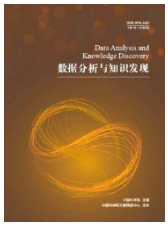


数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：　无监督引用文本自动识别与分析
作者：　金贤日，欧石燕
网络首发日期：　2020-09-03
引用格式：　金贤日，欧石燕．无监督引用文本自动识别与分析．数据分析与知识发现.
https://kns.cnki.net/kcms/detail/10.1478.G2.20200903.1009.002.html

网络首发时间：2020-09-03 14:09:12
网络首发地址：https://kns.cnki.net/kcms/detail/10.1478.G2.20200903.1009.002.html

无监督引用文本自动识别与分析

金贤日，欧石燕
(南京大学信息管理学院 南京 210023)

摘要：
[目的] 探索引用文本自动识别方法，并比较不同类型引用句在内容上的差别。
[方法] 提出了一种无监督引用文本识别方法，通过比较候选句与施引文献和被引文献的文本相似度来确定隐性引用句。为了精确计算文本相似度，提出了向量空间模型与词嵌入模型相结合的两种文档向量模型。
[结果] 分别对两篇高被引论文约 200 篇施引文献中的隐性引用句进行了识别，本文所提方法的 F 值均达到了 92%以上。通过对显性引用句和隐性引用句的内容进行比较，发现两者在引用功能和情感上有明显区别：表达研究背景和技术基础的隐性引用句比例要高于显性引用句，而表达研究基础和研究比较的隐性引用句比例要低于显性引用句；45.3%的显性引用句为正面引用，而 78.8%的隐性引用句为中性引用。
[局限] 本文目前只是对句子层面的引用文本识别进行了识别，在短语层面的引用文本识别还有待于进一步探索。
[结论] 在识别引用文本时有必要识别隐性引用句，本文提出的引用文本识别方法性能较高。
关键词： 引用文本识别；隐性引用句；引用内容分析
分类号： TP393，G250
DOI: 10.11925/infotech.2096-3467.2020.0548

The Unsupervised Identification and Analysis of Citation Texts

Kim Hyonil, Ou Shiyan
(School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract:
[Objective]This paper intends to explore the method for automatic identification of citation texts and compare the difference in the content of different types of citation sentences.
[Methods] This paper proposed an unsupervised method for identifiying citation texts, which determines implicit citation sentences by comparing the similarity of a candidate sentence wth a citing paper and that with a cited paper. To precisely calcuate text similarity, two document vector models were propsoed by combining the vector space model and the word embedding model.
[Results] while identifying the implicit citation sentences of two higly-cited papers respectively from over 200 citing papers, the proposed unsupervised method obtained the F-value of above 92%. By comparing the content of the explicit and implicit citaiton senstences, it was found that there are significant difference in citation function and citaiton sentiment between the two types of citation sentences: the proportion of implicit citation sentences expressing research background and technical basis is higher than that of explicit citation sentences, while the proportion of

implicit citation sentences expressing research basis and research comparison is lower than that of explicit citation sentence; 45.3% of explicit citation sentences were positive references while 78.8% of implicit citation sentences were neutral references.

[Limitations] This paper only identifies citation texts at sentence level. The clause-level and phrase-level identification should be explored further.

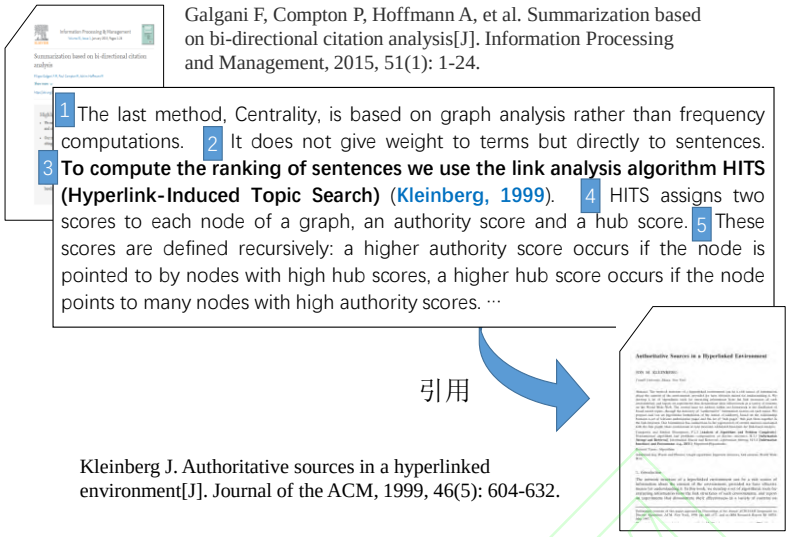
[Conclusions] It is necessary to contain implicit citation sentences while identifying citaion texts. The proposed similarity-based method is effective.

Keywords： Citation Text Identification; Implicit Citation Sentence; Citation Context Analysis

1 引言

长期以来，引文分析（Citation Analysis）作为文献计量学的主要分析方法，在评价科研成果、揭示学科结构、预测科学发展趋势等方面被广泛应用。基于被引频次的传统引文分析将所有的引用关系同等对待，对目的和态度不同的引用不加区分，这种粗粒度、浅层次的引文分析方法只能反映出表层的引用关系，无法深入揭示引用现象的本质，使得引文分析结果常引起许多争议[1-3]。在此背景下，图书情报领域的学者们提出了引用内容分析（Citation Context Analysis、Citation Content Analysis 或 Content-based Citation Analysis，CCA）的概念，试图通过对引用内容进行细粒度、深层次的分析，从内容和语义层面揭示引用现象的本质[4,5]。

进行引用内容分析的前提是从施引文献中抽取出反映引用内容的上下文文本（简称引用文本）。学术文献的作者在引用参考文献时，通常会以标准的引文著录和标记方式表示对参考文献的引用。引用多以完整的句子为单位，引用标记被置于被引内容的末尾，这种带有明显引用标记的句子我们称之为显性引用句。但是施引文献对参考文献的引用信息有时并不仅限于显性引用句，而是会涉及其周围若干个句子，但往往却并不再附加引用标记，这种没有带有引用标记但也包含被引参考文献相关内容的句子我们称之为隐性引用句。显性引用句与其周围的隐性引用句共同组成了引用文本，如图 1 所示。显性引用句和隐性引用句识别的方法与难度有很大不同：仅依靠引用标记就可以很容易地识别出显性引用句[6]，但隐性引用句的识别则没有明显的语言和句法线索，需要根据其深度语义特征来判定。因此引用文本的识别在很大程度上就是确定显性引用句周围隐性引用句的范围。



