基于引文上下文的学术文本自动摘要技术研究*

陈海华[1]，黄永[1]，张炯[1]，陆伟[1,2]
（1.武汉大学信息管理学院，武汉 430072；2.武汉大学信息检索与知识挖掘研究所，武汉 430072）

摘要：学术文本自动摘要是指对于给定学术文献，自动地抽取其核心内容，以提高用户撰写和阅读文献的效率。目前基于文本词频对句子重要性排序的自动摘要技术，无法从语义层面揭示学术文本的核心内容。本文在已有研究的基础上，引入引文上下文内容特征，并通过构建支持向量回归模型，综合考虑自动摘要系统中的各个特征对句子权重的影响，重新对句子重要性进行排序。基于WE-ROUGE的评测表明，相比于传统基于词频统计和图模型的方法，本文提出的算法能够有效提升自动摘要的准确度。

关键词：文本自动摘要；引文上下文；支持向量回归；词向量
中图分类号：G353.4
DOI：10.3772/j.issn.1673-2286.2016.8.007

1 引言

信息技术的发展促进了科学交流，导致科学产出极快增长。May的统计结果表明，学术公开出版物的平均年增长率为3.7%[1]，在一些热门领域，这个数字更为惊人[2]。如何在海量学术文献中快速而准确地获取信息成为研究热点，随之产生的自动摘要、文献推荐等信息过滤技术受到人们广泛关注。

学术文本自动摘要技术，一方面提高了用户检索和阅读文献的效率，另一方面提高了用户撰写学术文献的效率。其核心思想是对于给定文本或主题相近的文档集，计算机自动生成涵盖文本或文档集核心内容的摘要[3]。文本自动摘要技术主要分为基于全文内容抽取的自动摘要和基于全文内容理解的自动摘要，但由于自然语言理解与生成技术还有待发展，目前研究主要集中在基于全文内容抽取的自动摘要[4]。这一研究的主要思想是“原文句子的重要性可以通过关键概念的反复出现来评价，这些重要的句子能够用来生成自动摘要”[3]，但是反复出现的概念并不能准确完整地表达文章核心内容。

施引文献中的引文上下文包含丰富的信息，可以表现出被引文献具有代表性意义的方法、观点等内容特征[5]。这些内容可以直接被用于生成自动摘要[6]，然而只抽取引文上下文信息作为自动摘要存在两个方面的问题：一方面，由于作者在引用文献时目的不同，引文上下文的信息也有所侧重，所以直接抽取引文上下文信息生成的摘要并不一定能完全覆盖原始文献中的内容；另一方面，这种方法生成的摘要与基于全文内容抽取生成的摘要相比可读性较差，尤其是当引文中包含较多代词时。

本文基于施引文献的引文上下文信息对原始文本中的句子进行打分，并生成自动摘要。一方面，利用引文上下文中的有效信息，避免了原文中核心内容的遗漏；另一方面，使用原文句子生成摘要，保证摘要的可读性。本文设计相关实验，对比基于词频统计和基于图模型的方法，结果表明本文提出的算法能够有效提升自动摘要的准确度。

本文的主要贡献在于：（1）对国内外相关研究进行梳理，总结出用于重要句子识别的五类特征；（2）提出利用引文上下文信息对原文句子进行打分，避免重要信

* 本研究得到国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”（编号：71473183）资助。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

息遗漏的同时提高了摘要的可读性；(3) 利用SVR模型在自构建的数据集上进行实验，取得了很好的效果。

2 相关研究

目前，关于自动摘要技术的相关研究大多基于原始文本内容层面，将施引文献的引文上下文信息应用到摘要生成的研究则相对较少。本文将从这两方面进行相关文献的梳理。

2.1 基于文本内容的自动摘要技术

早期自动摘要技术都是基于包含高频词句子抽取的方法。然而，Baxendale发现句子在文本中出现的位置与其重要性间也存在某些关联[7]，例如，出现在文章段落开头或结尾的句子比其他句子更可能暗示出文章所要表达的信息；Edmundson则结合词频、位置信息（用一组线索词表示）、标题信息等抽取重要句子[8]。

