DOI：10.16353/j.cnki.1000-7490.2019.11.021

● 黄文彬，车尚锟（北京大学信息管理系，北京 100871）

计算文本相似度的方法体系与应用分析

摘 要：［目的/意义］文本间的相似度是信息检索、文档检测和文本挖掘等任务核心参考的指标之一。梳理现有计算文本相似度的方法、分类体系及应用，有助于研究人员选择合适的计算方法提高特定场景应用的性能。［方法/过程］文章将算法利用文本语义信息的程度、基础语义信息类型、模型类型以及关联关系类型作为划分依据构建方法体系，并从原理和应用上梳理算法间的异同。［结果/结论］将文本相似度计算方法分为无语义信息、基于浅层语义信息、基于深层语义信息三个大类，对参考的语义信息、算法的基本原理和该类的典型应用做了探索分析。［创新/价值］使文本相似度计算方法具有更清晰和完整的体系，使研究人员能更好地区分相似度计算方法间的计算需求与应用场景的差异。

关键词：文本挖掘；文本相似度；分类体系；语义信息；应用

Methodological System and Application Scenarios on Text Similarity Calculation

Abstract: [Purpose/significance] Text similarity calculation is a core technology in information retrieval, document detection and text mining. Researchers can improve the performance in their applications by using an appropriate calculating text similarity technologies according to the review of methodologies, classification system, and application scenarios. [Method/process] In the paper, the methodological system is constructed based on four characteristics of an algorithm, which are the degree of text semantics, and the type of semantic information, its mathematical model, and relationship. The similarities and differences among methods are concluded from their principle and application. [Result/conclusion] We divide the methods of text similarity calculation into three categories, no semantic information, shallow semantic information and deep semantic information. And the semantic information, the algorithmic principle, and the application scenarios in each categories are also explored and analyzed. [Originality/value] As a result, the methodological system of text similarity calculation is more clear and complete, and it makes researchers to better distinguish and apply an algorithm to their application scenario based on the computational requirements.

Keywords: text mining; text similarity; classification system; semantic information; application

信息数据多数以文本形态大量存在在各个应用领域，识别与寻找相关资源费时费力，信息自动抽取和文本检索成为协助人们最重要的应用技术之一，而文本相似度被用来衡量文本间的差异和共性，它是这些技术任务的核心环节[1-2]。近年来，文本相似度主要被应用在词义消歧[3]、抽取式自动摘要[4]、机器翻译自动评估[5]、数据库的模式匹配及语义异构问题[6]、论文抄袭检测[7]等研究议题，并且随着深度学习、本体与语义网等新技术的发展，计算文本相似度的方法也产生新的进展。

梁艳艳从信息科学的角度提供文本相似度在工程上的表达：如果文档 M 和 N 具有某些相似的或者相同的属性，则判定 M 与 N 相似[8]。而 Lin[9] 以信息论（信息熵）的观点论述了文本的共性、差异性及其相似性之间的关系，当共性越大而差异越小时文本具有更大的相似性。因此字符串匹配和词语匹配成为计算文本相似度的基础方法，随着自然语言处理技术的发展，词干提取、停用词移除、词性标注等方法被引入文本相似度计算中，再结合各种加权和正则化策略后，结合深入到语义层面的信息形成目前常用的方法。目前，存在 90 种以上基于内容计算文本相似度的方法[10]，其所构成的混合算法更无法计数。

因此，基于上述众多的应用以及大量文本相似度计算的方法，建立合理的方法体系并进行具体的应用分析，对于理解其算法的内在联系和针对特定应用场景的优化和改进具有重要意义。

1 文本相似度的计算方法体系

文本相似度主要被划分为词汇（Lexically）相似度和语义（Semantically）相似度两大类[11]，前者主要是基于字符串匹配的相似度比较，属于不包含语义信息的算法，后者则包括内容、语法、上下文语境等多种层面信息的算法。这种划分方式从“有无语义”信息的角度对所有概念进行划分，同时“无语义”信息到“有语义”信息的算法的过渡是符合算法发展的时间顺序的。然而，语义信息包含的范围比较广泛，从词语本身和词语的上下文到词

