Word2Vec 在文本相似度计算中的应用

黄云磊

浙江清桦智控科技有限公司　浙江杭州

【摘要】随着文本数据激增，高效计算文本相似度成为 NLP 关键任务。Word2Vec 作为词嵌入技术，将词语映射为低维稠密向量，捕捉语义信息，革新文本相似度计算。其核心包括连续词袋与跳字模型，通过负采样、层序 softmax 等算法优化性能。结合余弦相似度、欧氏距离等度量方法，已广泛应用于信息检索、文本分类等领域。未来，Word2Vec 与新兴技术融合，应用前景广阔。

【关键词】Word2Vec；文本相似度；词嵌入；自然语言处理；语义计算

【收稿日期】2025 年 5 月 16 日　　【出刊日期】2025 年 6 月 7 日　　【DOI】10.12208/j.aics.20250022

The application of Word2Vec in text similarity calculation

Yunlei Huang

Zhejiang Qinghua Intelligent Control Technology Co., Ltd, Hangzhou, Zhejiang

【**Abstract**】With the surge in text data, efficiently calculating text similarity has become a key task in Natural Language Processing (NLP). Word2Vec, a word embedding technique, maps words into low-dimensional dense vectors to capture semantic information, revolutionizing text similarity calculation. Its core components include continuous bag-of-words and skip-word models, with performance optimized through techniques such as negative sampling and hierarchical softmax. By integrating metrics like cosine similarity and Euclidean distance, Word2Vec has found widespread application in information retrieval and text classification. Looking ahead, the integration of Word2Vec with emerging technologies promises a broad range of applications.

【**Keywords**】Word2Vec; Text similarity; Word embedding; Natural language processing; Semantic computation

引言

在信息时代，文本数据海量涌现，如何快速准确地计算文本间相似度，挖掘文本蕴含的语义信息，成为自然语言处理领域亟待解决的重要问题。传统文本相似度计算方法在处理语义信息时存在局限性，难以有效捕捉词语间复杂语义关系。Word2Vec 的出现为文本相似度计算带来突破，其通过训练将词语转化为向量表示，能够反映词语语义相似性。探究 Word2Vec 在文本相似度计算中的应用，对推动自然语言处理技术发展、提高文本处理效率具有重要意义。

1 模型原理剖析

Word2Vec 作为自然语言处理领域的基石性技术，其核心思想源于对人类语言认知机制的深度模拟。传统的独热编码方式将每个词语表示为高维稀疏向量，虽然能够实现词语的数字化表达，但完全割裂了词语间的语义联系，无法捕捉语言的内在逻辑。Word2Vec 突破性地采用分布式表示，将词语映射到低维稠密向量空间，使得语义相似的词语在向量空间中形成“语义簇”。这种表示方式不仅大幅降低了数据维度，更通过向量间的几何关系直观反映词语间的语义关联，为自然语言处理任务搭建起语义理解的桥梁。

在具体实现层面，连续词袋模型（CBOW）与跳字模型（Skip-Gram）构成了 Word2Vec 的双轨架构。CBOW 模型基于“上下文决定词义”的语言学假设，通过聚合目标词语周围的上下文向量，预测中心词的概率分布。在句子“我喜欢阅读书籍”中，CBOW 模型会将“我”“喜欢”“阅读”“书籍”作为输入，尝试预测中心词“喜欢”。这种从语境到核心的预测逻辑，本质上是对人类语言理解过程的简化模拟——我们在阅读时，往往通过上下文信息推断生词含义。而 Skip-Gram 模型则反向操作，以目标词语为出发点，预测其上下文词语的概率分布，更侧重于捕捉词语在不

同语境下的语义延伸[1]。两种模型虽方向不同，但都通过神经网络的多层非线性变换，将词语映射为蕴含语义信息的向量表示。

训练过程中，Word2Vec 采用的梯度下降算法通过不断调整神经网络参数，最小化预测词语与真实词语间的误差。这个迭代优化的过程，本质上是让模型学习海量文本中词语共现的统计规律。当模型在大规模语料库上完成训练后，每个词语对应的向量不再是孤立的数字组合，而是凝聚了丰富的语义知识[2]。“国王”“皇后”“王子”“公主”等词语的向量在空间中会形成紧密的聚类，且“国王-皇后”与“王子-公主”的向量差值呈现出相似的方向，这种数学特性直观地反映了词语间的语义类比关系，为后续文本相似度计算奠定了坚实的语义基础。