注：第3到5句为引用文本，其中，第3句为显性引用句，第4和第5句为隐性引用句

图 1 包含显性引用句和隐性引用句的引用文本实例

Fig 1. A Sample Citation Text that Contains the Explicit and Implicit Citation Sentences

目前有关引用文本识别的研究中，许多研究者仅将带有引用标记的句子（即显性引用句）作为引用文本，完全忽略隐性引用句的存在，这不可以避免地会遗漏大量引用信息。有些研究者将引用标记左右一段固定长度的文本窗口（通常包含若干个句子）作为引用文本，但这会带来大量噪音。也有少量研究者视引用文本为可变长度的窗口，通过机器学习技术来判断引用标记周围的若干个句子是否属于引用句。但这类方法多为有监督学习，需要大量的人工标注语料，难以迁移到其他领域。

在本文中，我们首先提出了一种基于文本相似度的无监督隐性引用句识别方法，无需标注大量的训练语料，适用于不同领域的学术文献。在采用该方法识别出隐性引用句的基础上， 我们比较分析了显性引用句和隐性引用句在引用功能和引用情感上的差异，进一步说明了进行引用内容分析时识别隐性引用句的必要性。

2 相关研究

目前，引用文本识别研究主要是在宏观层面识别引用标记周围的一段文本（包含一个或若干个句子）是否包含引用信息，主要采用固定窗口法和可变窗口法两类方法。所谓固定窗口法，就是将引用标记周围一段固定长度的文本窗口作为引用文本。其关键点是如何确定最佳的窗口长度，该长度常以词语数量或句子数量来衡量。一些研究者将引用文本用于信息检索系统的文献标引当中，通过考察引用文本窗口长度对信息检索效果的影响，从而确定合适的窗口长度[7,8]。基于固定窗口的引用文本识别简单而实用，但会造成大量噪音。

基于可变窗口法的引用文本识别可分为基于规则与基于统计两类方法。基于规则的方法是基于线索词手工构建用于引用文本识别的规则[9,10]。然而，由于不同学科的学术文献撰写规范和习惯不同，此种方法迁移性较差，使用并不广泛。

基于统计的方法是引用文本识别的主流，根据采用的机器学习技术不同，又可进一步分为有监督学习和无监督学习两种。有监督的引用文本识别是在已标注的引用文本语料库中采用条件随机场（CRF）、隐马尔科夫模型（HMM）、支

持向量机（SVM）等机器学习算法训练出分类模型，用以判断一个句子是否属于引用文本，不同研究者探索了不同的分类特征。2009 年，Kaplan 等人采用支持向量机（SVM）从 MUC-7 语料库中训练出共指解析模型， 然后利用该模型识别句子中的共指链(Coreference Chain)，将与显性引用句位于同一共指链上的句子识别为隐性引用句[11]，但识别效果并不理想。有些研究者则利用词汇特征及引用标记位置特征训练分类模型，譬如，Angrosh 等人将训练出的 CRF 分类器用来识别论文文献综述部分的引用文本，获得了良好的效果，正确率高达 96.51%[12]；但 Athar 等人将训练出的 SVM 分类器用于识别全文中的引用文本，识别效果并不理想[13]。Sondhi 和 Zhai 认为隐性引用句与显性引用句具有较高的相关性，因此基于语言生成模型计算两者间相互生成的概率，以此构建 HMM 模型用于引用文本的识别，取得了较高的查准率（98.7%），但是查全率则比较低（50.3%）[14]。

国内针对引用文本识别的研究还刚刚起步，只有极少量的研究。2016 年，雷声伟等人综合利用指代、句子位置、引用标记位置、句子结构等特征从标注语料中训练出 CRF 和 SVM 分类器用于识别隐性引用句，并通过实验发现 SVM 分类器的识别效果要优于 CRF 分类器[6]。总体来说，有监督机器学习方法在引用文本识别中取得了一定效果，但缺点是需要标注大量训练数据，而引用文本的标注是一项难度很大的工作， 因此阻碍了这类方法的推广应用。

目前采用无监督学习方法进行引用文本识别的相关研究还较少，比较有代表性的研究有两项。2010 年 Qazvinian 等人采用马尔科夫随机场（Markov Random Field, MRF）的引用文本识别研究[15]。该研究没有使用训练数据训练 MRF 模型的参数，而是利用候选句与显性引用句之间的相似度和距离以及候选句的词汇特征来设置模型参数，实现了无监督引用文本识别，但识别效果并不理想。2018 年，Jebari 等人提出了基于文本相似性的无监督识别方法，把显性引用句后面的若干个句子作为候选句，采用 LDA 模型计算候选句和被引文献摘要间的相似度，并把最相似的句子作为隐性引用句[16]。一方面该方法没有进行评测，不知其效果如何，另一方面在于给每条引用都硬性分配了一个隐性引用句，显然与实际不符。与有监督学习方法相比，无监督学习方法不需要训练语料，在引用文本识别中具有较大的应用潜力，因此是值得探索的方向。

3 基于文本相似度的引用文本自动识别方法

目前，在引用文本识别的相关研究中，基本都只考虑施引文献中的信息，譬如隐性引用句和显性引用句之间的相关性，但是忽略了隐性引用句与被引参考文献之间的相关性。隐性引用句与显性引用句一样，在一定程度上反映了被引参考文献中某些方面的内容。根据我们对 7 篇施引文献中引用文本的手工识别与分析，发现显性引用句多是对被引内容进行简单概括，而隐性引用句则是对引用内容做进一步详述或评价。在这种情况下，隐性引用句和显性引用句之间的语义相似性很差，此时需要参考被引文献中的信息来判断该句是否提及了被引文献中的内容。

因此，本文提出如下假设：与所在的施引文献相比，隐性引用句与被引参考文献更加相似。基于该假设，本文提出了一种基于文本相似度的无监督隐性引用句识别方法，其流程如图 2 所示。首先根据引用标记识别出显性引用句，然后选择显性引用句周围若干个句子作为候选引用句，接下来比较每个候选引用句与施引文献和被引参考文献之间的内容相似度，最后将与参考文献更加相似的句子判

定为隐性引用句。候选句与文献之间内容的相似程度通过文本相似度来衡量，文献内容可采用全文文本或者摘要文本来表示。为了能够更好地反映文本间的相似情况，需要对不同的文本向量表示方法进行探索与比较。

