示。

4. 组织关系分析

词嵌入技术可用于挖掘不同组织在价值观和意识形态层面的关联，以此作为组织关系的推断依据。在此思路基础上，Spirling 等采用 GloVe 和 Word2Vec 的词嵌入模型，[33] 分析了美国共和党和民主党在各自公开发言稿中对部分政治议题的态度。例如，对于“堕胎”议题，两政党的理解存在较大争议，民主党认为“堕胎”是一种自愿选择，而共和党认为“堕胎”与“绝育”“公平”的话题相关；对于“税收”议题，两政党的理解则存在更多共识。由此可见，词嵌入技术不仅可以帮助我们了解政党组织在哪些政治议题上存在冲突，还可进一步推断党派政治关系。

Rheault 等分析了英国、加拿大和美国 20 世纪以来的议会记录文本，[34] 并构建了“党派嵌入”模型。学者利用词嵌入模型量化了不同党派与特定“意识形态”维度（如自由 vs. 保守、北部 vs. 南部）的相关性，从而对比不同党派组织的意识形态差异。从整体上看，美国民主党的意识形态更靠近自由派思想（如“民权”“种族”“枪支管控”），而美国共和党更具保守派和南部州色彩（如“官僚”“果农”“烟草”），且两党的意识形态差异正逐渐扩大。Pomeroy 等利用 GloVe 词嵌入模型分析了各个国家在联合国论坛的演讲文本，[15] 并使用词移距离来量化国家讲演文本间的总体相似性，以此来反映国家的政治立场。基于词嵌入技术的分析结果能比较真实地反映国家间的政治关系。例如，虽然土耳其和希腊两国在投票议程中表现出一致的国家态度，但实则两国在当年发生了边境军事冲突，而这一冲突能从两国的联合国讲演文本中捕捉到线索。[15] 本研究指出，有关词嵌入在主体网络关系的应用还可以分析其他情景下的主体关系，如社交网络关系、品牌竞争关系、组织内部关系等。

5. 个体的判断与决策心理

决策结果和决策信息线索之间具有表征关系，因而词嵌入技术能够通过挖掘概念间的内在关联，在一定程度上揭示个体在决策任务中的思维过程和决策依据。Bhatia 在自然语言处理的框架下，[16] 验证了以往决策研究中的相关性判断机制，即人们在进行判断性任务（如 A 多大可能属于 B）时，会出于直觉心理去思考问题与选项间的相关性或相似性，并以此作为判断依据。具体而言，综合 Word2Vec、CCA、GloVe 几项词嵌入技术，基于谷歌新闻和 GigaWord 文本库⑤训练生成词向量。[16] 进一步，作者通过对句子中每个“词”的向量求取平均值，分别对判断问题（如在德国的以下两座城市中，哪一个人口最多）与选项（如汉堡和科隆）实现表征。作者依据两者间的语义相关性来预测答案选项的概率，并据此模拟一般决策者的选择。例如，针对上述问题，基于词嵌入模型的预测结果为“汉堡”，与被试的选择高度相似。此外，词嵌入模型在其他测试任务（如经典的“Linda 问题”）⑥下也预测了决策者的选择倾向，[16] 这一现象与代表性启发理论（一种依赖相关性感知进行识别和判断的心理决策过程）相符。这说明词嵌入技术为我们理解人的判断和决策心理提供了信息参考，甚至对其中常见的认知偏差，如合取谬误、⑦基础概率忽略⑧也能够予以反映。[16]

另外，个体的风险感知和风险判断也是个体决策研究中的重要部分。Bhatia 利用词嵌入技术探讨了人们面对各类风险源时的风险评估机制。[35] 基于谷歌新闻文本

的预训练 Word2Vec 模型，作者量化了不同风险源（技术性风险源："新兴技术""能源"等；活动性风险源："运动""职业"等）与相关概念的语义联系，揭示了人们评估风险时的知识表征内容。例如，当评估药物风险时，人们会潜意识地联想到"毒品""无序"等具有高风险含义的概念（如图 6a 的词云图）；而评估运动风险时，人们容易联想到"碰撞""斗争"等风险事件（图 6b）。作者弥补了以往有关风险研究中难以预测样本外数据（如新型风险源）的缺陷，展现了词嵌入技术在理解和预测个体判断决策机制中的应用优势。（词嵌入技术应用总结见表 2）

(a) (b)

图6 个体评估"药物"风险和"运动"风险时的感知联想词云[35]

四、词嵌入分析的基本流程

词向量的表征学习存在两条路径：一是采用本地化的训练模型。二是使用预训练的词嵌入模型。针对第一条路径，通常需要经历如下预处理和模型训练步骤（见图 7）：（1）选择语料库。语料库是用于训练词嵌入模型的文本集，"词"的表征效果及后续的相关性分析依赖于训练文本的规模、质量及其语言环境。文本语料的选择依研究问题而定，使研究主题与文本主题相对应。文本数据的获取主要有以下三种途径：第一，公开且已初步整理规范的文本数据库。如，人民日报文本集（1946 年至今）、[11] 谷歌图书（包含 1500–2012 年公开出版的书籍，约占人类历史所有出版书目的 6%）、[12] 亚马逊评论集（包含 1996–2018 年亚马逊平台近 30 个产品品类超过 2 亿条评论）等。[13] 第二，借助爬虫程序收集文本数据。例如，众筹平台的项目材料、社交平台的历时推文、论坛上的互动文本、企业员工在 Glassdoor[14] 等职业资讯网的日志评论等。第三，研究者还可以将纸质文本转为电子文本形式，如员工日记、会议记录、心理咨询稿等。（2）语料预处理。常规的预处理流程包括：删除与文本内容无关的标点符号、特殊字符和停用词（如代词、连词）等。此外，还需要对文本分词，即将语料处理成由词这一最小语义单位所构成的列表。[15]（3）模型训练。在预处理后的语料文本中训练词嵌入模型，最终实现词的向量表征。当前主流的词嵌入模型有 Word2Vec、GloVe、fastText 等，可在 Python 环境下调用开源工具包（如 Gensim）中的算法模块，并自主调整相关参数。