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

语、句子、文档的生成模型，再延伸到词语与语料库其他词语的上下位类和同义反义关系。经过大量的文献调研发现有语义的文本相似度算法具有从基于浅层语义到基于深层语义的发展过程，各个层级的算法利用语义信息的递增使得计算文本相似度的效果也越好。深入探讨这些语义关系后，笔者把“有语义”信息的算法切分成“基于浅层语义”信息和“基于深层语义”信息的算法，基于浅层语义信息的算法主要利用文本自身的信息进行信息提取，而基于深层语义信息的算法则大量使用已构建好的语义信息（比如引入外部知识库）以及复杂算法来关联文中的语义关系。二者主要区别在于语义信息的来源和获取难度。因此，笔者构建的文本相似度计算方法体系见图 1：首先将利用语义信息的程度作为一级分类的依据，并为了避免类别间出现重合，下设多级分类进一步细化算法的种类，利用所使用的语义类型或关联关系划分各种算法，最后以模型类型作为三级分类依据，使得类内的算法相似性更高。

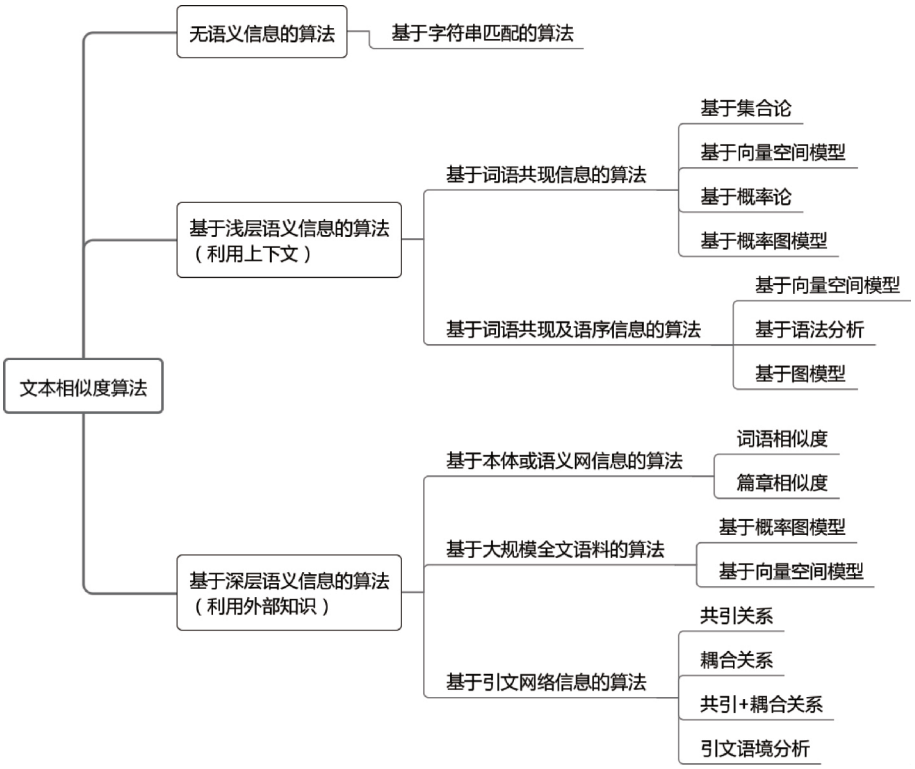

图 1　文本相似度的计算方法体系

本文提出基于语义信息利用程度的分类体系既尊重了 Gomaa[11]、陈二静[12]等研究的词汇相似度和语义相似度的划分观点，又将文本相似度计算方法结合 Mahmood[1] 的分类思想，按模型类型、基础语义信息类型和关联关系类型进行组织，具有明确的逻辑关系。研究人员可以利用此方法体系直观地理解文本相似度计算方法的发展思路和趋势，并根据已有文本信息的种类和数量，从这一体系中选择合适的算法进行运用。接下来分别对无语义信息算法、基于浅层语义信息算法以及基于深层语义信息算法进行深入的描述与分析。

2　无语义信息的算法

无语义信息的相似度计算方法特指基于字符串匹配的算法。字符串匹配属于文本处理领域最基本的要素，Mahmood[1] 总结大量现存的研究方法发现基于字符串匹配的相似度算法不会被直接应用于文档相似度或句子相似度等文本级别的相似度计算中，而是作为各类文本处理算法的底层实现算法。无语义信息的代表性相似度算法见表 1，这些算法主要基于字符串的匹配或搜索操作，其中算法时间复杂度的字母 m 和 n 表示算法涉及的字符串长度。