20世纪90年代末期，随着自然语言处理和机器学习技术的发展，一些经典的机器学习算法，如朴素贝叶斯模型[9]、支持向量机[10]、隐马尔可夫模型[11]和PageRank[12]等被运用于文本自动摘要中。例如，Kupiec等通过训练朴素贝叶斯模型来计算文章中句子成为摘要句的概率[9]；Zhou等则认为句子是否成为摘要句不仅与其自身的重要性有关，还与其相邻句子是否被选为摘要句有关，基于该思想运用隐马尔可夫模型完成自动摘要并取得很好效果[11]；此外，Chuang等利用线索词将文章分割，通过有监督的机器学习算法抽取句子组成摘要[13]；Amini等引入机器学习排序的方法对文本中出现的句子进行排序，抽取排名靠前的句子作为摘要句[14]。

近年来，深度学习技术受到人们的广泛关注，一些学者开始利用词向量模型代替传统模型中基于空间向量模型，来计算句子之间的相关性。Rush等将图模型中的Attention机制引入句子重要内容的识别中，弥补传统模型应对长句时的缺陷[15]；Cao等指出利用深度学习技术来表示特征，会比简单地利用传统特征对文本中的句子进行打分排序更加科学，并提出将AttSum用于查询式摘要的生成[16]；AttSum联合相关性和显著性两个指标，测试结果表明在没有使用人工特征的情况下对句子进行打分，生成的摘要效果显著；Nallapati等将文本自动摘要看作将输入文本序列映射为另一个目标文本序列的问题，并利用神经网络改善摘要生成的质量，这也是未来基于理解的自动摘要发展方向[17]。

中文信息处理技术和汉语自身表达多样性等问题，给中文自动摘要生成增加了难度。刘挺等是较早进行中文自动摘要研究的学者，对原文进行词语频率、词语分布和修辞结构分析并结合用户需求，抽取原文部分内容填写文摘框架[18]；吴中勤等则利用最大熵模型和支持向量机模型对文本中的句子进行分类[19]；万小军提出一种基于簇排列的半监督学习算法，提高自动生成摘要的质量[20-22]。其他方面，沈焕生等考虑到词性和句子修辞等信息，利用文摘框架对生成的摘要进行加工组织，使生成的摘要更加连贯[23]。罗文娟等从文档中抽取熵和相关度这两组特征用以权衡摘要的信息覆盖度和紧凑性[24]。

2.2 引文上下文在自动摘要中的应用

基于内容的自动摘要技术在处理内容较为分散的文本时，由于摘要长度有限，直接通过抽取原始文本中的若干句子很难覆盖到文本中的核心内容。因此，Elkiss等认为可将施引文献的引文上下文作为自动摘要的补充[25]。

Qazvinian等通过构建文献间的引文网络，对描述被引文献的引文上下文进行聚类，在此基础上，分别利用Cluster Round-Robin和Cluster Lexrank方法抽取重要的引文上下文构建被引文献的摘要[26]。Tandon等也认为一个完整的引文上下文必然是对被引文献主要内容简洁而精确的总结[27]，因此可以直接作为抽取摘要的句子；基于这种思想，Tandon等首先对引文上下文按照Summary、Strengths、Limitations、Related work、Applications进行分类，再从各个类别中抽取句子生成一个全面介绍文章内容的结构化摘要[27]。不同的是，Cohan等则抽取完整的引文上下文（包括显式和隐式引文上下文）而非直接使用引文句作为引文上下文[28]，这样能涵盖更加完整的信息，然后将包含同一主题的引文上下文分组，从每组引文上下文中抽取核心内容组成摘要，一定程度上缓解信息遗漏和可读性差的问题，在TAC2014 dataset上的ROUGE值比之前最好结果提升了近30%。Galgani等同时利用施引文献和参考文献的上下文信息挖掘原始文献中的重要事实信息[29]，这些信息由多个概念组成，然后利用这些高频概念对原

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

文句子进行打分，在法律文本上取得很好的效果，并在学术文本上验证了该方法的通用性。胡珀将引文上下文内容引入句子权重的评价，较好地改善了自动摘要的质量[30]。

事实上，早在2009年，Mei等就提出利用引文上下文去自动生成基于学术影响力的摘要，其基于引文上下文信息，利用语言模型和KL距离计算原文中句子的重要程度，并抽取重要句子组成摘要，然而当摘要句为3句时，ROUGE-1值仅为0.323[31]。

本文通过对相关研究进行全面的调研和总结，认为基于引文上下文的文本自动摘要技术是当前较新的研究方向，融合施引文献对被引文献叙述和评论的上下文信息，可以更加准确完整地反映出被引文献的主要思想和核心内容，从而提升自动生成摘要的质量。