表2 词嵌入技术在社会科学领域的应用主题总结

主题	作者	年份	词嵌入模型	研究内容	语料库/训练集	时间跨度
社会偏好 / 刻板印象	Garg 等 [12]	2018	SVD、GloVe 和 Word2Vec	探讨美国女性、少数族裔在职业和人格特质方面所面临的社会偏见，及其历史变化趋势	谷歌新闻、纽约时报、谷歌书籍和 COHA	1900-1990
	Boghrati 等 [11]	2020	Word2Vec（CBOW）	美国流行歌曲的歌词文本中对女性的偏见及其历史变化；不同音乐流派的歌词文本中所反映的性别偏见存在差异	美国音乐公告榜（Billboard）上榜歌曲的歌词文本	1965-2018
	Caliskan 等 [17]	2017	GloVe 和 WEAT	构建 WEAT 分析框架，通过度量系列目标词与属性词之间的语义相关性，以揭示隐含在人类认知层面的事物偏好	爬虫数据集（Common Crawl Dataset）	/
	Defranza 等 [31]	2020	fastText 和 WEAT	在跨文化背景下对比了不同地域社会中的性别偏见，发现性别偏见与当地的语言性别属性高度相关，结论支持了萨皮尔—沃夫假说	49 种语言文本 - 维基百科和爬虫数据语料库（Common Crawl Dataset）	/
文化概念及联系	Kozlowski 等 [28]	2018	Word2Vec（Skip-gram）	构建标度各类等级概念的维度，借此探讨美国社会对七类等级文化维度的基本共识、等级文化概念间的相互作用及其历时变化	Google Ngrams [9] 文本集	1900-1999
	Kurdi 等 [13]	2019	fastText	探讨美国社会对不同群体的内隐态度与内隐信念及两者间的关系，揭示了内隐社会认知与外显社会认知的不一致性；对比了词嵌入技术与 IAT 测试方法的研究结论	爬虫数据库（Common Crawl Dataset）	2017-2018
语义演变	Hamilton 等 [14]	2016	PPMI、SVD 和 SGNS（Word2Vec）	利用 SGNS 计算标准化语义历时位移值；利用 PPMI 构建文本词汇的多义性衡量指标；总结一致性和新颖性的词义演化法则	ENGALL、ENGFIC、COHA、FREALL、GERALL、CHIALL [10]	1900-1990
	Rodman[10]	2020	Word2Vec	追踪美国社会中的"平等"概念与性别、种族、经济等话题的关联变化，由此揭示"平等"在不同时代背景下的语义内涵；结合具体史事，发现词嵌入技术能够对美国民权运动的发展提供预示	纽约时报、路透社报道和美联社报道	1855-2016
组织关系	Pomeroy 等 [15]	2019	GloVe	利用"词移距离"的相关性计算方法，分析不同国家讲演文本的文本语义结构，挖掘国家间在政治立场上的相似程度，以此衡量国家间的政治关系，辅助国际关系网络的构建	联合国大会一般性辩论演讲文本	1970-1990
	Spirling 等 [33]	2014	GloVe 和 Word2Vec	分别探讨美国两大政党对部分社会议题的理解，并根据两大党派对特定议题的共识范围，即对社会议题的认知的相似性来衡量党派关系	国会记录（共和党、民主党议事文本）	102nd-111th 美国国会会议
	Rheault 等 [34]	2019	Word2Vec（CBOW）	利用词嵌入技术构建"党派嵌入"模型；计算英国、加拿大和美国国内的不同党派与"自由—保守""北部州—南部州"两项维度的相对距离，据此分析党派组织的意识形态差异	国会 / 议会会议记录（英国、美国、加拿大）	1873-2016—美国 1935-2014—英国 1901-2018—加拿大
判断决策	Bhatia [16]	2017	Word2Vec CCA（Eigenwords vectors） GloVe	将词嵌入技术运用至相关性判断任务之中，通过分析判断性问题与选项在文本上的语义相关性，以此预测答案的概率分布，从而说明词嵌入技术的分析逻辑实则模拟了人的心理及行为模式	谷歌新闻、GigaWord 语料库和维基百科数据库	/
	Bhatia[35]	2019	Word2Vec	利用词嵌入技术解释了人们面对各类风险源时的内在风险评估机制和评判依据，揭示了人们对部分风险源的知识表征的具体内容，即内隐联想	谷歌新闻文本集	/