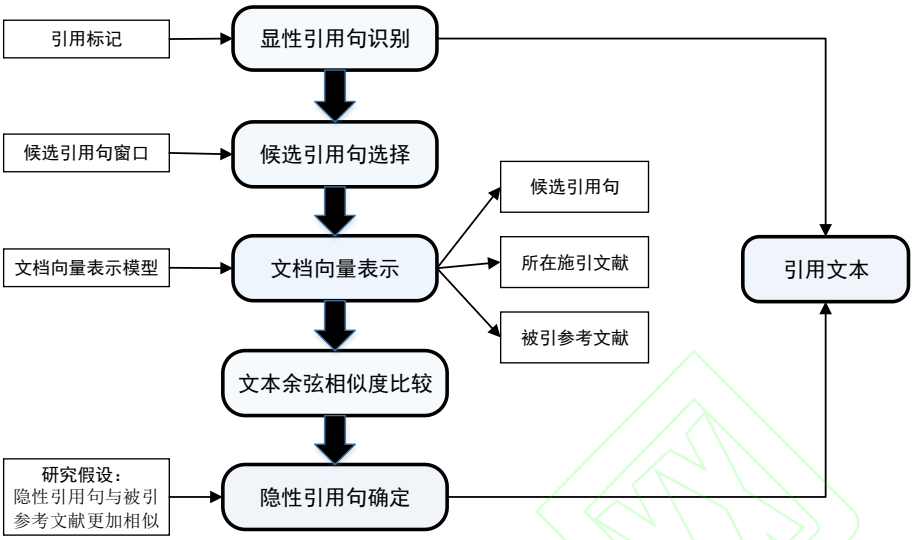


图 2 隐性引用句识别框架

Fig 2. A Framework for Implicit Citation Sentence Identification

3.1 基于不同文本向量表示的文本相似度计算方法

在当前各种文本相似度计算方法当中，最普遍采用的是基于 TF-IDF 权重的向量空间模型（Vector Space Model）表示文本，然后计算两个文本向量间的余弦距离。但是，这种基于词袋的文本向量表示方法只考虑文本间词汇的共现情况，并不考虑词汇的语义，这一缺点在隐性引用句识别中尤为严重。因为作者在施引过程中，往往通过概括或者转述的方式来提及被引参考文献原文中的内容，而非直接重复原文中的表述，这导致引用内容和原文内容虽然在语义上仍然相似，但是使用的相同词汇却大大减少。

随着深度学习技术的发展与进步，Mikolov 于 2013 年提出了词嵌入（Word2Vec）模型 CBOW 和 Skip-gram [17]，可以从无标注的大规模语料库中为其中的每个单词训练出一个表达语义的词向量，从而为计算词汇之间的语义相似度带来了重大突破。在词向量基础上， Le 和 Mikolv 又于 2014 年提出了两个文档嵌入（Doc2Vec）模型 PV-DM 和 PV-DBOW [18]。在这两个模型中，文档（或者句子、段落）被当作一个“特殊词汇”加入到文档所有局部上下文中，然后利用词嵌入模型来推测出文档的向量表示。但是， Doc2Vec 模型的缺点在于无法精确反映每个词汇在文档中的权重。

综上所述，向量空间模型和基于深度神经网络的文档向量表示模型在表示文本向量时各有优缺点。向量空间模型能够精确计算词汇在文本中的权重，但是不考虑词汇间的语义关系；深度神经网络模型虽然能够捕捉到词汇的语义，但是不考虑词汇在文本中的权重。因此，在本文中，为了能够精确计算候选引用句与文献之间的语义相似度，我们探索向量空间模型与深度神经网络模型相结合的文本向量表示方法，提出了两种结合方式：一种是在基于深度神经网络的文档向量表示模型中，嵌入 TF-IDF 权重；另一种是在基于 TF-IDF 的向量空间模型中，采用基于深度神经网络的词向量代替原模型中词的独热表示。

（1）基于 TF-IDF 权重和词向量的文档向量表示模型

在传统向量空间模型中，文档被看作是由一组词汇构成的词袋。针对词袋中的每个词汇，采用基于深度神经网络的词嵌入方法可从大规模未标注语料中训练出其词向量表示。我们考虑采用词汇的词向量线性加权组合来预测词袋（即文档）的向量。但是何种加权组合能够更精确地反映词袋的实际向量则需要做进一步验证。鉴于在大规模文本语料库中进行此验证实验计算量非常大，且文档的实际向量表示也难以获得，因此，在本文中，我们采用多义词语料库来探索词向量的不同线性组合与词袋向量之间的关系。

仿照 Le 和 Mikolov 的做法[18]，我们也将文档看作一个“特殊词汇”。因为文档能够表达多方面的语义，所以这个“特殊词汇”也是一个多义词。多义词的每个语义可被看作是一个特殊的“语义词汇”，那么一个多义词则是由它所包含的所有“语义词汇”构成的词袋。为了探索多义词的词向量表示，我们采用多义词语料库 SENSEVAL[1] 进行多义词词向量表示实验。 SENSEVAL 是由计算语言学协会（Association of Computational Linguistics, ACL）构建的一个多义词语义消歧语料库。在该语料库中，针对每个多义词的不同语义，均提供一个示例句展示该语义的用法。以多义词“line”为例，其有 6 个语义：*cord*（绳子）、*division*（分隔）、*formation*（编队）、*phone*（电话）、*product*（产品）和 *text*（文本）。将上述每个语义看作是一个特殊“语义词汇”，分别命名为 *line_cord*、*line_division*、*line_formation*、*line_phone*、*line_product* 和 *line_text*, 然后将示例句中的多义词替换为相应的“语义词汇”，则构成包含原始示例句和替换后示例句的训练语料, 如表 1 所示。基于该语料，我们利用 Google 开发的词向量训练算法 Word2Vec[2]训练出语料中每个多义词及其每个“语义词汇”的词向量表示。

表 1 多义词词向量训练语料（以多义词“line”为例）

Table 1 An example from the training corpus for multi-sense word vectors

语义	示例句
cord	原始：He hung on to his ***line*** and landed the fish.
	替换：He hung on to his ***line_cord*** and landed the fish.
division	原始：Further, blur the legal line separating commercial and investment banking.
	替换：Further, blur the legal line_division separating commercial and investment banking.
formation	原始：Correspondent said in the passport ***line*** at Moscow’s Sheremetyevo airport.
	替换：Correspondent said in the passport ***line_formation*** at Moscow’s Sheremetyevo airport.
phone	原始：He made another call and came back on the ***line*** with the news that …
	替换：He made another call and came back on the ***line_phone*** with the news that …
product	原始：In addition, Mr. Frashier will push for development of a ***line*** of protein-based adhesive and coating products.
	替换：In addition, Mr. Frashier will push for development of a ***line_product*** of protein-based adhesive and coating products.
text	原始：Clients reportedly get a one-page bill on which is written a single ***line***.
	替换：Clients reportedly get a one-page bill on which is written a single ***line_text***.