3 问题及方法

3.1 问题定义

基于全文内容抽取的自动摘要最重要的是从原文中抽取可作为摘要来源的句子，对于每个原文的句子，可构造一个从句子到句子重要性得分的映射函数。

若原文本记为D，该文本中的每个句子记为s_i，即$D=\{s_1, s_2, \cdots, s_i\}$，$i=1, 2, \cdots, n$，存在函数$f$对每个句子进行评分，使得得分前$m$个句子组成摘要句集合$S$。不同的自动摘要技术对句子评分不同。

3.2 句子评分算法

本文选择支持向量回归（Support Vector Regression，SVR）模型预测文本中句子的得分。支持向量回归假设模型能够允许$f(x)$与y之间最多有ε的偏差，即仅当$f(x)$与y之间的差别绝对值大于ε时才计算损失，SVR问题可以形式化为：

$$\theta^* = argmin_\theta \frac{1}{2}(y-f(x))^2 + C\sum_{i=1}^{m} l_\varepsilon(f(x)-y)$$

其中x为特征值，C为正则化常数，l_ε为ε-不敏感损失（ε-insensitive loss）函数：

$$l_\varepsilon(Z) = \begin{cases} 0, & if\ |Z| \leqslant \varepsilon; \\ |Z| - \varepsilon, & otherwise \end{cases}$$

$f(x)$为回归函数：

$$f(x) = \theta^T x$$

考虑到特征与结果间可能是非线性的，支持向量回归的可通过个核函数将特征映射到高维空间，其计算公式：

$$f(x)=\theta^T \Phi(x)$$

其中，$\theta^T \in F$。

在实际的应用中，最常用的核函数有3种：多项式核、径向基（Radial Basis Function）核、多层感知机核等。

3.3 特征描述

3.3.1 基于文本内容的特征

（1）句子在全文的位置特征。一般来说，出现在全文开头和结尾的句子更可能包含重要信息。其计算公式：

$$\text{Weight(location)} = \frac{|N_{sentences} - 2\times Number_s|}{N_{sentences}}$$

其中$N_{sentences}$为文本所包含的句子的总数，$Number_s$为句子在文本中的编号，编号从1开始。

（2）句子在段落的位置特征。同样的，出现在段落开头和结尾的句子也可能包含重要信息。其计算方式同上，不过此时$N_{sentences}$为该段落所包含的句子的总数，$Number_s$为句子在段落中的编号，编号从1开始。

（3）句子长度特征。句子越长，则包含的实词越多，信息量越大，因而其权重也应该越高，但摘要的长度有限。综合考虑以上两点，给出句子长度计算：

$$\text{Weight(location)} = \frac{count(s)}{length(s)}$$

其中，$count(s)$指句子包含实词的个数，$length(s)$为句子的长度，包括实词和虚词。

（4）句子基于词的TF*IDF特征。不同的词在句子中的权重不同，本文利用信息检索中的TF*IDF模型计算文本中出现实词的得分，以此得到句子的权值：

$$\text{Weight(tfidf)} = \sum_{w_i \in s} tfidf(w_i)$$

其中w_i为句子中出现的实词，$tfidf(w_i)$为w_i在该篇

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

文本中TF-IDF值。

（5）句子基于标题相似度的特征。标题暗示文章的核心内容，文本中出现的句子与标题越相似，则越可能成为摘要句。据此基于标题相似度的特征计算公式：

Weight(title)=cosinesimi(s，title)

其中计算余弦相似度时，采用基于TF-IDF的布尔模型。

（6）句子的PageRank值。本文将每个句子当作图模型中的一个顶点，通过反复迭代，得到其PageRank值：

Weight(PageRank)=PageRank(s)

在计算PageRank的构建中，两个顶点间的相关度是TF-IDF的余弦相似度。

3.3.2 基于引文上下文的特征

本文提出了基于引文上下文的TF-IDF句子加权（C-TF-IDF）和基于引文上下文的相似度加权（C-Simi）两个特征。

引文上下文内容反映施引文献对被引文献内容的评价信息，通过对引文上下文中出现的词频进行统计，出现频率越高的词越能反映文章的主题。本文利用引文上下文的词频统计信息，对文献中出现的句子重新打分以获取一个权值。