基于词嵌入模型的迁移学习能力，也可直接使用预训练的词嵌入模型（如谷歌的 GloVe[23] 和 BERT[26]），从而获得基于其他大型语料库充分训练的词向量，并基于研究问题对模型或表征结果进行微调。但无论采用何种词向量表征路径，在词向量相关性分析之前，都有必要对词嵌入模型的训练结果进行评估。常见的评估方式是通过与人工标注的词相关性评分进行对比，以此判断词嵌入模型是否能够捕捉一般化的语义关系。目前，已有大量成熟的针对“词对”相似性或相关度的人工标注测试集，如MEN–3000（英文）、[36]Wordsim240/297（中文）。[37]

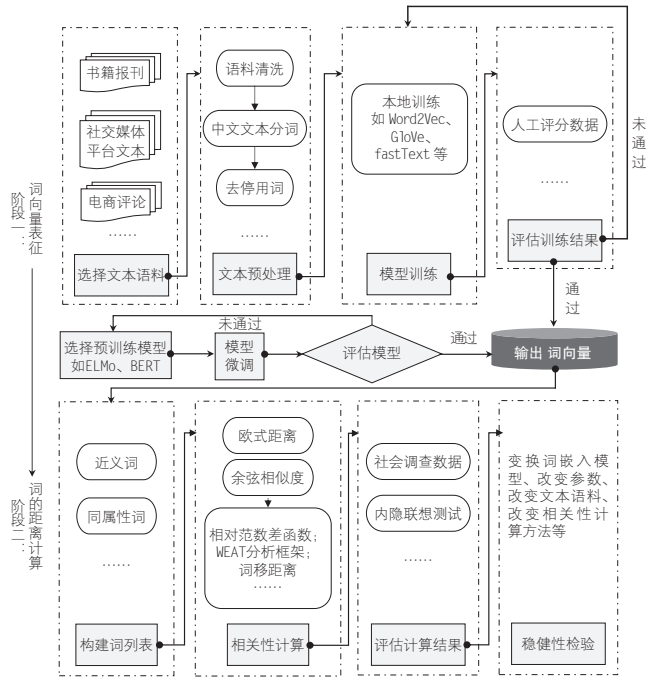

图7 词嵌入技术在文本分析中的应用流程

词嵌入模型训练完成后，通过构建词列表、计算相关性、有效性检验和稳健性检验四个步骤来完成相关性分析阶段。（1）构建词列表。依据词典、量表等方式挑选特定概念的常用近义词或同属性词来组成该概念的词列表，并保证词列表的区分效度。（2）计算词向量的相关性。针对具体的研究问题，衡量词、概念或文档之间的向量距离，主要包括余弦相似度、欧式距离两种基本的计算方法。（3）有效性检验。针对计算机的分析结果，我们有必要开展进一步的检验，以保证结论的可靠性及方法的有效性。具体包括两类检验方法：①与对应年代的相关调查数据比对，以判断词与词的相关程度、变化是否与相应的指标数据、社会事件吻合；[12,16] ②与其他研究方法对比，如内隐联想测试（IAT）、主题模型（LDA），以检验词嵌入模型能否重复已有研究结果。[10,13]（4）稳健性检验。作为一种无监督的探索性分析方法，词嵌入的分析结果会因文本、模型等因素的不同而产生差异。通过变换词嵌入模型、参数、文本语料或相关性计算方法可以检验研究结果的一致性。

五、词嵌入分析方法的应用挑战与展望

1. 词嵌入分析方法应用挑战

（1）词嵌入的分析效果依赖文本数据的体量和语言环境。①通常来说，文本数据规模越大越有利于词嵌入学习和提取更充分的语义信息。[10] 针对体量较小的文本，Rodman 提出了两种解决思路：一是采用基于超大型文本数据的预训练模型。[10] 二是采取自举法 ⑯ 生成规模更大的文本数据集，并对不同的抽样过程下的词向量结果求其平均值。②文本情景、社会文化环境、观点立场等背景信息在很大程度上影响着文本词汇的分布，可能导致结果偏差。正如 Spirling 等发现不同党派的议事文本展现出不同的政治视角。[33] 再如，企业的官方书面文本与员工的口述文本可能也有观点差异。就学者的普遍实践来看，文本选择需要“有的放矢”，即依据研究问题对文本的背景信息进行筛选，在扩大语料规模和类型的同时尽可能聚焦同一视角和语境。[10,28,33]

（2）中文文本的分析需要预先“分词”，对于某些特定领域的文本而言，如专业学术文章、古代汉语文本等，由于其文本内容及结构与标准的训练语料存在较大差异，使得文本分词的过程面临困难。近年来，文本预处理技术获得了长足进步，例如，Deng 等开发了 TopWORDS 分词器，[38] 在小型文本中实现了部分低频词的精确识别，亦能处理含有专业语汇的文本，进而可以探讨古代社会背景下如围绕性别、权力、道德、宗教等的伦理课题及社会价值观念的演变。

（3）词与词之间的组合搭配能够涌现更加丰富而抽象的语义信息（如段落含义、文本主题），这一类信息难以通过词向量间简单的结构化公式运算来体现。[19] 词嵌入技术所建构的是词与词之间的关联，侧重于表达“单词级别”的语义信息。学界也在积极探索“组合式分布语义”的实现方法，即如何利用词表征的组合实现对句子、段落和文档的表示。[18,39] 鉴于词嵌入模型较强的扩展能力，大量学者将有关文本整体特征的信息融入模型训练过程。例如，Liu 等结合 LDA 算法，⑰ 使 Word2Vec 模型训练下的词向量包含更多的主题特征，如“Apple”在电子产品的背景信息下表示“苹果公司”，而在食品背景下表示“苹果”这类水果。[40] 此外，词嵌入的基本原理在长文本表征领域也得到了发展。例如，Le 等将 Word2Vec（Skip–gram）的算法运用至句子和短文本的表征学习；[41] 词向量模型界的新秀——BERT 语