[1] 注：此语料库来自 https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/ senseval.zip。

[2] 注：Word2Vec 算法的地址为 https://github.com/tmikolov/word2vec。

鉴于一个多义词可被看作是它所包含的所有“语义词汇”的词袋，因此可采用这些“语义词汇”词向量的线性组合来计算（预测）多义词的词向量。我们定义了两种线性组合模型：平均模型 AWV 和加权平均模型 TF-AWV，前者是“语义词汇”词向量的平均，后者则是“语义词汇”词向量的加权平均， 这两个模型的数学表示如下所示。

假设词袋 **U** 表示为 $\mathbf{U}=\{u_1,u_2,\cdots,u_i,\cdots,u_m\}$，其中 u_i 为词袋中的第 i 个词语。基于平均模型的词袋 **U** 的词向量表示为公式（1），基于加权平均模型的词袋 **U** 的词向量表示为公式（2）。

$$AWV(\mathbf{U})=\frac{1}{m}\sum_{i=1}^{m}V_{u_i} \tag{1}$$

$$TF-AWV(\mathbf{U})=\frac{1}{\sum_{i=1}^{m}g_i}\sum_{i=1}^{m}g_i \bullet V_{u_i} \tag{2}$$

在公式（1）和（2）中，V_{u_i} 表示词汇 u_i 的词向量，g_i 则表示词袋 **U** 中第 i 个词汇 u_i 的权重，通常为 u_i 的出现频次或者出现概率（Term Frequency, TF）。

我们分别采用上述两种线性组合模型基于“语义词汇”的词向量计算出多义词的预测词向量，通过将其与基于 SENSEVAL 语料库训练出的多义词真实词向量进行对比，可以获知哪个线性模型能够更好地表示多义词的词向量。表 2 所示是 4 个多义词的真实词向量与其“语义词汇”的词向量、基于两种线性组合模型的预测词向量之间的余弦相似度。

表 2 多义词真实词向量与基于线性组合模型的预测词向量之间的余璇相似度

Table 2 The cosine similarity between the real word vector of a multi-sense word and its predicted word vectors based on the two linear combination models

多义词	语义词汇	余弦相似度	多义词	语义词汇	余弦相似度
line	line_cord	0.45	interest	interest1	0.59
	line_division	0.54		interest2	0.62
	line_formation	0.48		interest3	0.53
	line_phone	0.57		interest4	0.49
	line_product	0.92		interest5	0.57
	line_text	0.46		interest6	0.86
	AWV	0.74		AWV	0.78
	TF-AWV	**0.96**		**TF-AWV**	**0.92**
server	server2	0.79	hard	hard1	0.98
	server6	0.62		hard2	0.81
	server10	0.79		hard3	0.61
	server12	0.78		AWV	0.94
	AWV	0.91		**TF-AWV**	**0.99**
	TF-AWV	**0.93**			

从表 2 可以看出，针对上述四个多义词（即词袋），基于加权平均模型计算得出的词向量与其真实词向量最为接近，余弦相似度均达到 0.9 以上。因此说明，可以使用 TF 加权平均模型表示词袋的词向量。

文档虽然也被看作是词袋，但文档这种词袋与上述多义词这种词袋有一个重要不同：文档中的词汇，除了词频（TF）这一权重，还有一个更重要的权重—词频-逆文档频率（TF-IDF）权重，能够更加精确地反映词汇在文档中的重要程度。因此，我们在前述 TF 加权平均模型的基础上，将其中每个组成词汇的 TF 权重替换为 TF-IDF 权重，得到预测文档向量的 TF-IDF 加权平均模型 TFIDF-AWV， 如公式（3）所示：

$$TFIDF-AWV(\mathbf{d})=\frac{1}{\sum_{w\in\mathbf{d}}tfidf(w,\mathbf{d})}\sum_{w\in\mathbf{d}}tfidf(w,d)\bullet V_w \tag{3}$$

（2）基于 TF-IDF 权重和词向量的向量空间模型

如前所述，在基于传统空间向量模型计算文本间的相似度时，只考虑词汇在文本中的共现情况。如果一个词汇不出现，则其权重为 0，完全忽略其在文本中可能存在语义相同或相近的替代词汇。对于传统向量空间模型的这一局限性，本文采用语义相似的词语来进行相互替代，提出了基于 TF-IDF 权重和词向量的向量空间模型 PTFIDF-VSM，其数学表示如下所示。

假设 $V_{\mathbf{d}}=(v_1,v_2,\cdots,v_i,\cdots,v_n)$ 是文档 d 的基于向量空间模型的向量表示，其中，v_i 表示文档中的第 i 个词汇的权重。如果该词汇在文档中出现，以其 TF-IDF 权重来赋值，如果不出现，则采用文档中出现的与该词汇语义最相似的词汇（常为同义词或近义词）的 TF-IDF 权重来代替，但该值需采用两者间的语义相似度来进行修正。其具体计算如公式（4）和（5）所示。

$$v_i=\begin{cases}TFIDF(w_i,\mathbf{d}),w_i\in\mathbf{d}\\ p_i\bullet TFIDF(w_i,\mathbf{d}),w_i\in\mathbf{d}\end{cases} \tag{4}$$

$$p_i=\max_{w_j\in\mathbf{d}}sim(V_{w_i},V_{w_j}) \tag{5}$$

在公式（5）中，p_i 表示文档 **d** 中与第 i 个词汇最相似的词汇与第 i 个词汇的语义相似度， 可采用基于 Word2Vec 训练出的词向量的余璇相似度来计算。

综上所述，本文探索了两种文档向量表示方法，即基于 TF-IDF 权重和词向量的文档向量表示模型（TFIDF-AWV）和基于 TF-IDF 权重和词向量的向量空间模型（PTFIDF-VSM）。 在下文中，我们将分别采用这两种模型对候选引用句和文献进行向量表示，用于文本相似度计算。

3.2 隐性引用句自动识别方法测评

根据 3.1 节提出的文档向量表示模型，在训练出词向量的基础上，可以将候选引用句、施引文献和被引参考文献表示为文档向量，然后利用余弦相似度，比较引用句与施引文献和被引参考文献间的语义相似程度。

（1）数据准备

首先，为了训练词向量，我们从计算语言学论文语料库(ACL Anthology Network Corpus)[3]中随机采集了 23,500 余篇科学论文并下载其全文，使用 Apache PDFBox 工具将 PDF 格式的全文转换为计算机可处理的文本，然后利用 Word2Vec 工具 训练出所有词汇的词向量。