由于在编写引文句时存在一些特定的格式和搭配方式，所以存在某些词在引文中出现的频率较高，但是并不能作为特征词存在。本文将一篇被引文献的所有引文上下文作为一类，利用TF*IDF的思路对出现在引文上下文中的词进行打分，根据这些词的TF-IDF对原始文本中的每个句子进行打分，计算公式如下：

$$\text{Weight}(C_{\text{tfidf}}) = \sum_{w_i \in s} TF * IDF_{citation}(w_i)$$

利用引文上下文对句子打分的另一种方法是计算句子与引文上下文之间的相似度，句子与引文上下文越相似，则越能反映文章的核心内容。之所以提出该特征是因为学术论文中的引用行为呈现高度复杂性，引用目的不同也导致作者对引文句的描述存在较大差别，在施引文献中的重要性也有较大差异。因此，本文使用PageRank算法对文献的每一个引文上下文构建图模型以计算其权重值，然后利用向量空间模型计算原文中的句子与每一个引文上下文的相似度，取最大值作为MaxSimilarity。基于此，句子基于引文上下文相似度的特征值计算公式：

$$\text{Weight}(CitationSimilarity)=MaxSimilarity*Weight_{pagerank}$$

在PageRank的计算过程中，顶点间的边通过引文上下文的TF*IDF相似度构建，而句子与引文上下文的相似度则通过原始文本的TF*IDF来计算的。

3.4 摘要抽取

如果两个句子所包含的词过于相似，它们所包含的特征值以及最终的得分也很可能相同。由于最终生成的摘要长度受到限制，所以通过回归模型预测句子的权重后还需对句子进行筛选，避免因信息冗余而降低摘要整体质量。本文采用基于最大边缘相关（Maximum Marginal Relevance，MMR）算法。（1）首先将所有的句子按其得分（由3.2节评分算法计算）进行排序；（2）依次选择句子，并将当前句子与已被选中的句子进行比较，如果当前句子与被选中的任一句子相似度较高（在本文中设定的相似度的阈值为0.7），则直接跳至下一个句子，反之则加入被选摘要句中；（3）以此类推，直至满足摘要所需长度。

自动摘要长度的设置一般有两种方式，分别是按比例抽取和固定长度抽取。根据摘要的粒度，又可分为词和句子两个层面。本文是直接抽取原文中固定数量的句子组成摘要，根据对语料库中文本原始摘要的统计结果，最终确定抽取句子的数目为4。

4 实验与评测

4.1 实验设置

4.1.1 数据集

本研究所使用的原始数据集为ACL（Association for Computational Linguistics）选集[32]，包含3.4万多篇从计算语言学和自然语言处理相关的各种期刊和会议中选择的英文文献。本文所使用的数据集是Schäfer等通过OCR等技术对ACL数据集进行识别得到的结构完整的XML文档数据[33]。文档包含完整的标题、摘要、章节、段落和参考文献等信息。

实验首先对每篇文档中的引文和参考文献进行配对，然后将参考文献与数据集中的文档进行匹配，获取

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

被引文献、引文上下文和施引文献间的关系图。本文随机抽取123篇被引次数在10次以上的文本构成被引文献和引文上下文数据集，然后对每篇文档的正文进行句子分割，并进行清洗和过滤，去掉句子中的噪音（包括引文标记等信息），剔除较短的句子片段（所含单词个数小于5）。最后对句子进行分词和词性标注。利用词性标注，对原文中的词进行筛选。本文只对文本中出现的名词、动词、形容词和副词进行统计分析，同时，利用词性标注也可以作为停用词的补充和扩展。

4.1.2 特征分析

对经过预处理的数据集进行特征抽取，实验选择的特征是上文描述的两类共8个特征，通过计算皮尔逊系数（Pearson correlation coefficient）和最大信息相关系数（Maximal information coefficient）来对比其预测句子得分的相关性。从图1可见，基于引文上下文的C-TF-IDF和C-Simi无论是皮尔逊系数还是最大信息相关系数均较高，尤其是基于引文上下文的C-TF-IDF特征的各项指数均明显高于其他特征，表明本文提出的特征在回归分析时与句子得分权重间存在较高相关性。