此类算法主要都集中在字符串的搜索、匹配、重复检测或差异检测上。针对具体应用，根据字符串长度、模式串数量等的具体要求分成两类场景分别说明不同算法的适用性：

1）用于字符串匹配方面，Levenshtein 编辑距离只适用于一般的拼写检查，而 Jaro-Winkler 算法的赋分方式更适合人名等短字符串的重复检测。LCS 采用求公共序列的方式更适合版本控制系统。

2）用于字符串检索方面，KMP 算法适用于子串查找，BM 系列算法主要针对目标串进行处理而适用文本编辑器内的搜索匹配，并且 BM 搜索算法比 KMP 快 3 倍以上，更适合较长文本的搜索匹配。Aho-Corasick 算法复杂度与模式字符串的数量和长度都无关，因此更适合于操作系统级的文件搜索。总体而言，这类算法仅利用了字符串级别的信息，缺少词语、句子的概念，因此不适用于计算贴近人类理解的语义相似度。

3　基于浅层语义信息的算法

浅层语义信息是指组成文本的词语本身、词语的顺序、词语的上下文搭配以及词语的形态变化等基本的信息。基于浅层语义信息的相似度计算方法不引入外部知

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表1　无语义信息的相似度计算方法

类别	算法名称	核心原理	时间复杂度	应用
基于字符串匹配的算法	Levenshtein（编辑距离）	两个字串间由一个转成另一个所需的最少编辑操作[13]次数	可由 $O(m*n)$ 改进为 $O(n)$	拼写检查；DNA 相似度匹配；UNIX 下的 diff 操作
	Jaro-Winkler	在编辑距离基础上给予起始部分相同的字符串更高分数	可优化至 $O(m)$	主要用于短字符串的重复记录检测如人名是否重复
	Hamming（汉明距离）	表示两个等长字符串在对应位置上不同字符的数目	$O(n)$	主要用于通信领域的误差检测与信号处理
	Longest Common Sequence（LCS）	比较多个序列的最长公共子序列，常使用动态规划实现	可由 $O(n^2)$ 优化至 $O(n \log n)$	GitHub 等版本控制系统中用于比较不同版本的改动关系[14]
	KMP	字符快速匹配算法	$O(m+n)$	主要用于子串查找与替换等
	BM（Boyer-Moore）	字符串搜索算法，只对搜索目标字符串进行预处理	平均 $O(n)$，最坏 $O(mn)$	常用文本编辑器中的搜索匹配采用此算法
	Boyer-Moore-Horspool[15]	比 BM 算法的内部表更小，时空复杂度和匹配复杂度降低	同 BM 算法，但初始化速度更快	用于文本的匹配式搜索，如文本编辑器中搜索功能
	Aho-Corasick	用有限状态自动机结构来处理模式集合中的所有字符串	$O(n)$，与模式串的数量和长度无关	UNIX 系统中文件搜索指令“fgrep”的基础

识，仅使用文本本身信息进行比较并计算相似度。根据是否利用词语的语序信息将算法分成两个大类：第一类方法仅使用词语共现信息，而第二类方法同时利用词语的共现信息和语序信息进行计算。这种划分方式符合算法的演进发展顺序，每一类别的代表性算法见表2：其中时间复杂度中的 m、n 为两字符串的长度，对于需要建立语义空间的算法，n 特指空间中不重复的词语总数。

基于词语共现信息的算法又称为基于词袋模型（Bag of Words，BOW）的算法，主要切分出每篇文档的词语并将其放入同集合中，利用所构成的词语集合来表示每篇文档的特征信息。基于词语共现和语序信息的算法主要在 BOW 方法的基础上将词语出现的先后顺序和词语之间的间隔距离等信息加入相似度计算。其中基于向量空间模型的方法是在 BOW 方法上做简单的扩展，例如 HAL 算法。基于语法分析的方法利用了自然语言处理中复杂的句法依存和语义依存分析等方法，在依存结构的基础上计算语义相似度。基于图模型的方法通过图相关的算法抽取关键节点，得到能够表征文本语义的关键信息进而可以计算文本相似度。此类算法不引入外部数据源，在计算相似度时需要使用的外部资源较少。