接下来，为了比较不同文档向量模型的隐性引用句识别效果，我们需要采集样本论文并对其所含的所有引用文本（主要是隐性引用句）进行手工标注以构建实验语料。显性引用句根据引用标签可容易识得，但隐性引用句的识别并不容易，需要阅读并理解被引文献以及施引文献引用上下文的内容。因此从一篇论文中识别出所有被引参考文献的隐性引用句（如果存在）是一件非常耗时的任务。有鉴于此，我们只从计算机领域的 4 本英文期刊[4]中随机选取 2014 至 2017 年间发表的 7 篇论文，对每篇论文中的每条引用文本进行了手工识别，生成一个小型语料库。7 篇论文中共包含 207 条引用文本，其中 139 条引用文本（占比 67.1%）中只包含显性引用句，其他 68 条引用文本（占比 32.9%）则由显性和隐性引用句共同组成, 共涉及 98 个隐性引用句。

鉴于前面构建的实验语料规模较小，不足以评测隐性引用句的最终识别效果，我们选取了两篇高被引论文—Krizhevsky 等人于 2012 年发表的有关深度神经网络的会议论文和 Blei 等人于 2003 年发表的有关 LDA 主题模型的期刊论文[5]，随机爬取其各自约 200 篇施引文献，手工标注每篇施引文献中针对高被引论文的引用文本（含显性和隐性引用句），构建最终评测语料。考虑到不同领域文献的引用风格或有所差异，因此爬取的施引文献尽量来自多个领域。

（2）研究假设的验证

首先，我们采用上述实验语料对 3.1 节提出的假设“与所在的施引文献相比，隐性引用句与被引参考文献更加相似”进行验证。在实际中，由于文献的全文往往难以获得，因此在该实验中我们采用摘要和全文两种方式来表示文献。首先，采用不同的文档向量表示模型表示隐性引用句、施引文献（全文或摘要）和参考文献（全文或摘要），然后比较每个隐性引用句与其所在的施引文献和所提及的被引参考文献之间的余璇相似度。表 3 列出了采用不同文档向量表示模型的比较结果。

从表 3 可以看出，无论采用哪种文档向量表示模型，无论文献采用摘要还是全文表示，超过一半的隐性引用句（至少 57.11%）都与其被引参考文献更加相似。采用基于 TF-IDF 权重和词向量的文档向量表示模型和参考文献摘要，效果更加显著，有超过 80%的隐性引用句与被引参考文献更加相似。此外，采用文献的摘要代替其全文，隐性引用与其相似程度更加明显，这是因为引用句往往是对被引参考文献内容的概括，而摘要同样是概括性的，因此引用句与摘要的语义相

[3] ACL 语料库在 https://www.aclweb.org/anthology/中可以下载。

[4] 四本英文期刊分别为：Computer Speech and Language、Information Processing & Management、 Artificial Intelligence 和 Fuzzy and System。

[5] 两篇高被引论文分别为：①Krizhevsky A, et al. ImageNet Classification with Deep Convolutional Neural Networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. ② Blei D M, et al. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003: 993-1022.

似度要比全文更高。实验结果表明，本文提出的基于文本相似度的隐性引用句识别方法具有很大程度的合理性和可行性。

表 3 基于各种文档向量表示模型的隐性引用句与施引/参考文献的相似度比较结果

Table 3 Similarity Comparison Results between Implicit Citation Sentences and Citing/Cited Papers based on Various Document Vector Models

文档向量表示模型	简称	隐性引用句与被引参考文献更加相似的比例	
		文献表示为摘要	文献表示为全文
传统向量空间模型	TFIDF-VSM	69.33%	57.11%
基于 TF 或 TF-IDF 权重和词向量的文档向量表示模型	TF-AWV	79.37%	59.20%
	TFIDF-AWV	**80.32%**	70.82%
基于 TF-IDF 权重和词向量的向量空间模型	PTFIDF-VSM	73.62%	70.65%

（3）候选引用句范围确定

本文所提出的基于文本相似度的隐性引用句识别方法， 其关键之处有两点：一方面是候选引用句范围的设定: 如果范围设定过大，会带来大量噪音，导致低的查准率，如果范围设定过小，则会遗漏真正的隐性引用句，造成查全率偏低；另一方面是文档的向量表示方法，其直接影响了文本相似度计算的准确性，在前文提出了两种文档向量表示模型，需要确定哪种模型有更好的识别效果。

在本节中，我们通过实验语料来确定候选引用句的范围。候选引用句范围有左右两个窗口，左窗口是指位于显性引用句之前的句子，右窗口是指位于显性引用句之后的句子， 左右两个窗口长度的设定相对独立。我们分别考察左右两个窗口长度变化对识别效果的影响，评测指标为 F1 值。首先，固定右窗口的长度为 10（即显性引用句之后的 10 个句子），调节左窗口的长度由 1 到 9（即显性引用句之前的 1 到 9 个句子），将左右窗口内的句子作为候选引用句，不同左窗口大小下的识别效果如图 3 所示，不同曲线表示采用不同文档向量表示模型的结果。从该图可以看出，针对所有的文档向量表示模型，当左窗口的长度为 2 时隐性引用句识别的 F1 值达到最高，因此将候选引用句范围的左窗口长度设定为 2。接下来，固定左窗口的长度为 2，调节右窗口的长度由 1 到 14，不同右窗口大小下的识别效果如图 4 所示。从该图可以看出，针对所有的文档向量表示模型，随着右窗口长度的扩大，F1 值逐渐升高，当右窗口长度为 10 时，F1 值的增长趋于平缓，不再发生明显变化，因此将候选引用句范围的右窗口长度设定为 10。至此，设定候选引用句的范围为显性引用句之前的 2 个句子和之后的 10 个句子。

