第 46 卷 第 7 期 计 算 机 工 程 2020 年 7 月
Vol. 46 No. 7 Computer Engineering July 2020

·人工智能与模式识别· 文章编号：1000-3428(2020)07-0058-07 文献标志码：A 中图分类号：TP391

基于布谷鸟搜索优化算法的多文档摘要方法

周诗源，王英林

（上海财经大学 信息管理与工程学院，上海 200433）

摘 要： 为最大化生成摘要的信息量，提出一种基于布谷鸟搜索（CS）算法与多目标函数的多文档摘要方法。对多文档数据进行预处理，通过句子分割、分词、移除停用词和词干化将文档转化为词语的基本处理形式，计算经数据预处理后的句子信息量得分并将其作为 CS 算法的输入，再基于多目标函数生成包含原始文档重要信息的句子以组成最终的摘要。实验结果表明，与基于粒子群优化算法和双层 K 最近邻算法的多文档摘要方法相比，该方法在最大化生成摘要信息量的前提下，保证了高可读性和低冗余性，并且在 DUC 基准数据集上的摘要平均准确度高达 0.99。

关键词： 多文档摘要；布谷鸟搜索算法；数据预处理；多目标函数；信息量

开放科学（资源服务）标志码（OSID）：

中文引用格式： 周诗源，王英林. 基于布谷鸟搜索优化算法的多文档摘要方法[J]. 计算机工程，2020，46(7)：58-64，71.

英文引用格式： ZHOU Shiyuan，WANG Yinglin. Multiple document summarization method based on optimized Cuckoo search algorithm[J]. Computer Engineering，2020，46(7)：58-64，71.

Multiple Document Summarization Method Based on Optimized Cuckoo Search Algorithm

ZHOU Shiyuan，WANG Yinglin

（School of Information Management and Engineering，Shanghai University of Finance and Economics，Shanghai 200433，China）

【Abstract】 To maximize the amount of information of generated summary，this paper proposes a multiple document summarization method based on the Cuckoo Search（CS） algorithm and multiple objective function. The method pre-processes data of multiple documents by using sentence segmentation，word segmentation，removal of stop words and word drying to transform the documents into a basic processed form of words. Then the score of information amount of pre-processed sentences is calculated to serve as the input of the CS algorithm. Based on the multiple objective function，the sentences including key information of original texts are generated to form the ultimate summarization. Results show that compared with multiple document summarization methods based on Particle Swarm Optimization（PSO） algorithm and Double-layer K Nearest Neighbor（DKNN） algorithm，the proposed summarization method maximizes the amount of information in the generated summary while keeping high readability and low redundancy. Its average accuracy rate on the DUC benchmark dataset reaches 0.99.

【Key words】 multiple document summarization；Cuckoo Search（CS） algorithm；data preprocessing；multiple objective function；amount of information

DOI： 10.19678/j.issn.1000-3428.0054780

0 概述

随着移动互联网的快速发展，网络中的信息量呈指数级增长，大量资讯、知识、视频、音乐等资源可供用户选择，信息过载问题日益突出，而文档摘要是解决信息过载的有效方式。文档摘要[1]是指在不丢失原文主要内容的前提下创建摘要的过程[2]，其能够提供相关信息的快速导向，帮助用户确定文档是否具备可读性，因此成为近年来研究的热点问题。

根据摘要生成过程中的文档数量，可将摘要分为单文档摘要[3]和多文档摘要[4]，由于多文档摘要的搜索空间大于单文档摘要，因此提取重要句子的难度更大。为生成包含原始文档重要信息句的最优

基金项目： 国家自然科学基金（61375053）。

作者简介： 周诗源（1982—），男，博士研究生，主研方向为语义分析、文档摘要、机器学习；王英林，教授、博士生导师。

收稿日期： 2019-04-30 **修回日期：** 2019-07-18 **E-mail：** zhou.shiyuan@foxmail.com

摘要,文献[5]设计基于整数线性规划的概括式多文档自动摘要方法,优选出每个主题下的重要主题语义信息,生成新的摘要句,并对候选摘要句进行润色加工,解决了生成概括式摘要的信息覆盖和可读性问题。文献[6]提出一种基于 K 最近邻句子-图模型的动态文本摘要方法,根据 K 近邻规则构建一个双层句子图模型,使用基于密度划分的增量图聚类方法对句子进行子主题划分,并结合时间因素提高句子新颖度以抽取动态文摘。文献[7]使用粒子群优化(Particle Swarm Optimization,PSO)算法设计单文档摘要器,通过内容覆盖度和冗余度特征的加权平均,获得目标函数,进而设计文档摘要器。文献[8]提出一种句子话题重要性计算方法,通过分析句子与话题的语义关系,判断句子是否描述了话题的重要信息,同时设计一种基于排序学习的半监督训练框架。此外,现有研究采用 PSO 算法[9]、差分进化(Differential Evolution,DE)算法[10]、遗传算法(Genetic Algorithm,GA)[11]、LDA 主题模型[12]及子主题划分模型[13]等自动生成文档摘要。鉴于布谷鸟搜索(Cuckoo Search,CS)算法[14-15]在多目标优化问题中的性能优势,本文提出基于 CS 算法的多文档摘要方法,主要过程包括预处理、输入表示和摘要表示。