针对此类算法的场景适用性而言，基于集合论和向量空间模型的方法仅能利用文本的基本统计信息（例如，词语数量、频率、逆文档频率和词语组合关系等），导致无法精确获取主题或结构等更深入的信息，使得其应用集中在文本挖掘领域的基础工作（例如，文本集合的相似性计算等），常作为其他复杂算法的底层实现。基于概率论或概率图模型的方法能够挖掘词语、主题、文档等多个级别之间的概率转移或生成关系，经常作为主题提取、主题相似度计算、主题聚类和信息抽取等工作的实践方法。基于语法分析和图模型的方法能够挖掘文档或句子间的结构关系，进而识别文档主题或词语间的依存与搭配关系，因此常被多主题识别、关键词抽取、语义消歧和篇章理解等对语义要求更大的复杂任务所采用。最后，针对语义要求较高的相似度计算而言，更精准和较低复杂度的自然语言处理技术中语义图分析算法将成为主流，经常应用在信息检索、主题识别、信息抽取等场景，且优化概率图模型（如 LDA、LSA）将成为未来技术演进的趋势。

4　基于深层语义信息的算法

人类常利用已有的知识体系理解或分析文本的词语所包含的深层意义与关联关系，而许多研究人员认为机器仅从文本本身数据获取准确的信息较不现实，为了提高机器理解文本的能力，则需要引入足够的外部语料或知识库识别词语间隐含的关系，或引入合适的本体或语义网络挖掘词语间更深层的语义关系。本文将此类方法统称为基于深层语义信息的算法。根据不同模型类型又可将上述每类算法划分成数个子类，其划分的体系与每个子类的代表算法见表3。

本体和语义网络一般被用来计算词语间的语义相似度，基于本体和语义网络的相关算法主要利用词语在语义网中的最短路径、或在本体树结构中的最近祖先节点、或词语定义的重合度、词语在词语集合中出现的概率等特征信息。目前，许多的研究常使用 HowNet[44]、同义词词林[45]等中文语义网，以及英文的 WordNet 作为外部知识库[46]。

基于大规模全文语料的方法利用大量外部数据进行词语的特征发现与相似度计算，依据外部语料数据类型划分成概率图模型和向量空间模型的子类方法。概率图模型一方面可利用大规模语料库（例如，网页、检索词等）计算词语的互信息或共现的词语列表来获得两词语的相似

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 2　基于浅层语义信息的算法

类别	模型	算法名称	核心原理	时间复杂度	应用
基于词语共现信息的算法（词袋模型）	基于集合论	Dice's Coefficient	两集合交集去除以两集合的大小之和并乘 2	取决于求交集的算法，直接匹配为 $O(mn)$，最低可优化到 $O(n)$	都用于计算样本集合间的相似度，Dice 和 Jaccard 系数早期用于比较物种间的相似度，Matching 系数可用于购物篮分析
		Jaccard similarity	利用交集的模除以并集的模		
		Matching coefficient	两集合匹配属性数除以属性总数		
	基于向量空间模型	Euclidean istance	M 维空间中两个点之间的真实距离	$O(n)$	常用于如神经网络等研究中基本的距离计算
		Manhattan distance	欧式空间内两点对轴的投影距离总和	$O(n)$	
		cosine-similarity	是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量	$O(n)$	信息检索早期研究的常用方法
		soft-cosine-similarity [16]	每两个词语相似度作为新维度加入原有向量。从 N 维扩展到了 $N+N*N$ 维	$O(n^2)$	信息检索、相似度计算
	基于概率论	BM25	利用词在文档中相关度、词在查询关键字中的相关度以及词的权重信息	$O(n^2)$	广泛应用于信息检索 [17] 和文档聚类 [18]
		KL 散度	计算文档间词语概率形成向量的分布差异	分为求概率分布和计算散度，后者为 $O(n)$	度量使用基于理论分布 Q 的编码来编码真实分布 P 的样本平均所需额外比特数
		JS 散度	KL 散度的改进，JS 散度是对称的，可以用于衡量两种不同分布之间的差异		
	基于概率图模型	LSA	提取文档潜在语义表示：通过矩阵奇异值分解（SVD）选取进行降维	主要代价在 SVD 上，复杂度高达 $O(n^3)$	用于文档索引、文本聚类、信息抽取等
		P-LSA [19]	引入主题层，采用最大期望算法来训练主题，形成概率潜在语义索引		
		LDA	LDA 有 3 层生成式贝叶斯网络结构：为词语—主题—文档 3 层构成	迭代 n 次，共 m 篇文档，每篇约 N 个词，主题数为 k，复杂度为 $O(nmNk)$	文档主题生成，P-LDA 实现文档多标签判定，S-LDA 实现文档类别判定和连续数据的回归分析
		P-LDA [20]	将文档类别映射成为多个主题的组合		
		S-LDA [21]	将文档类别标记或对应的连续变量映射为由主题混合方式产生的响应变量		
基于词语共现与语序信息的算法	基于向量空间模型	HAL [22]	使用词语共现和语序信息来建立高维语义空间，每个词语都对应了一个点，用空间中求距离的方法来计算词语之间的关系	$O(n^2)$，分为建立语义空间和求向量相似度	适用于具有大型语料库的文本相似度计算
	基于语法分析	TED [23]	将段落切分成多个 N-gram 片段，每个片段表示为句法依存树，通过依存树之间树编辑距离倒数的累加得到段落相似度	取决于 NLP 中的句法依存和语义依存分析算法，一般为 $O(n^3)$	算法复杂度很高，但是能提取到文本中的语义信息和语义之间的转移关系，适用于对文本语义相似性的要求高，例如篇章理解、语义消歧等复杂任务
		基于依存图分析的算法	将句子各个语言单位之间的语义关联以依存结构呈现，在此基础上计算语义相似度		
	基于图模型	NGG [24]	文本解析为 N-gram 作为节点，边用来表示 N-gram 之间的位置关系，关系频率作为边的权重。通过计算 Z-score 识别主题	取决于图分析算法的复杂度，一般为 $O(n^3)$	