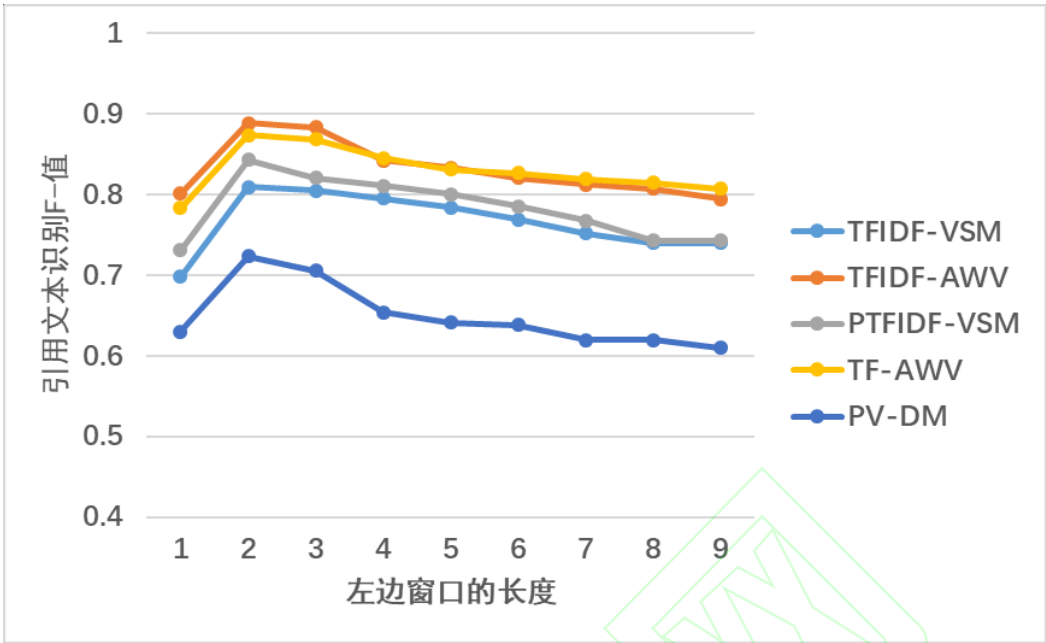


图 3 右窗口长度固定为 10 的条件下隐性引用句识别性能随左窗口长度变化的情况

Fig 3. The Performance Change of the Implicit Sentence Identification with Different Left Boundaries while the Right Boundary is Fixed to 10

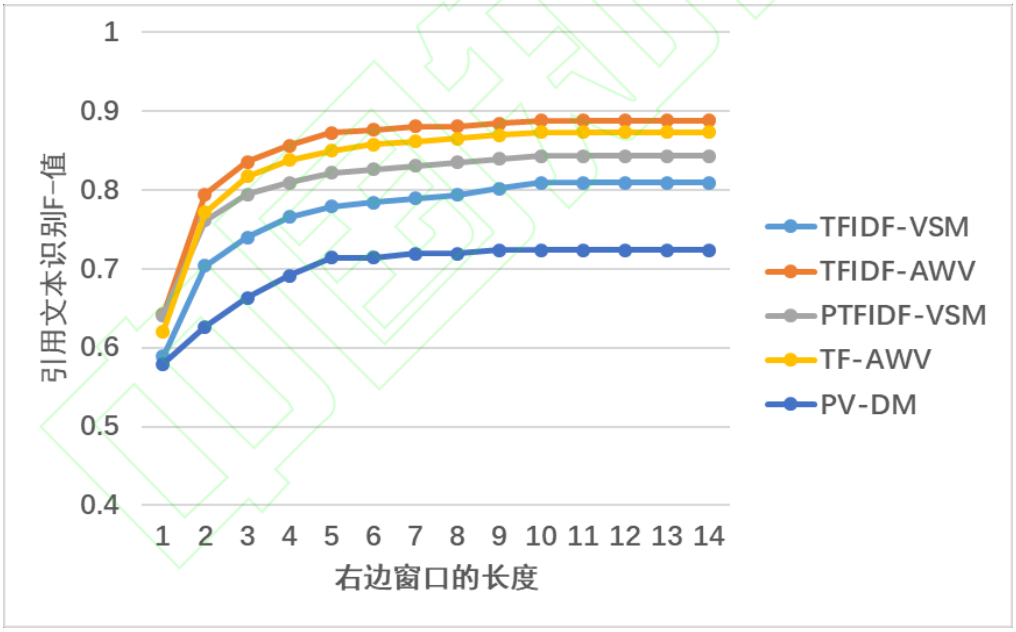


图 4 左窗口长度固定为 2 的条件下隐性引用句识别性能随右窗口长度变化的情况

Fig 4. The Performance Change of the Implicit Sentence Identification with Different Right Boundaries while the Left Boundary Is Fixed to 10

（4）基于不同文档向量表示模型的隐性引用句识别效果测评与结果分析

在确定了候选引用句范围之后，我们采用实验语料对前文提出的两个文档向量表示模型在隐性引用句识别中的性能进行评测。作为对比，将传统的向量空间模型以及 Mikolov 提出的 Doc2Vec 文档向量表示模型[18]作为基准模型。

表 4 所示为基于各种文档向量表示模型的隐性引用句识别性能，以查全率（R）、查准率（P）和 F1 值表示。从该表中可以看出，识别效果最好的模型是基于 TF-IDF 权重和词向量的文档向量表示模型，尤其是基于 TF-IDF 权重的模型，F1 值达到了 88.86%。 此外，当采用摘要表示文献时，识别效果要优于采用全文。在现实中，文献的摘要远比全文更容易获得，因此这一发现对于实际应用具有非常大的意义。针对所有文档向量表示模型，识别的查准率（P 值）都非常高，全部达到了 70%以上，有些模型甚至达到 99%以上，但是识别的查全率（R 值）并不理想，最高也只有 80%。这说明还是有部分隐性引用句被遗漏，因此要想提高识别的整体性能关键在于提高查全率。

表 4 基于各种文档表示模型的隐性引用句识别的性能

Table 4. The Identification Performance of the Implicit Citation Sentences Based on Different Document Vector Models

文档向量表示模型	简称	文献被表示为摘要			文献被表示为全文		
		R(%)	P(%)	F1(%)	R(%)	P(%)	F1(%)
传统向量空间模型	TFIDF-VSM	69.33	97.25	80.95	57.11	100.00	72.70
Doc2Vec 模型	PV-DBOW	63.06	84.96	72.39	54.28	97.66	69.77
基于 TF 或 TF-IDF 权重和词向量的文档向量模型	TF-AWV	79.78	96.52	**87.36**	59.20	99.25	74.16
	TFIDF-AWV	80.32	99.43	**88.86**	70.82	98.57	**82.42**
基于 TF-IDF 权重和词向量的向量空间模型	PTFIDF-VSM	73.62	98.76	84.36	70.65	100.00	**82.80**

为了提高查全率， 我们将基于不同文档向量表示模型的识别方法进行组合。首先，采用第一种文档向量表示模型从候选引用句中识别出引用句和非引用句；接下来，对第一步过滤出的非引用句采用第二种文档向量表示模型进一步进行识别，从中识别出前一步遗漏的隐性引用句。表 5 为不同组合模式的隐性引用句识别性能。从该表中可以看出，基于组合模式的隐性引用句识别方法能够进一步提高查全率，使得识别的整体性能大大提高。 最好的组合模式是基于 TF-IDF 权重和词向量的文档向量表示模型（TFIDF-AWV）和基于 TF-IDF 权重和词向量的向量空间模型（PTFIDF-VSM）的组合，F1 值达到 94%以上。这两个模型的组合顺序对识别性能有稍许影响，但两者区别不大，可忽略不计。

表 5 基于不同组合模式的隐性引用句识别性能

Table 5 The Identification Performance of the Implicit Citation Sentences Based on Various Hybrid Models