言模型，能够有效表征“单词级别”以上的文本语义概念，推动了对更高文本单位的关系的理解。

（4）传统的社会科学研究方法具备词嵌入技术所无法提供的分析视角，尤其是相对于文本细读法，词嵌入的机器学习技术难以捕捉单词的细粒度语义。[42] 例如，对同义词、反义词、多义词、上下位词等词义的区分和表征还有待优化。据此，相关学者提出利用有监督的学习过程，在词嵌入的算法层面引入某些先验知识库（如描述词义关联信息的 WordNet 语义网 [43]），帮助模型更好地捕捉单词多元的属性信息。（词嵌入技术应用挑战及应对方法见表 3）

表3 词嵌入技术面临的挑战及应对方法

面临挑战	具体困难	应对方法或思路
文本数据的选择	（1）对于小体量的文本语料而言，模型的训练效果会受到一定限制	（1）采用预训练的词嵌入模型 （2）采取“自举法”生成规模更大的文本数据集 [10]
	（2）词表征效果和结论推演严格依赖于文本所处的语言环境，表达视角和立场不同的训练文本可能会产生不同的研究结论	依据研究问题对文本的背景信息进行分析和筛选。训练文本的来源和类型越丰富，越反映一般化、综合性问题；训练文本越聚焦，越体现特定的视角与立场
中文文本的分词处理	某些领域的文本（专业学科、古代文本体裁）在其内容或结构上与标准的训练语料存在较大差异，使得文本的分词过程会存在一定困难，进而影响模型的训练效果	（1）邀请领域专家辅助构建相关的核心词词典，作为分词阶段的参考依据，提高分词结果的可靠性 （2）改进模型算法。例如，TopWORDS 中文分词系统能够在小型文本语料中精确地识别低频词，能有效处理含有未知专业语汇的文本 [38]
单词语义信息的表征层次	（1）词嵌入模型建构词与词之间的关联，侧重于表达“单词级别”的语义信息。基于词向量的简单组合模式（如加法、乘法、卷积法）难以刻画文本整体且抽象的语义信息（如主题、背景、情绪）	（1）纳入额外的知识库信息。将有关文本整体特征的信息（如文本主题、情绪标签）融入词嵌入的学习算法过程（如主题词嵌入模型 TWE）[40] （2）借鉴词嵌入技术的基本原理，并直接运用至文本层面的语义表征学习（如句子嵌入 Sentence2Vec、文档嵌入 Doc2Vec）[41]
	（2）词嵌入作为一种计算机化的分析方法，尚难以达到人工水平的对细微的语义信息的识别与区分（如同义词、多义词、上下位词）	（1）采取有监督的模型训练。在词嵌入技术的算法层面引入某些先验知识库（如语义网 WordNet[43]），帮助模型更全面地捕捉单词词义的属性信息 （2）调整词嵌入模型的学习机制，使之尽可能包含文本的全局信息，挖掘更深层次更多元化的词义信息（如 ELMo 语言模型 [25]）

2. 词嵌入分析应用例举

（1）数“智”品牌管理

词嵌入技术能用于刻画品牌—消费者关系，辅助企业的品牌管理。借助词嵌入的分析方法，企业可以透过用户生成文本纵观市场对企业品牌的态度、评价 [9, 44] 及品牌个性的挖掘，如测量品牌与个性维度间的相关性（如真诚型 vs. 粗犷型）。在跨文化视角下，词嵌入技术能够帮助企业考察不同文化背景下的市场对其品牌的感知差异。[45] 其他相关话题，诸如品牌依恋、品牌文化和品牌联想等研究也将受益于词嵌入的分析方法。

（2）组织内部管理

在管理学领域，有关组织行为的研究大多采用问卷访谈、自然观察和开展田野实验的研究方式。这些研究路径在理解和预测个体行为的过程中存在较强的主观性和外显性，本研究提出，词嵌入技术可以用于分析组织内成员的心理及行为规律，通过挖掘组织内的文本（如会议记录、员工评述、领导讲演文本），揭示员工的内隐认知信息（如动机、信念、情绪），甚至包括领导力、员工创新力、员工的组织支持感和企业文化等主题。

（3）中国传统智慧与管理问题

中国社会背景下的众多管理问题、思想乃至组织行为领域的话题，也能够从历史事件中获得借鉴。例如，Huang 等基于对《资治通鉴》这一古籍的人工编码，探讨了中国家族式企业内的领导—员工关系。[46] 此外，通过对《二十四史》展开词嵌入分析，也能帮助学者了解中国各朝代的管理层在应对人事、外部环境、组织治理等方面的管理思想与策略。对此，本文展望利用词嵌入方法对中华古籍文本展开分析，挖掘诸如组织领导风格、组织文化、管理者与下属间关系、人员激励政策等研究话题，探索中国本土的管理智慧和组织话题。

参考文献

[1] Solso, R. L., Maclin, O. H., Maclin, M. K., Pearson.. Cognitive Psychology: Pearson New International Edition. European Journal of Epidemiology, 2008, 23(6): 411-422.

[2] Kramsch, C. J.. Language and Culture. AILA Review, 1998, 27(1): 30-55.