度；另一方面可通过机器学习模型对目标词以及更复杂的上下文之间关系进行建模并形成词语的特征向量（称为词向量，Word Embedding），直接利用词向量来计算文本的相似度，或者在此基础上将段落句子转化成语义向量后再计算整段文本相似度。向量空间模型利用外部的概念知识库（如 Wikipedia）建立词语构成的向量空间，利用概念知识库中词语之间的关联关系将词语表示为多个相关词语组成的向量，通过计算向量间的相似度得到词语相似度。

基于引文网络信息的方法主要使用共引或耦合等关系来优化文章间词语关系进而计算文章相似度。共引关系是指不同文章共同引用同文章，耦合关系则指不同文章被同文章所引用；在此基础上加入引文标记及其周围文本的内容信息产生了“引文语境分析”（Citation Context Analysis，CCA）的概念 [43]。

所利用的外部信息类型的差异导致不同算法的适用场景有较大差异。基于本体或语义网的算法较能深入到知识层面，但多领域的本体提取和语义网的构建难度很大，因此仅能在特定领域内的主题识别、信息抽取和信息检索等任务上获得比其他类型算法优越的效果。基于大规模全文语料的算法主要是在已有大规模全文语料且具有充足的计算资源为前提条件的背景下才具有优势，而又因为不同的