组合模式	文献被表示为摘要			文献被表示为全文		
	R(%)	P(%)	F_1(%)	R(%)	P(%)	F_1(%)
TFIDF-AWV + PTFIDF-VSM	90.51	99.45	**94.77**	87.63	99.05	**92.99**
PTFIDF-VSM + TFIDF-AWV	90.10	98.90	**94.30**	88.04	99.52	**93.43**
TFIDF-AWV + PV-DBOW	89.54	96.67	92.97	80.73	97.92	88.49
TFIDF-AWV + TFIDF-VSM	87.78	98.47	92.82	80.49	98.76	88.69

（5）基于最佳组合模型的隐性引用句识别效果最终评测

为了对隐性引用句的最终识别效果进行评测， 我们采用最佳识别模型（即组合模型 TFIDF-AWV + PTFIDF-VSM）在评测语料上对两篇高被引论文的隐性引用句进行自动识别，比较的施引文献和被引文献均采用文摘，结果分别如表 6 和表 7 所示。实验结果表明，两篇高被引论文隐性引用句识别的总体效果都比较理想，F 值都高达 92%，说明我们提出的隐性引用句识别方法是非常有效的。相比较而言，深度神经网络高被引论文隐性引用句的查准率（89.0%）要低于 LDA 主题模型高被引论文（96.6%），但召回率（96.4%）要高于后者（88.3%）。不同领域的施引文献中隐性引用句的识别效果没有明显区别。

表 6 深度神经网络高被引论文的隐性引用句识别结果

Table 6 The Identification Performance of the Implicit Citation Sentences of the Highly-Cited Paper on Deep Neural Network

施引文献领域	施引文献篇数	显性引用句数量	隐性引用句数量	隐性引用句识别结果		
				P(%)	R(%)	F(%)
计算机	89	118	214	89.5	97.6	93.4
工程学	65	86	136	91.0	95.9	93.4
物理	25	40	53	89.2	97.8	93.3
医学	24	31	54	89.8	92.9	91.3
其他	22	32	48	80.9	95.0	87.4
合计	225	307	505	-	-	-
平均	-	-	-	89.0	96.4	92.6

表 7 LDA 主题模型高被引论文的隐性引用句识别结果

Table 7 The Identification Performance of the Implicit Citation Sentences of the Highly-Cited Paper on LDA Topic Model

领域	施引文献篇数	显性引用句数量	隐性引用句数量	隐性引用句识别结果		
				P(%)	R(%)	F(%)
计算机	92	146	253	97.3	88.7	92.8
工程学	39	58	89	96.3	87.8	91.8
管理学	28	41	82	95.5	87.0	91.1
医学	13	24	45	95.3	88.4	91.7
商学	10	11	22	100.0	90.9	95.2
其他	25	47	90	96.3	88.1	92.0
合计	207	327	581	-	-	-
平均	-	-	-	96.6	88.3	92.3

4 显性引用句和隐性引用句对比分析

目前的引用内容分析研究主要是对引用文本的引用功能（或动机）、引用情感、引用主题进行分析。引用文本是由显性引用句及其周围的隐性引用句共同组成。但这两类引用句在表达作者引用内容时是否存在着差异尚未可知。我们仍以 Krizhevsky 等人的深度神经网络高被引论文为例，分析这篇高被引论文在被其他文献引用时显性引用句和隐性引用句在表达引用功能和情感上的差异。为了构建分析语料，我们首先从 Elsevier 网站中获取引用该文献的 1203 篇科学论文的全文（XML 格式）， 然后采用本文提出的引用文本识别方法（即采用引用标记识

别出显性引用句，然后采用 TFIDF-AWV 和 PTFIDF-VSM 模型的组合基于相似度识别出其周围的隐性引用句），从这些施引文献中共识别出 1622 条引用文本，其中包含 1622 个显性引用句， 3563 个隐性引用句，共计 5185 个引用句。

4.1 引用功能对比分析

我们根据 Dong 等人的分类标准，将引用功能分为四类：阐述研究背景（“背景”类别）、提供技术基础（“使用”类别）、激发现有研究的基本思想（“基于”类别）以及比较现有研究与其他研究（“比较”类别）[19]。利用本研究团队开发的引用功能自动分类工具[20]，对所有引用句（包含显性和隐性）的引用功能进行自动分类。不同引用功能在两种类型引用句中的分布如表 8 所示。

表 8 不同引用功能在显性引用句和隐性引用句中的分布对比

Table 8 The Citation Function Distribution in the Explicit Citation Sentences in comparison with that in the Implicit Citation Sentences

引用句类别	引用功能类别							
	背景		使用		基于		比较	
	数量	占比(%)	数量	占比%)	数量	占比(%)	数量	占比(%)
显性引用句	1223	75.4	306	18.9	60	3.6	33	2.3
隐性引用句	2762	77.5	755	21.2	12	0.3	34	0.9

针对表 8 中的统计数据，我们采用卡方检验方法检验显性引用句和隐性引用句在引用功能上的分布是否一致，也即检验引用句类别和引用功能类别是否存在着相关性。根据统计结果，卡方值为 104.4，大于临界值（自由度为 3，可信度为 99.95 时临界值为 8.05），这说明显性引用句和隐性引用句在引用功能分布上有明显差异。“背景”和“使用”这两类引用功能在隐性引用句中的占比要稍高于显性引用句，反之，“基于”和“比较”这两类引用功能在显性引用句中的占比要明显高于隐性引用句。

4.2 引用情感对比分析

我们根据目前普遍的引用情感分类标准，将引用情感分为 3 类：正面引用、负面引用和中性引用。同样利用本研究团队开发的引用情感自动分类工具[20]，对所有引用句（包含显性和隐性）的引用情感进行自动分类。不同引用情感在两种类型引用句中的分布如表 9 所示。

表 9 不同引用情感在显性引用句和隐性引用句中的分布对比

Table 9 The Citation Sentiment Distribution in the Explicit Citation Sentences in comparison with that in the Implicit Citation Sentences

引用句类别	引用情感类别					
	正面		负面		中性	
	数量	占比(%)	数量	占比(%)	数量	占比(%)
显性引用句	734	45.3	83	5.1	805	49.6
隐性引用句	546	15.3	208	5.8	2809	78.8

针对表 9 的数据，同样进行卡方检验，获得的卡方值为 541.9，大于临界值（自由度为 2，可信度为 99.95%时临界值为 5.98），这说明显性引用句和隐性引用句在引用情感分布上有明显差异。显性引用句中正面引用的占比（45.3%）非

常突出，远高于隐性引用句（15.3%）；而隐性引用句中接近于 80%的引用为中性引用。这说明，显性引用句对被引参考文献的主观评价较多，而隐性引用句则对参考文献的客观叙述较多。