[3] Samovar, L., Porter, R., Mcdaniel, E., Roy, C.. Communication between Cultures. Boston: Cengage, 2000.

[4] Tausczik, Y. R., Pennebaker, J. W.. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 2010, 29(1): 24-54.

[5] 曹奔，夏勉，任志洪，林秀彬，徐升，赖丽足，王琪，江光荣. 大数据时代心理学文本分析技术——“主题模型”的应用. 心理科学进展，2018, 26(5): 770-780.

[6] Buyalskaya, A., Gallo, M., Camerer, C. F.. The Golden Age of Social Science. Proceedings of the National Academy of Sciences, 2021, 118(5); e2002923118.

[7] Loughran, T., Mcdonald, B.. Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 2016, 54(4): 1187-1230.

[8] Humphreys, A., Wang, J. H.. Automated Text Analysis for Consumer Research. Journal of Consumer Research, 2018, 44(6): 1274-1306.

[9] Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., Schweidel, D. A.. Uniting the Tribes: Using Text for Marketing Insight. Journal of Marketing, 2020, 84(1): 1-25.

[10] Rodman, E.. A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. Political Analysis, 2020, 28(1): 87-111.

[11] Boghrati, R., Berger, J.. Quantifying 60 Years of Misogyny in Music. 2020, Working Paper.

[12] Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceed-

ings of the National Academy of Sciences, 2018, 115(16): E3635.
[13] Kurdi, B., Mann, T. C., Charlesworth, T. E. S., Banaji, M. R.. The Relationship between Implicit Intergroup Attitudes and Beliefs. Proceedings of the National Academy of Sciences, 2019, 116(13): 5862-5871.
[14] Hamilton, W. L., Leskovec, J., Jurafsky, D.. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, (1): 1489-1501.
[15] Pomeroy, C., Dasandi, N., Mikhaylov, S. J.. Multiplex Communities and the Emergence of International Conflict. PLos One, 2019, 14(10): e0223040.
[16] Bhatia, S.. Associative Judgment and Vector Space Semantics. Psychological Review, 2017, 124(1): 1-20.
[17] Caliskan, A., Bryson, J. J., Narayanan, A.. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Science, 2017, 356(6334): 183-186.
[18] Mikolov, T., Chen, K., Corrado, G., Dean, J.. Efficient Estimation of Word Representations in Vector Space. 2013, arXiv:1301.3781.
[19] 白长虹 . 管理研究与学术研究公式化 . 南开管理评论 , 2020, 23(6): 1-2.
[20] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.. Indexing by Latent Semantic Analysis. Journal of the Association for Information Science & Technology, 2010, 41(6): 391-407.
[21] Harris, Z. S.. Distributional Structure. Word, 1954, 10(2-3): 146-162.
[22] Rong, X.. Word2vec Parameter Learning Explained. 2014, arXiv:1411.2738.
[23] Pennington, J., Socher, R., Manning, C.. Glove: Global Vectors for Word Representation. Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.
[24] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.. Bag of Tricks for Efficient Text Classification. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference, 2016, (2): 427-431.
[25] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Zettlemoyer, L.. Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018, (1): 2227-2237.
[26] Devlin, J., Chang, M. W., Lee, K., Toutanova, K.. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2018, (1): 4171-4186.
[27] Collobert, R., Weston, J., Bottou, Léon, Karlen, M., Kavukcuoglu, K., Kuksa, P.. Natural Language Processing (almost) from Scratch. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.
[28] Kozlowski, A. C., Taddy, M., Evans, J. A.. The Geometry of Culture: Analyzing Meaning through Word Embeddings. American Sociological Review, 2018, 84(5): 905-949.
[29] Bhatia, S., Richie, R., Zou, W.. Distributed Semantic Representations for Modeling Human Judgment. Current Opinion in Behavioral Sciences, 2019, (29): 31-36.
[30] Kusner, M. J., Sun, Y., Kolkin, N. I., Weinberger, K. Q.. From Word Embeddings to Document Distances. 32nd International Conference on Machine Learning, 2015, (2): 957-966.
[31] Defranza, D., Mishra, H., Mishra, A.. How Language Shapes Prejudice against Women: An Examination across 45 World Languages. Journal of Personality and Social Psychology, 2020, 119(1): 7-22.
[32] McClelland, D., Koestner, R., Weinberger, J.. How do Self-attributed and Implicit Motives Differ? Psychological Review, 1989, 96(4): 690-702.
[33] Spirling, A., Rodriguez, P. L.. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. Journal of Politics, 2014, 84(1): 101-115.
[34] Rheault, L., Cochrane, C.. Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. Political Analysis, 2019, 28(1): 1-22.
[35] Bhatia, S.. Predicting Risk Perception: New Insights from Data Science. Management Science, 2019, 65(8): 3800-3823.
[36] Bruni, E., Boleda, G., Baroni, M., Tran, N. K.. Distributional Semantics in Technicolor. Meeting of the Association for Computational Linguistics, 2012, (1): 136-145.
[37] Chen, X., Xu, L., Liu, Z., Sun, M., Luan, H.. Joint Learning of Character and Word Embeddings. In Proceedings of the 24th International Conference on Artificial Intelligence, 2015: 1236-1242.
[38] Deng, K., Bol K. P., Li K. J., Liu J. S.. On the Unsupervised Analysis of Domain-specific Chinese Texts. Proceedings of the National Academy of Sciences, 2016, 113(22): 6154-6159.
[39] Bhatia, S., Richie, R., Zou, W.. Distributed Semantic Representations for Modeling Human Judgment. Current Opinion in Behavioral Sciences, 2019, (29): 31-36.
[40] Liu, Y., Liu, Z., Chua, T. S., Sun, M.. Topical Word Embeddings. Proceedings of the National Conference on Artificial Intelligence, 2015, (3): 2418-2424.
[41] Le, Q., Mikolov, T.. Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on International Conference on Machine Learning, 2014, (32): 1188-119.
[42] 孙飞 , 郭嘉丰 , 兰艳艳 , 徐君 , 程学旗 . 分布式单词表示综述 . 计算机学报 , 2019, 42(7): 1605-1625.
[43] Miller, G. A.. WordNet. Communications of the ACM, 1995, 38(11): 39-41.
[44] Netzer, O., Feldman, R., Goldenberg, J., Fresko, M.. Mine Your Own Business: Market-structure Surveillance through Text Mining. Marketing Science, 2014, 31(3): 521-543.
[45] Camiciottoli, B. C., Ranfagni, S., Guercini, S.. Exploring Brand Associations: An Innovative Methodological Approach. European Journal of Marketing, 2014, 48(5-6): 1092-1112.
[46] Huang, X., Chen, L., Xu, E., Lu, F., Tam, K-C.. Shadow of the Prince: Parent-incumbents' Coercive Control over Child-successors in Family Organizations. Administrative Science Quarterly, 2020, 65(3): 710-750