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表3　基于深层语义信息的算法

类别	模型	算法名称或作者		核心原理	应用
基于本体或语义网信息的算法	词语相似	C. Leacock 等[25]		sim = - log(length/2* D) ，length 是两个概念之间的距离，D 是分类系统的最大深度	主要用于有特定领域的语义网或本体的情况，一般都针对特定领域的文本相似度计算、信息检索、信息抽取等
		Z. Wu 和 M. Palme[26]		LCS 为最近共同祖先，Depth（ ）为深度，sim = 2* Depth(LCS) /(Depth(Concept1) + Depth(Concept2))	
		M. E. Lesk[27]		指词语在语义网或词典上的定义之间的重叠度	
		P. Resnik[28]		sim = IC(LCS) ，IC(c) = - logP(c) ，P(c) 是概念 c 出现的概率	
		D. Lin[3]		sim = 2* IC(LCS) /(IC(Concept1) + IC(Concept2))	
		J. J. Jiang 和 D. W. Conrath[29]		sim = 1 /(IC(Concept1) + IC(Concept2) - 2* IC(LCS))	
	语篇相似	Walk Through 算法[30]		对于第一篇文档中每个词，在第二篇文档中找到与它最相似的词语求其在语义网内的相似度取值并累加	
基于大规模全文语料的算法	概率图模型	PMI 系列算法	PMI-IR[31]	利用每个网页中两个特定词语共现的频率来判定其相似度	常用于情感倾向识别领域
			SCO-PMI[32]	从大规模语料库中使用计算互信息的办法找出与两个特定词语经常共现的两个词语列表以计算两个词语间相似度	
		Word Embedding 系列算法	Word Embedding[33]	利用神经网络语言模型对目标词以及更复杂的上下文之间关系进行建模，得到词向量	适用于具有大规模语料库的情形，从而通过神经网络模型得到准确的 embedding 表达，用于篇章理解等复杂任务
			Joint Word Embedding[34]	用特定领域的数据与通用的数据分别训练然后将模型进行融合，在针对特定领域的文本进行相似度比较时取得了效果的提升	
			Para2vec[35]	在 Word Embedding 网络结构上增加段落向量，对段落进行向量化表示	
		QACNN	QACNN[36]	将问答系统引入相似度计算中，使用基于问答（Query-Answer）的 CNN 算法来得到文本相似度	
	向量空间模型	ESA 系列算法	ESA[37]	Explicit Semantic Analysis 利用 Wikipedia 将词语表达为高维向量，然后用余弦相似度求词语相似度	用于有大规模概念知识库作为基础的文本分类等任务
			CL-ESA[38]	ESA 多语言泛化的版本，利用基于 Wikipedia 的方法将文本表示成与语言无关的概念向量	
基于引文网络信息的算法	共引关系	SimRank[39]		两文献 A 与 B 相似度等于：阻尼系数 C* 同时引用 AB 的文献数/（引用 A 的文献数* 引用 B 的文献数）	用于只依赖引文网络信息的论文主题的相似性计算
	耦合关系	rvs-SimRank[40]		两文献 A 与 B 相似度等于：阻尼系数 C* AB 共同引用的文献数/A 和 B 引用的文献总数	
	共引+耦合关系	P-Rank[41]		结合共引关系和耦合关系，同时考虑入度和出度，然后从中心节点开始递归整个图，求出所有子节点两两之间相似度	
		C-Rank[42]		为了连接年代差异久远的论文，加入了一个中间节点的概念，表示被 p 引用同时引用 q	
	引文语境	引文语境分析[43]		它在传统引文分析的基础上，基于引证标记及其周围的文本内容进行引文分析工作	论文主题、方法、结论等相似度研究

语料使得应用场景不同：例如 PMI 系列算法利用网络知识百科提取主题概念在文本分类、自动摘要生成等领域应用。而 ESA 系列算法利用网页搜索引擎挖掘到词语的共现信息来进行情感倾向分析、主观性分析等应用。基于引文网络分析的算法多数时候仅在文本数据具有引文关系的场景使用，引文语境分析方法通过利用引用的语境信息在文本的主题、方法或结论等具体模块的相似度计算等任务中可以提供较佳的效果。

总体来说，基于本体和语义网络和基于大规模全文语料的方法在语义上有较高的准确性，但对可利用的信息数量与质量的要求比较高（比如需要语义网、大规模有标注训练语料等），使得在实际场景中往往受到这些资源不足的限制而难以实践应用。因此，自动构建语义网或领域本体，以及通过小规模标注语料实现准确性更高的半监督学习方法将是未来重要的研究方向。同时，利用本体知识库来获得词语之间的深层语义信息并结合其他类型算法形成“混合类型方法”也是未来的研究趋势。例如，ADW 算法使用基于 WordNet 的随机游走模型来产生语义标签，并进而用余弦相似度和 Top-k Jaccard 等方法来计算语义标签的相似度[47]。Joint Word Embedding 方法利用 AAN Corpus 数据人工构建领域本体形成词语向量来提高词嵌入的准确性[34]。

5　讨论与展望

文本相似度计算对信息检索、信息抽取、自动摘要等多个领域具有重要意义，对其方法建立完善的研究体系并

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

进行深入分析有利于该领域算法被合理应用与发展。笔者探索近年来文本相似度计算的重要文献，并对算法进行较全面的总结与分析，借鉴前人研究的思想，深入理解算法利用语义信息程度、基础语义信息类型、模型类型以及关联关系类型，构建出较为清晰的文本相似度计算方法体系，并对各类的算法进行计算复杂度与应用分析。该体系利于研究者直观了解文本相似度算法的类型、演进以及原理的异同，便于开发者通过当前数据的情况和实际需求，从中选择合适的相似度算法，提高相关研究或应用的质量与效率。

依据本文的总结分析，计算文本相似度已经成为文本挖掘复杂且重要的任务，越来越多的研究人员提出新的文本建模方法。例如，晋耀红[48]提出的基于文本形式化的语境框架、Crockett 等[49]提出的模糊逻辑（Fuzzy Logic）的概念等。利用已有的语义网、引文网络、深度学习算法等方法进行组合搭配，以及新的文本建模模型将成为未来文本挖掘方法的研究趋势。□

参考文献

[1] MAHMOOD Q，QADIR M A，AFZAL M T. Application of COReS to compute research papers similarity [J]. IEEE Access，2017，99：1.