2 算法优化路径

面对互联网时代海量的文本数据，Word2Vec 的原始训练算法面临着计算效率与内存占用的双重挑战。传统的 softmax 函数在计算词表中所有词语的概率分布时，需要对每个训练样本进行全量计算，这在百万级甚至亿级词表规模下几乎无法实现。为突破这一瓶颈，负采样与层序 softmax（HierarchicalSoftmax）两种优化策略应运而生，它们从不同角度对计算复杂度进行降维，使 Word2Vec 能够在普通硬件上高效处理大规模语料。

负采样策略的核心理念在于采用“以少代多”的抽样思想。在传统的训练过程中，模型需要对整个词表进行详尽的计算以更新其参数，这不仅耗时而且计算量巨大。然而，负采样技术的引入，通过随机选取少量与目标词语无关的“负样本”，使得模型仅需针对这些样本以及正样本进行参数更新。这种方法基于一个关键的假设：在大规模的语料库中，词语之间的共现关系是相对稀疏的，即便只使用少量的负样本，也足以捕捉到词语间的区分特征。举例来说，在训练“苹果”这个词的向量表示时，负采样可能会随机选取“汽车”、“天空”等与“苹果”无关的词语作为负样本[3]。通过这样的对比学习，模型能够强化对“苹果”语义边界的认知，从而更准确地理解其含义。采用负采样技术后，模型在训练过程中无需遍历整个词表，这使得计算量大幅度减少，计算效率显著提升。即便是在减少了计算量的情况下，模型依然能够保留词语间的关键语义信息，确保了模型的准确性和效率。

层序 softmax 则另辟蹊径，利用二叉树结构对词表进行层次化组织。在这棵树中，每个叶节点对应一个词语，根节点到叶节点的路径构成了词语的概率计算路径。模型通过判断路径上节点的概率分布，逐步定位到目标词语，将原本的全局概率计算转化为路径搜索问题。这种树形结构的优势在于，每次计算仅需遍历从根节点到目标节点的路径，计算复杂度从线性降低为对数级[4]。对于包含 100 万个词语的词表，传统 softmax 需要进行百万次计算，而层序 softmax 通过二叉树结构，最多仅需进行约 20 次计算（$\log_2 1000000 \approx 20$）。这种优化不仅大幅提升了训练速度，还显著降低了内存占用，使得 Word2Vec 能够高效处理大规模文本数据，为高质量词向量的生成提供了技术保障。

3 计算方法实现

基于 Word2Vec 生成的词向量，文本相似度计算进入了语义度量的新阶段。传统基于词频的相似度计算方法，如 TF-IDF，仅关注词语的统计分布，无法捕捉语义层面的关联。而词向量的引入，使得文本相似度计算能够深入到语义空间，通过向量间的几何关系量化文本的语义相似程度。在众多计算方法中，余弦相似度与欧氏距离成为最常用的度量工具，它们从不同维度刻画向量间的关系，为文本相似度评估提供了灵活的选择。

余弦相似度是一种衡量两个向量在方向上一致性的有效方法，它通过计算这两个向量夹角的余弦值来实现。在高维的语义空间中，向量的方向往往比它们的长度更能准确地反映出语义的本质。即便是在两个文本中，词语的数量可能存在较大的差异，只要它们的核心语义相似，那么这两个文本所对应的向量方向就会表现出高度的一致性。当我们考察“人工智能的发展”与“机器学习的进步”这两个文本时，尽管它们使用了不同的词汇，但它们的核心语义都聚焦于智能技术的不断进步和演进，这两个文本的向量夹角的余弦值将会非常接近于 1[5]。这种计算方式的一个显著优势在于它能够摆脱文本长度的限制，更加专注于语义的匹配程度，这使得它特别适合于短文本相似度的计算。

欧氏距离则从空间几何的角度，直接度量两个向量在高维空间中的绝对距离。与余弦相似度不同，欧氏距离同时考虑向量的方向与长度，更强调文本在语义空间中的绝对位置关系。在评估两篇论述同一主题但详略不同的文章时，欧氏距离能够通过向量长度差异反映文本信息量的差别，而余弦相似度可能因方向一致而高估其相似度[6]。实际应用中，常将欧氏距离与余弦相似度结合使用：先用余弦相似度进行快速筛选，定位语义相关的文本；再用欧氏距离进行精确排序，区分