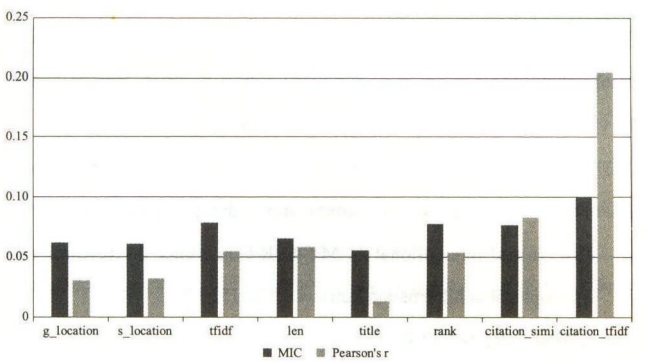

图 1 实验特征的皮尔逊系数和最大相关系数对比

4.2 结果与评测

当前自动摘要的评测方法有很多，但应用最为广泛的是ROUGE[34]，其利用n-gram在生成的摘要和参考摘要间的共现来判断摘要的质量，以召回率作为标准，主要包括ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-SU四种评测标准。

ROUGE-N是基于n-gram的召回率，即生成的摘要与人工生成的摘要中共同出现的n-gram在人工生成摘要之中所占的比值，计算公式：

$$R_n=\frac{\sum_{S\in\{ReferemceSummaries\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{ReferemceSummaries\}}\sum_{gram_n\in S}Count(gram_n)}$$

其中Reference Summaries为参考摘要，n为n-gram模型的长度，$Count_{match}(gram_n)$为自动摘要与参考摘要基于n-gram共现的个数最大值。

ROUGE-L（Longest Common Subsequence）的核心思想是如果自动摘要与参考摘要间存在最长公共子串，那么公共子串越长，两个摘要越相似，摘要效果越好。

ROUGE-W是在ROUGE-L的基础上考虑子串的连续性，一般来说，子串越连续，则其相似度越高。

ROUGE-SU则更进一步考虑到对于同一个二元词组，插入不同形容词或介词后，虽然句子的意思相近，但是ROUGE-2的得分明显低于ROUGE-1（譬如句子“我喜欢你”和“我非常喜欢你”），而ROUGE-SU正是考虑到词汇在摘要中存在跳跃时ROUGE-2的值。

本文选取WE-ROUGE-1，WE-ROUGE-2，WE-ROUGE-SU4指标对结果进行评测，是因为Flick等的研究表明这3个类别在自动摘要中的评价效果最好[34]，并将得到的结果分别与SumBasic和TextRank进行比较[35-36]。评测结果如图2所示（C1指引文上下文的TF-IDF特征，C2指引文上下文的相似度权重）。

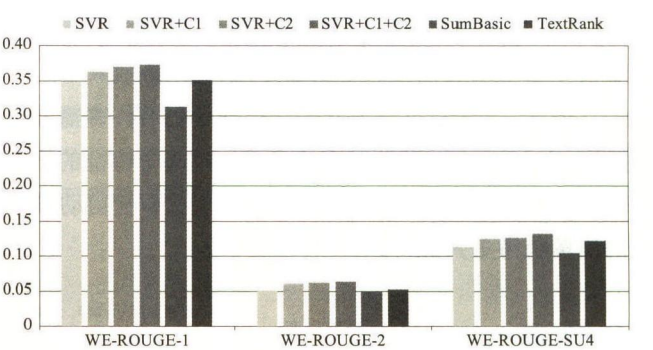

图 2 摘要在各个系统上的WE-ROUGE评测结果

从上述结果可以看出，本文提出的支持向量回归模型在WE-ROUGE测评上的表现均优于SumBasic。在使用回归模型抽取摘要的结果中，通过加入引文上下文的特征，WE-ROUGE的得分较BasicSum分别提升19.3%、26.9%和27.6%，较TextRank分别提升6.1%、7.2%和8.5%，其中基于引文上下文的TF-IDF构建的特征增幅最大。对两种特征综合考虑时，其对应WE-ROUGE的评分最高，这与特征分析的结果保持一致，

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

即引文上下文所包含的信息与文本摘要间存在较大的关联。值得关注的是，相比于只关注图模型的TextRank算法，SVR算法综合考虑句子的PageRank值以及词频统计等其他信息，但是在最终测评结果上，SVR的评分要略低于TextRank，可能是由于回归模型的整体的训练并没能达到一个较好的效果，还有改进空间。

5 结语

本文通过对文本抽取的相关特征构建回归模型来预测句子得分，在有序的句子列表中通过最大相关算法抽取句子生成摘要。相较于传统模型，摘要的质量有一定的提高，但仍有一些不足。