5 结论与展望

本文聚焦于引用文本的识别，重点是识别不带有引用标记的隐性引用句。本文提出了一种基于文本相似度的无监督隐性引用句识别方法，通过比较文本相似度，将显性引用句周围与被引参考文献内容更加相似的句子识别为隐性引用句。为了精确地计算文本相似度，本文提出了不同的文档向量模型，并通过实验比较，得出了基于 TF-IDF 权重和词向量的文档向量表示模型和基于 TF-IDF 权重和词向量的向量空间模型的最佳组合模型。使用该模型对两篇高被引论文部分施引文献中的隐性引用句进行了自动识别，获得的 F 值高达 92%，说明了本文所提方法的有效性。相比已有的引用文本识别方法，此方法的优点在于无需要标注语料，只需从大规模无标注训练语料中训练出词向量，为文本相似度计算提供基础。

在显性引用句和隐性引用句识别的基础上，本文对显性引用句和隐性引用句表达的引用内容进行了比较分析。结果表明：在引用功能上，隐性引用句多表达“阐述研究背景”和“提供技术基础”，而显性引用句则主要表达“激发研究思想”以及“与其他研究进行比较”；在引用情感上，隐性引用句偏向于客观叙述参考文献的被引内容，多为中性引用，而显性引用句则较多包含对被引文献的正面评价，有明显的情感倾向。鉴于两者在表达内容上的区别，在引用文本识别时包含隐性引用句非常有必要。

本文目前只是对宏观层面的引用文本进行了识别，未来研究将关注微观层面的引用文本识别，即从带有引用标记的显性引用句中进一步精确识别出与指定参考文献内容相关的文本片断（如短语或从句）。

参考文献

[1] Chen C M. Eugene Garfield's Scholarly Impact: A Scientometric Review [J]. Scientometrics, 2018, 114(2): 489-516.

[2] 刘浏, 王东波. 引用内容分析研究综述[J]. 情报学报, 2017, 36(6): 637-643. (Liu Liu, Wang Dongbo. Review on Citation Context Analysis [J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(6): 637-643.)

[3] 陈颖芳, 马晓雷. 基于引用内容与功能分析的科学知识发展演进规律研究[J]. 情报杂志, 2020, 39(3): 71-80. (Chen Yingfang, Ma Xiaolei. Measuring the Developmental Trend of a Knowledge Domain through Citation Content and Citation Function Analysis [J]. Journal of Intelligence, 2020, 39(3): 71-80.)

[4] Tahamtan I, Bornmann L. What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published between 2006 and 2018[J]. Scientometrics, 2019, 121(3): 1635-1684.

[5] 吴素研, 吴江瑞, 李文波. 大规模科技文献深度解析和检索平台构建[J]. 现代情报, 2020, 40(1): 110-115. (Wu Suyan, Wu Jiangrui, Li Wenbo. Construction of Deep Resolution and Retrieval Platform for Large Scale Scientific and Technical Literature [J]. Journal of Modern Information, 2020, 40(1): 110-115.)

[6] 雷声伟, 陈海华, 黄永, 等. 学术文献引文上下文自动识别研究[J]. 图书情报工作, 2016, 60(17): 78-87. (Lei Shengwei, Chen Haihua, Huang Yong, et al. Research on Automatic Recognition of Academic Citation Context [J]. Library and Information Service, 2016, 60(17): 78-87.)

[7] Bradshaw S. Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes[C]// Proceedings of the International Conference on Theory and Practice of Digital Libraries. Lecture Notes in Computer Science, vol 2769. Heidelberg, Berlin: Springer, 2003: 499-510.
[8] Ritchie A, Robertson S, Teufel S, et al. Comparing citation contexts for information retrieval[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York, NY: Association for Computing Machinery, 2008: 213-222.
[9] O'connor J. Citing statements: Computer recognition and use to improve retrieval [J]. Information Processing and Management, 1982, 18(3): 125-131.
[10] Nanba H, Okumura M. Towards Multi-paper Summarization Using Reference Information[C]//Proceedings of the 16th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1999: 926-931.
[11] Kaplan D, Iida R, Tokunaga T. Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach[C]// Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL). 2009: 88-95.
[12] Angrosh M A, Cranefield S, Stanger N, et al. Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries[C]//Proceedings of the 10th Joint Conference on Digital Libraries (JCDL), New York, NY: Association for Computing Machinery, 2010: 293-302.
[13] Athar A. Sentiment Analysis of Citations using Sentence Structure-Based Features[C]//Proceedings of the ACL-HLT 2011 Student Session, Stroudsburg, PA: Association for Computational Linguistics, 2011: 81-87.
[14] Sondhi P, Zhai C X. A constrained hidden Markov model approach for non-explicit citation context extraction[C]// Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2014: 361-369.
[15] Qazvinian V, Radev D R. Identifying non-explicit citing sentences for citation-based summarization[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 555-564.
[16] Jebari C, Cobo M J, Herreraviedma E, et al. A New Approach for Implicit Citation Extraction[C]// Proceedings of the 19th International Conference on Intelligent Data Engineering and Automated Learning. Lecture Notes in Computer Science, vol 11315. Cham, Switzerland: Springer, 2018: 121-129.
[17] Mikolov T, Chen K, Corrado G S, et al. Efficient Estimation of Word Representations in Vector Space[J]，arXiv:1301.3781[cs.CL], 2013. [2020-08-10]. https://arxiv.org/pdf/1301.3781.pdf.
[18] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on Machine Learning, volume 32. 2014: 1188-1196.
[19] Dong C, Schafer U. Ensemble-style Self-training on Citation Classification[C]//Proceedings of the 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand. Asian Federation of Natural Language Processing, 2011: 623-631.
[20] 凌洪飞. 基于引文文本自动分类的引用内容分析研究[D]. 南京大学, 2020. (Ling Hongfei. A Study on Citation Context Analysis based on Automatic Citation Text Classification [D]. Nanjing University, 2020.)

(通讯作者: 欧石燕，ORCID：0000-0001-8617-6987，E-mail: oushiyan@nju.edu.cn。)

基金项目：本文系国家社科基金重点项目“基于关联数据的学术文献内容语义发布及其应用研究（项目编号 17ATQ001）”的研究成果之一。

作者贡献声明：

金贤日：采集数据，算法实现，进行实验，论文修改；

欧石燕：提出研究思路，设计研究方案，论文撰写与修改。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail: oushiyan@nju.edu.cn。

[1] 欧石燕. Experiment_corpus.xls. 包含 7 篇施引论文中所有引文文本的标注

[2] 欧石燕. Evaluation_corpus.xls. 包含 432 篇施引论文中针对 2 篇高被引论文的引文文本标注