(下转第 79 页)

1995, 48(1): 1-23.
[31] 张峰 , 黄玖立 , 王睿 . 政府管制、非正规部门与企业创新 : 来自制造业的实证依据 . 管理世界 , 2016, (2): 95-111,169.
[32] Porter, M. E., Der Linde, C. V.. Green and Competitive: Ending the Stalemate. Long Range Planning, 1995, 6(28): 128-129.
[33] Porter, M. E.. America's Green Strategy. Scientific American, 1991, 264(4): 193-246.
[34] Jaffe, A. B., Palmer, K.. Environmental Regulation and Innovation: A Panel Data Study. Review of Economics and Statistics, 1997, 79(4): 610-619.
[35] 蒋为 . 环境规制是否影响了中国制造业企业研发创新——基于微观数据的实证研究 . 财经研究 , 2015, 41(2): 76-87.
[36] 彭雪蓉 , 魏江 . 利益相关者环保导向与企业生态创新——高管环保意识的调节作用 . 科学学研究 , 2015, 33(7): 1109-1120.
[37] 王锋正 , 姜涛 , 郭晓川 . 政府质量、环境规制与企业绿色技术创新 . 科研管理 , 2018, 39(1): 26-33.
[38] 李婉红 . 排污费制度驱动绿色技术创新的空间计量检验——以 29 个省域制造业为例 . 科研管理 , 2015, 36(6): 1-9.
[39] 王旭 , 褚旭 , 王非 . 绿色技术创新与企业融资契约最优动态配置——基于高科技制造业上市公司面板数据的实证研究 . 研究与发展管理 , 2018, 30(6): 12-22.
[40] Chen, E., Miller, G. E.. Stress and Inflammation in Exacerbations of Asthma. Brain Behavior & Immunity, 2007, 21(8): 993-999.
[41] 赵丽娟 . 政府 R&D 投入、环境规制与农业科技创新效率 . 科研管理 , 2019, 40(2): 76-85.
[42] 林玲 , 赵子健 , 曹聪丽 . 环境规制与大气科技创新——以 SO_2 排放量控制技术为例 . 科研管理 , 2018, 39(12): 45-52.
[43] Ang, J. S., Cheng, Y., Wu, C.. Does Enforcement of Intellectual Property Rights Matter in China? Evidence from Financing and Investment Choices in the High-tech Industry. Review of Economics and Statistics, 2014, 96(2): 332-348.
[44] 王惠 , 王树乔 , 苗壮 , 李小聪 . 研发投入对绿色创新效率的异质门槛效应——基于中国高技术产业的经验研究 . 科研管理 , 2016, 37(2): 63-71.
[45] 许昊 , 万迪昉 , 徐晋 . 风险投资、区域创新与创新质量甄别 . 科研管理 , 2017, 38(8): 27-35.
[46] 王兰芳 , 王悦 , 侯青川 . 法制环境、研发"粉饰"行为与绩效 . 南开管理评论 , 2019, 22(2): 128-141,185.
[47] 冯戈坚 , 王建琼 . 企业创新活动的社会网络同群效应 . 管理学报 , 2019, 16(12): 1809-1819.

注释

① 因滞后 2 期和滞后 3 期模型的拟合优度均低于滞后 1 期的模型，因此在"内外部参照的优先级分析"仅针对滞后 1 期的回归结果展开讨论。

(上接第 56 页)