[2] PARASCHIV I C，DASCALU M，TRAUSAN-MATU S，et al. Analyzing the semantic relatedness of paper abstracts：an application to the educational research field [C] //International Conference on Control Systems and Computer Science. IEEE，2015：759-764.

[3] LESK M. Automatic sense disambiguation using machine readable dictionaries：how to tell a pine cone from an ice cream cone [C] //Acm Special Interest Group for Design of Communication，1986：24-26.

[4] SALTON G，SINGHAL A，MITRA M，et al. Automatic text structuring and summarization [J]. Information Processing & Management，1997，33（2）：193-207.

[5] PAPINENI S. BLEU：A method for automatic evaluation of machine translation [C] //Meeting of the Association for Computational Linguistics. Association for Computational Linguistics，2002.

[6] MADHAVAN J，BERNSTEIN P A，DOAN A H，et al. Corpus-based schema matching [C] //International Conference on Data Engineering. ICDE 2005. IEEE，2005：57-68.

[7] LABBÉ C，LABBÉ D. Duplicate and fake publications in the scientific literature：how many SCIgen papers in computer science? [J]. Scientometrics，2013，94（1）：379-396.

[8] 梁艳艳. 基于主题域的文本相似度系统分析与设计 [D]. 济南：山东科技大学，2015.

[9] LIN D. An information-theoretic definition of similarity [C] //International Conference on Machine Learning，1998：296-304.

[10] BEEL J，GIPP B，LANGER S，et al. Research-paper recommender systems：a literature survey [J]. International Journal on Digital Libraries，2015，17（4）：1-34.

[11] GOMAA W H，FAHMY A. A survey of text similarity approaches [J]. International Journal of Computer Applications，2013，68（13）：13-18.

[12] 陈二静，姜恩波. 文本相似度计算方法研究综述 [J]. 数据分析与知识发现，2017，1（6）：1-11.

[13] RISTAD E S. Learning string-edit distance [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence，1998，20（5）：522-532.

[14] BERGROTH L，HAKONEN H，RAITA T. A survey of longest common subsequence algorithms [C] //International Symposium on String Processing and Information Retrieval. Spire 2000. IEEE，2002：39-48.

[15] WU S，MANBER U. A fast algorithm for multi-pattern searching [D]. Arizona：Department of Computer Science，University of Arizona，1994.

[16] SIDOROV G，GELBUKH A，GÓMEZADORNO H，et al. Soft similarity and soft cosine measure：similarity of features in vector space model [J]. Computación Y Sistemas，2014，18（3）：491-504.

[17] JONES K S，WALKER S，ROBERTSON S E. A probabilistic model of information retrieval：development and comparative experiments [M]. Pergamon Press，Inc，2000.

[18] BOYACK K W，NEWMAN D，DUHON R J，et al. Clustering more than two million biomedical publications：Comparing the accuracies of nine text-based similarity approaches [J]. PLoS One，2011，6（3）：e18029.

[19] HOFMANN T. Probabilistic latent semantic analysis [C] //Fifteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc，1999：289-296.

[20] BLEI D M，MCAULIFFE J D. Supervised topic models [J]. Advances in Neural Information Processing Systems，2010，3：327-332.

[21] RAMAGE D，MANNING C D，DUMAIS S T. Partially labeled topic models for interpretable text mining [C] //Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. DBLP，2011.

[22] BURGESS C，LIVESAY K，LUND K. Explorations in context space：Words，sentences，discourse [J]. Discourse Processes，1998，25（2/3）：211-257.

[23] SIDOROV G，GOMEZ-ADORNO H，MARKOV I，et al. Computing text similarity using tree edit distance [C] //Fuzzy Information Processing Society. IEEE，2015：1-4.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[24] JOHN V，KONSTANTINOS T，IRAKLIS V，et al. Text classification using the N-Gram graph representation model over high frequency data streams [J]. Frontiers in Applied Mathematics and Statistics，2018.

[25] LEACOCK C，CHODOROW M，FELLBAUM C. Combining local context and WordNet similarity for word sense identification [J]. Wordnet An Electronic Lexical Database，1998.