首先，本文选择被引10次以上的文献进行研究，是因为只有这些文献才有足够的引文上下文信息，然而，大量文献被引次数很少甚至没有被引，对于此类引文上下文稀疏的文献本文无法处理。

其次，本文直接以文献中的原始摘要作为实验最终的参照标准，但是不同作者在撰写摘要时侧重点存在差异，这就意味着针对同一篇文献，人工编写的摘要也存在较大差别，而将自动生成的结果直接参照原始文献的摘要进行评价，只能在一定程度上反映摘要的质量。也就是说，摘要的实验结果会略低于真实情况。

除此之外，本文在特征的选择上直接将句子的位置、长度等信息作为连续变量的做法，是否优于传统的离散变量或者布尔模型并未做更深入地探讨。

因此，基于引文上下文的自动摘要仍有许多需要改进和探讨的可能，若能从多个角度，对每个特征的研究和挑选做更细致地分析和扩充，或许能够使自动摘要的质量获得更大提升。

参考文献

[1] MAY R M. The scientific wealth of nations[J].Science,1997,275(5301):793-796.

[2] MALLIK A,MANDAL N.Bibliometric analysis of global publication output and collaboration structure study in microRNA research[J]. Scientometrics,2014, 98(3):2011-2037.

[3] LUHN H P. The automatic creation of literature abstracts[J]. Ibm Journal of Research & Development,1958, 2(2):159-165.

[4] NENKOVA A,MCKEOWN K. A survey of text summarization techniques[M]//AGGARWAL C C,ZHAI C X.Mining Text Data. Springer US, 2012:43-76.

[5] 刘洋,崔雷.引文上下文在文献内容分析中的信息价值研究[J].图书情报工作, 2014,58(6):101-104.

[6] KAPLAN D,IIDA R,TOKUNAGA T. Automatic extraction of citation contexts for research paper summarization: a coreference-chain based approach[C]//The Workshop on Text & Citation Analysis for Scholarly Digital Libraries.Association for Computational Linguistics,2009:88-95.

[7] BAXENDALE P B.Machine-made index for technical literature: an experiment[J].Ibm Journal of Research & Development,1958, 2(4):354-361.

[8] EDMUNDSON H P.New methods in automatic extracting [J].Journal of the ACM,1969,16(2):264-285.

[9] KUPIEC J,PEDERSEN J,CHEN F.A trainable document summarizer[C]// Sigir'95, Proceedings of the18th International ACM SIGIR Conference on Research and Development in Information Retrieval,July 9-13,1995, Seattle.Special Issue of the SIGIR Forum,1995:68-73.

[10] HIRAO T,ISOZAKI H,MAEDA E, et al. Extracting important sentences with support vector machines[C]//International Conference on Computational Linguistics.2010:342-348.

[11] ZHOU L,HOVY E. A Web-Trained extraction summarization system[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.2003,30(3): págs.331-336.

[12] MIHALCEA R.Graph-based ranking algorithms for sentence extraction, applied to text summarization[C]//Interactive Poster & Demonstration Sessions. Association for Computational Linguistics,2004:170-173..

[13] CHUANG W T,YANG J.Extracting sentence segments for text summarization: a machine learning approach[C]//In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,2000:152-159.

[14] AMINI M R,USUNIER N,GALLINARI P. Automatic text summarization based on word-clusters and ranking algorithms[J].Lecture Notes in Computer Science,2005, 3408:142-156.

[15] RUSH A M,CHOPRA S,WESTON J.A neural attention model for abstractive sentence summarization[J].Computer Science,2015.

[16] CAO Z,WEI F,LI S,et al.Learning Summary Prior Representation for Extractive Summarization[C]//Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing.2015.

[17] NALLAPATI R,ZHOU B,SANTOS C N D,et al.Abstractive text summarization using sequence-to-sequence RNNs and beyond[EB/OL]. (2016-04)[2016-07-01].http://arxiv.org/pdf/1602.06023v2.pdf.

[18] 刘挺,吴岩,王开铸.基于信息抽取和文本生成的自动文摘系统设计[J].

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

情报学报, 1997(S1):31-36.

[19] 吴中勤,黄萱菁,吴立德.基于有监督分类技术的文本自动摘要研究[C]//全国信息检索与内容安全学术会议.2005.

[20] WAN X,YANG J.Multi-document summarization using cluster-based link analysis[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval . New York:ACM,2008:299-306.