注释

① 维度灾难是指在涉及向量计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象。
② 即词频—逆向文件频率，是一种用于信息检索与文本挖掘的加权技术。TF-IDF 常用以评估一个字词对于一个文件集或一个语料库中其中一份文件的重要程度。
③ 即 COHA（Corpus of Historical American English）数据集，该数据集包含美国 1810-2009 年的各类历史文本，语言以英语为主，单词条目总数超过 40 亿。
④ IAT 是基于生理的神经网络模型，以被试的"反应时长"为指标，用于测量被试对特定概念间的相关性的潜在认知。每一个参与 IAT 的被试都会被分配若干组的分类任务，并被要求以尽可能快的速度完成。如果该分类任务符合被试的内隐态度，则被试将会表现出更快的反应速度；反之则会更慢。通过对比被试在面对不同任务反应时长的差异，研究者即可探知到个体的内隐态度。
⑤ GigaWord 文本库包含近 400 万篇新闻和公开发表的文章，常用于文本总结和标题生成任务。
⑥ "Linda 问题"是指"琳达, 31 岁，单身，一位直率又聪明的女士，主修哲学。在学生时代，她就对歧视问题和社会公正问题较为关心，还参加了反核示威游行。请问琳达更有可能是下面哪种情况？"有两个选项："A. 琳达是银行出纳；B. 琳达是银行出纳，同时她还积极参与女权运动"。相比于 A 选项，B 选项所塑造的女性形象更贴近问题所提供的信息，因而人们会倾向于选择 B 选项。
⑦ 合取谬误是指人们总是认为两个事件的联合出现比只出现其中一件事的可能性要大。以"Linda 问题"为例，人们会更多地选择 B，虽然从实际概率角度来讲，B 选项的概率应低于 A 选项。
⑧ 基础概率忽略是人们在进行主观概率判断时，倾向于使用当下的具体信息而忽略掉一般常识的现象。
⑨ Google Ngrams 文本集由 1500-2008 年公开出版的书籍构成，包含 8 种语言，书目总数超过 800 万本，约占人类历史所有出版书目总数的 6%。
⑩ 代表英语、法语、德语、中文四大语种下的历史文本数据集。文本集主要由公开发表的各类型书籍构成，横跨 1800-1999 年近 200 年的时段。
⑪ http://data.people.com.cn/rmrb/。
⑫ http://storage.googleapis.com/books/ngrams/books/datasetsv2.html。
⑬ https://nijianmo.github.io/amazon/index.html。
⑭ Glassdoor(https://www.glassdoor.com) 是美国的一家提供企业点评与职位搜索功能的职场社区网站，任何企业员工均可以在该网站匿名评论雇主。
⑮ 当前学界常用的中文分词程序有 Jieba、SnowNLP、HanLP、PkuSeg、THULAC 等。
⑯ 通过对原样本数据集的重置抽样以生成新的随机样本及统计量的方法，在原始样本数据有限的情况下可以建立起足以代表总体样本分布的新样本。
⑰ 即隐含狄利克雷分布（Latent Dirichlet Allocation），是一种主题模型算法。该模型能够从整个语言系统中学习词的主题信息，并用于推测文档的主题及主题词分布。

大数据时代社会科学研究方法的拓展

——基于词嵌入技术的文本分析的应用

摘要 在大数据时代的背景下，基于大数据的分析处理技术为以“数据驱动”的社会科学研究创造了契机。一方面，人们在网络世界贡献了大量包含思维、情感、观点的信息，为社会科学领域的各类研究提供了可及的数据来源。其中，文本这一类非结构化数据不仅是网络信息普遍的表现形式，也因其高度浓缩人类知识和广泛承载个体心理活动等抽象信息而具备重要的研究价值。另一方面，文本语言的表达天然具有规律性和共性特征，这让计算机学习和理解文本中的语义信息成为可能。以自然语言处理 (Natural Language Processing,NLP) 为核心的计算文本分析技术 (Computerized Text Analysis) 在近几年取得了长足发展，极大地推动了基于文本数据视角的计算社会科学研究。

其中，词嵌入 (Word Embeddings) 以其高效的词表征能力和强大的迁移学习能力在文本分析领域获得越来越多的关注。不同于传统的、依赖于词频统计因而难以捕捉不同语境下语义的文本分析路径，词嵌入技术依据词的分布规律，将文本词汇映射为高维向量空间中的点集，不仅实现了词的数字表征，还将全局文本语义信息融入了表征过程。因此，通过度量词向量之间的几何关系便能够刻画词汇之间在现实语义中的关系。与传统的社会科学研究方法对比，基于词嵌入技术的方法优势在于：第一，可以高效且自动化地处理大规模文本数据；第二，依据文本内在的分布规律学习和提取信息，结果更加客观；第三，能够利用外部信息和领域知识改进算法，可拓展性和重复性强；第四，可以实现对跨时间、跨文化文本中深层次文化信息的挖掘。本文回顾并梳理相关文献，发现词嵌入技术在社会学、语言学、心理学和政治学等领域得到了广泛应用，将现有研究总结为社会偏见、概念联想、语义演变、组织关系和个体判断机制五大主题，以期为不同领域的相关研究提供大数据方法支持。

随后，本文从实际使用出发，介绍了应用词嵌入技术展开社会科学分析的常见流程——词向量表征和相关性计算，并总结了学界广泛采用的语料库资源、模型选择、评估方法和测试任务集等，以期对研究人员提供清晰的流程指引。此外，本文针对词嵌入技术在实际应用中面临的文本数据的选择、中文文本的分词处理、单词语义信息的表征层次这三个方面的挑战归纳了相应的应对思路与方法。最后，基于词嵌入技术的强大适应能力，本文提出其对未来管理研究可能带来的独特贡献：第一，评估产品或品牌的市场表现和形象，探讨品牌依恋、品牌文化和品牌联想等话题；第二，基于组织内文本，分析组织内成员的心理及行为，探讨领导力、组织支持感和企业文化等话题；第三，利用词嵌入方法对中华古籍文本展开分析，探索中国传统的管理智慧。