[26] WU Z，PALMER M. Verbs semantics and lexical selection [C] //Meeting on Association for Computational Linguistics. Association for Computational Linguistics，1994.

[27] LESK M E. How to tell a pine cone from an ice cream cone [C] //Acm Sigdoc Conference，1986.

[28] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy [C] //International Joint Conference on Artificial Intelligence，1995.

[29] JIANG J J，CONRATH D W. Semantic similarity based on corpus statistics and lexical taxonomy [J]. Rocling，1997：11512.

[30] MIHALCEA R，CORLEY C，STRAPPARAVA C. Corpus-based and knowledge-based measures of text semantic similarity [C] //National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference，2006：775-780.

[31] TURNEY，PETER D. Mining the Web for synonyms：PMI-IR versus LSA on TOEFL [C] //Proc European Conference on Machine Learning，2001：491-502.

[32] ISLAM M A，INKPEN D. Second order co-occurrence PMI for determining the semantic similarity of words [C] //In Proceedings of the International Conference on Language Resources and Evaluation，2006：1033-1038.

[33] BENGIO Y，SCHWENK H，SENÉCAL J S，et al. A neural probabilistic language model [J]. Journal of Machine Learning Research，2003，3（6）：1137-1155.

[34] LIU M，LANG B，GU Z，et al. Measuring similarity of academic articles with semantic profile and joint word embedding [J]. Tsinghua Science and Technology，2017，22（6）：619-632.

[35] LE Q，MIKOLOV T. Distributed representations of sentences and documents [C] //International Conference on International Conference on Machine Learning. JMLR. org，2014：II-1188.

[36] LIU T C，WU Y H，LEE H Y. Query-based attention CNN for text similarity map [C] //IEEE International Conference on Computer Vision. IEEE，2017.

[37] GABRILOVICH E，MARKOVITCH S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis [C] //Proc. International Joint Conference on Artificial Intelligence，2007：1606-1611.

[38] POTTHAST M，STEIN B，ANDERKA M. A Wikipedia-based multilingual retrieval model [C] //Advances in Information Retrieval，European Conference on Ir Research，ECIR 2008，Glasgow，Uk，March 30-April 3，2008. 2008：522-530.

[39] JEH G，WIDOM J. SimRank：a measure of structural-context similarity [C] //Eighth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM，2002.

[40] HU X，ZHANG X，LU C，PARK E K，et al. Exploiting Wikipedia as external knowledge for document clustering [C] //Proc 15th ACM SIGKDD，Jun. 2009.

[41] ZHAO P，HAN J，SUN Y. P-Rank：a comprehensive structural similarity measure over information networks [C] //ACM Conference on Information & Knowledge Management. ACM，2009：553-562.

[42] YOON S H，KIM S W，PARK S. C-Rank：A link-based similarity measure for scientific literature databases [J]. Information Sciences，2016，326：25-40.

[43] SMALL. Citation context analysis [J]. Progress in communication science，1982，74（3）：287-310.

[44] 董振东，董强. 知网 [EB/OL]. [2012-03-20]. www. keenage. com.

[45] 梅家驹，竺一鸣，高蕴琦，等. 同义词词林 [M]. 上海：上海辞书出版社，1983.

[46] SHIRAKAWA M，NAKAYAMA K，HARA T，et al. Wikipedia-based semantic similarity measurements for noisy short texts using extended naive Bayes [J]. IEEE Transactions on Emerging Topics in Computing，2015，3（2）：205-219.

[47] PILEHVAR M T，JURGENS D，NAVIGLI R. Align，disambiguate and walk：A unified approach for measuring semantic similarity [C] //Meeting of the Association for Computational Linguistics，2013.

[48] 晋耀红. 基于语境框架的文本相似度计算 [J]. 计算机工程与应用，2004，40（16）：36-39.

[49] CROCKETT K，ADEL N，O'SHEA J，et al. Application of fuzzy semantic similarity measures to event detection within tweets [C] //IEEE International Conference on Fuzzy Systems. IEEE，2017：1-7.

作者简介： 黄文彬（ORCID：0000-0002-9174-5467），男，1977 年生，博士，副教授。研究方向：数据分析。车尚锟（ORCID：0000-0001-5012-6515，通讯作者），男，1997 年生。研究方向：文本挖掘，信息抽取。

作者贡献声明： 黄文彬，提出研究思路，设计研究方案，论文修订。车尚锟，数据采集与分析，设计分类体系，论文撰写。

录用日期： 2019－05－28

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net