[21] 万小军.基于集成技术的多文档自动摘要[EB/OL].(2011-03-16)[2016-05-16]. http://www.paper.edu.cn/releasepaper/content/201103-710.

[22] Wan X. Update Summarization Based on Co-Ranking with constraints[C]// Coling:Posters,2012.

[23] 沈焕生,朱磊.基于信息抽取的自动摘要生成技术[C]//中国信息技术应用学术研讨会.2009年中国信息技术应用学术研讨会论文集,2009: 227-232.

[24] 罗文娟,马慧芳,何清,等.权衡熵和相关度的自动摘要技术研究[J].中文信息学报, 2011, 25(5):9-16.

[25] ELKISS A,SHEN S,FADER A, et al. Blind men and elephants:what do citation summaries tell us about a research article?[J].Journal of the American Society for Information Science & Technology,2008, 59(1):51-62.

[26] QAZVINIAN V,RADEV D R. Scientific paper summarization using citation summary networks[C]//International Conference on Computational Linguistics.Association for Computational Linguistics, 2008:689-696.

[27] TANDON N,JAIN A.Citation context sentiment analysis for structured summarization of research papers[C]//35th German Conference on Artificial Intelligence,September 24-27,2012,Saarbrucken.2012:98.

[28] COHAN A,GOHARIAN N. Scientific article summarization using citation-context and article's discourse structure[C]//Conference on Empirical Methods in Natural Language Processing.2015.

[29] GALGANI F,COMPTON P,HOFFMANN A.Summarization based on bi-directional citation analysis[J].Information Processing & Management,2015, 51(1):1-24.

[30] 胡珀.融合上下文信息的自动文摘研究[D].武汉:武汉大学, 2013.

[31] MEI Q,ZHAI C X. Generating impact-based summaries for scientific literature[C]//Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus. ACL,2008:816-824.

[32] RADEV D R,MUTHUKRISHNAN P,QAZVINIAN V,et al.The ACL anthology network corpus[J].Language Resources & Evaluation, 2013, 47(4):919-944.

[33] SCHÄFER U,WEITZ B. Combining OCR outputs for logical document structure markup: technical background to the ACL 2012 contributed task[C]//Acl-Special Workshop on Rediscovering 50 Years of Discoveries. Association for Computational Linguistics, 2012:104-109.

[34] FLICK C. ROUGE: A package for automatic evaluation of summaries[C]// The Workshop on Text Summarization Branches Out.2004:25-26.

[35] VANDERWENDE L,SUZUKI H,BROCKETT C, et al. Beyond SumBasic:task-focused summarization with sentence simplification and lexical expansion[J]. Information Processing & Management, 2007, 43(6):1606-1618.

[36] MIHALCEA R,TARAU P.TextRank:bringing order into texts[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, July 25-26 2004, Barcelona. 2004:404-411.

作者简介

陈海华，男，1990年生，硕士研究生，研究方向：信息检索、知识挖掘等，E-mail：chh2014@whu.edu.cn。
黄永，男，1991年生，博士研究生，研究方向：信息检索、机器学习等。
张炯，男，1991年生，硕士研究生，研究方向：信息检索、知识挖掘等。
陆伟，男，1974年生，武汉大学信息管理学院教授，副院长，研究方向：信息检索、知识管理、数据挖掘等，E-mail：reedwhu@gmail.com。

Research on Citation Context Based on Automatic Summarization of Academic Literature

CHEN HaiHua[1], HUANG Yong[1], ZHANG Jiong[1], LU Wei[1,2]
(1.School of Information Management, Wuhan University, Wuhan 430072, China;
2. Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072, China)

Abstract: Text summarization of academic literature refers to automatically generate abstract for a given paper. With the aid of automatic summarization, authors can improve the efficiency of writing and reading academic literature. Existing works evaluate and rank the sentences based on term frequency, they can' t reveal the main idea of an article from a deeper semantic dimension. Based on previous research, this article introduces citation context as an enhanced feature. Combining it with other existing features, we conduct an automatic re-scoring of each sentence by utilizing support vector regression (SVR) model. A significant improvement over traditional term frequency-based and graph-based method based on WE-ROUGE shows the effectiveness of citation context in automatically text summarization.

Keywords: Automatic Summarization; Citation Context; Support Vector Regression; Word Embedding

（收稿日期：2016-08-04）

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net