关键词 词嵌入；自然语言处理；文本分析；社会科学；管理领域应用

The Development of Social Science Research Methods in the Era of Big Data: An Application of the Word Embeddings Technique

Ran Yaxuan[1], Li Zhiqiang[1], Liu Jiani[1], Zhang Yishi[2]

1. School of Business Administration, Zhongnan University of Economics and Law; 2. School of Management, Wuhan University of Technology

Abstract In the era of Big Data, computational technologies using the big data have provided a lot of opportunities for the “data-driven” social science researches. On the one hand, people have contributed a vast amount of information about what they know, feel and think on the internet, which provides an available data source for researches in social science. Among them, the unstructured text data not only is a general form of network information, but also has significant research value, since it contains human knowledge and a lot of abstract information such as individual mental activities. On the other hand, the composition of text has its own language rules and some universal characteristics, which makes it possible for computers to learn the semantic information in the text. The computational text analysis technology, as an application of Natural Language Processing (NLP), has made great progress in recent years, and the

analysis of large digitized corpora has already drawn widespread attentions in a range of social scientific works.
Word Embeddings, one of the computational text analysis techniques, has received increasing attention due to its representation capability and powerful transfer learning ability. This method transfers each word into a low-dimensional embedding space, in which each word in the corpus is represented geometrically as a vector, according to its distributions and usage in the corpus. The Word Embeddings technique not only realizes the digital representation of words, but also takes the context information into consideration. Therefore, the geometric relationship between different word vectors can describe the semantic relationship between words. Because words are located together in the semantic space if they appear in similar contexts, adjacent words in the vector space tend to share similar meanings. Based on this rationale, previous research has employed two metrics of word distance (i.e., cosine similarity and euclidean distance) and three paradigms of calculating construct distance (i.e., relative norm distance; Word Embeddings Association Test; word mover's distance). Comparing with traditional social science methods, the Word Embeddings technique can help to extract implicit and deep cultural information that might otherwise be difficult to acquire through self-reports or interviews, and has the following advantages: First, it can process large-scale text data efficiently and automatically. Second, it can learn and extract a variety of textual features according to the actual distributions of words, and the results are more objective. Third, external information and domain-specific knowledge can be integrated into the model training for algorithm optimization, so Word Embeddings models are highly scalable. Fourth, because a great deal of semantic and cultural information is available by examining the word vectors that surround the word of interest, Word Embeddings can shed light on questions regarding to cross-culture and temporal change. In light of Word Embeddings' outstanding advantages, an increasing amount of research has applied this method in sociology, linguistics, psychology and political science, etc. This paper therefore systematically reviewed related publications on this method, and the topics across multiple disciplines include social bias and stereotypes (e.g., gender bias and racial bias in news, song lyrics, books, etc), concept associations (e.g., social class culture in America; the relationship between Implicit belief and implicit attitude), semantic evolution (e.g., laws of semantic change; the changing meaning of "Equality" with social movements), polity relationship (e.g., ideological differences among political parties; international relations) and individual decision-making (e.g., individual mindset and cognitive biases in decision-making tasks; risk perception).
Subsequently, this paper presents an overview of the typical application procedure of the Word Embeddings, which includes two main steps---model training and relationship analysis. In order to provide researchers with a clear guideline, this paper also introduces the corpus resources that are widely-used, the basic steps of corpus pre-processing, criterions of model selection and parameter settings, principles and a series of test tasks about how to evaluate the performance of model and how to do the validition check and robustness check. In addition, this paper mentions two pre-trained Word Embedding models trained that can be used directly. It is worth noting that there are three challenges faced by the Word Embeddings technique, and this paper correspondingly proposes some solutions: First, the requirement of text data. For small-scale texts, pre-trained models can be used to address the problem of insufficient semantic information learning, or bootstrapping the texts to generate a larger corpus. Apart from that, background information such as textual context, culture, emotions and viewpoints of writers have impact on the semantics of word vectors, these factors should be considered to prevent biased conclusions. Second, the difficulty of word segmentation. Word segmentation is still hard for texts containing professional phrases or ancient books. In recent years, the text preprocessing methods have been optimized so that it can process out-vocabulary words. Last, the limitation in meaning representation. Word vectors can not represent topic information beyond word level very well. Nowadays, a large number of scholars manage to integrate some domain knowledge into the learning algorithm of Word Embeddings. In addition, Word Embeddings is not good as capturing the fine-grained semantics of words as 'close reading' unless combining some supplementary information.
Finally, based on the strong adaptability of word embedding technique, future researches in management can further focus on the following three aspects: First, words are part of almost every marketplace interaction. Online reviews, customer service calls, press releases, marketing com- munications, and other interactions create a wealth of textual data. The Word Embeddings technique may also be valuable for evaluating the company's market performance and formulating marketing strategies. Also, this method can help address challenges of measuring consumers' perceptions on a product and a brand, which is useful for targeting new products and overseas market. Second, in the context of organizational and strategic management, researchers could use Word Embeddings to analyze the psychology and behavior of organizational members through organizational texts (e.g., conference transcripts, employee comments, leadership speech texts), and explore topics such as leadership, organizational support, and corporate culture. Furthermore, we look forward to using Word Embeddings to analyze Chinese ancient books (e.g., The Twenty-four Histories), and to mine the traditional Chinese management wisdom.
Key Words Word Embeddings; Natural Language Processing; Text Analysis; Social Science; Applications in Management Field