语义相似但信息量不同的文本。对于由多个词语向量构成的文本，可通过加权平均、主成分分析等方法将词语向量聚合为文本向量，进一步拓展了相似度计算的应用场景。

4 领域应用实践

Word2Vec 与文本相似度计算的结合，为自然语言处理技术在多领域的落地应用提供了核心驱动力。在信息检索领域，传统的关键词匹配方式常因同义词、近义词的存在导致检索结果偏离用户需求。而基于 Word2Vec 的相似度计算，能够将用户查询与文档映射到同一语义空间，通过向量匹配找到语义相关的文档。当用户搜索“智能手表的功能”时，系统不仅能返回包含“智能手表”的文档，还能检索到论述“可穿戴智能设备特性”的相关内容，显著提升检索的召回率与准确率。这种语义级检索能力，使得搜索引擎能够更好地理解用户意图，提供更精准的信息服务。

在文本分类任务中，Word2Vec 的应用有效解决了传统方法对语义理解不足的问题。传统分类模型依赖于人工提取的特征或词频统计，难以捕捉文本的深层语义。而基于词向量的相似度计算，能够将待分类文本与各类别模板进行语义匹配，自动判断文本所属类别[7]。在新闻分类中，系统可通过计算新闻文本与“科技”“财经”“体育”等类别向量的相似度，快速完成文本归类。这种方法不仅减少了人工标注的成本，还能适应新出现的文本模式，提升分类模型的泛化能力。在情感分析、垃圾邮件识别等任务中，基于相似度的文本分类方法同样展现出强大的适应性，能够有效处理语义模糊、表达多样的文本数据。

在机器翻译领域，Word2Vec 与文本相似度计算为跨语言语义对齐提供了新的解决方案。传统的基于短语或统计的翻译模型，常因语言结构差异导致翻译质量不稳定。而通过计算源语言文本与目标语言文本的语义相似度，系统能够找到语义匹配的翻译候选，辅助生成更自然流畅的译文。在将“他在公园散步”译为英文时，模型可通过相似度计算，从语料库中检索到“heiswalkinginthepark”等语义相近的表达，避免逐词翻译导致的语法错误。这种基于语义匹配的翻译策略，不仅提升了翻译的准确性，还能处理习语、隐喻等复杂语言现象，推动机器翻译技术向更高水平发展[8]。在跨语言信息检索、多语言文本聚类等任务中，Word2Vec 的应用同样展现出巨大潜力，为自然语言处理的多语言应用开辟了新的道路。

5 结语

Word2Vec 在文本相似度计算中展现出强大优势，通过创新的词向量表示与优化算法，有效解决了传统方法在语义处理上的不足，在多个领域取得良好应用成果。未来，随着自然语言处理技术不断发展，Word2Vec 有望与预训练模型、知识图谱等新兴技术融合，进一步提升文本相似度计算的准确性与效率。在跨语言文本处理、情感分析等更多场景的应用探索，也将为 Word2Vec 在文本相似度计算中的发展带来新机遇与挑战。

参考文献

[1] 王延敏,张爱儒.基于 LDA 与 Word2vec 主题模型的草畜平衡研究主题演化与热点主题识别[J/OL].草业科学,1-29[2025-06-23].

[2] 林伟鸿,贺超波,呼增.基于 EWord2Vec-TextCNN-SE 的食品安全新闻文本分类[J].现代计算机,2024,30(21):156-160.

[3] 任伟建,徐海杰,康朝海,等.基于 Word2vec 与注意力机制的情感分析研究[J].计算机与数字工程,2024,52(10):2991-2995+3147.

[4] 陈宇.Word2Vec 新闻推荐系统设计与实现——基于 Attention 机制与 Embedding 优化[J].情报探索,2024,(10):88-96.

[5] 孙晶.分类数据的 Word2Vec 与 Jaccard 相似度聚类方法的比较分析[J].软件,2024,45(09):49-51.

[6] 余如辰.基于 Word2Vec 的我国青少年体质健康研究的可视化分析[J].文体用品与科技,2024,(16):90-93.

[7] 王捷,周迪,左洪福,等.结合 Word2vec 和 BiLSTM 的民航非计划事件分析方法[J].合肥工业大学学报(自然科学版),2024,47(07):917-924.

[8] 张昱,冯亚寒,丁千惠.融合 Word2Vec 词嵌入的多核卷积神经网络音乐歌词多情感分类方法[J].科学技术与工程,2024,24(20):8598-8605.

版权声明：©2025 作者与开放获取期刊研究中心（OAJRC）所有。本文章按照知识共享署名许可条款发表。
http://creativecommons.org/licenses/by/4.0/

OPEN ACCESS