1　多文档摘要

多文档摘要是从多个文档中自动创建一个简明文档(称为摘要)的过程。多文档摘要过程具体包括预处理、输入表示和摘要表示 3 个步骤,处理流程如图 1 所示。摘要系统的输入为多个文档,先对这些文档进行预处理,再通过输入表示和摘要表示提取最终摘要。

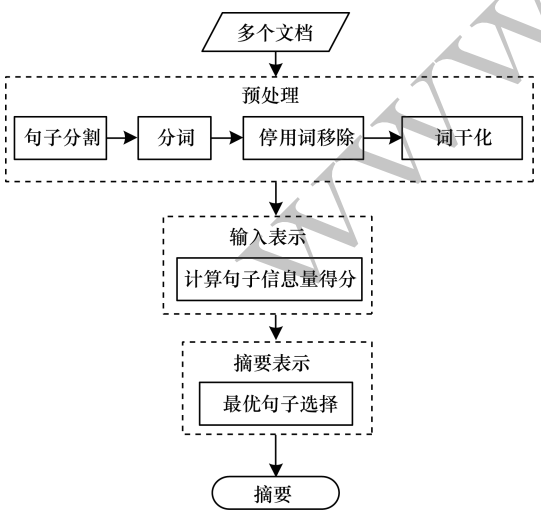

图 1　多文档摘要处理流程

Fig. 1　Procedure of multiple document summary processing

1.1　预处理

预处理流程如图 2 所示,具体步骤如下:

1)句子分割:从输入的文档集合中,将每个文档 D 单独分割为 $D=\{S_1,S_2,\cdots,S_n\}$,其中,S_j 表示文档中的第 j 个句子,n 表示文档中的句子数量。

2)分词:将每个句子的术语标记为 $T=\{t_1,t_2,\cdots,t_m\}$,其中,$t_k(k=1,2,\cdots,m)$ 表示 D 中出现的不同术语,m 表示术语数量。

3)停用词移除:将文档中出现频率较高但没有实际检索作用的词进行移除,如“的”“了”“在”等。

4)词干化:将句子转换为词的基本形式。

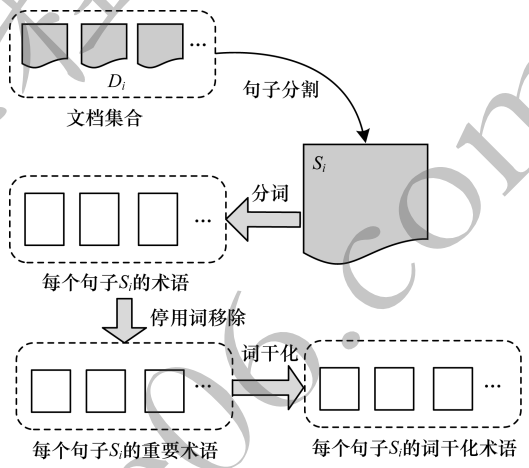

图 2　预处理流程

Fig. 2　Procedure of preprocessing

1.2　输入表示

在输入表示阶段,使用预处理后的数据计算每个句子的权重(术语频率之和),即句子信息量得分,将句子信息量得分作为算法输入,其流程如图 3 所示。

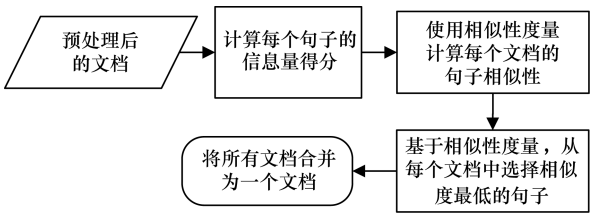

图 3　输入表示流程

Fig. 3　Procedure of input representation

1.3　摘要表示

摘要表示的目的是为文档集合生成包含有用信息的摘要。在最优句子的选择过程中,基于预定义阈值对 CS 优化算法得出的句子信息量得分进行比较,选出能够代表摘要的重要句子,其流程如图 4 所示。

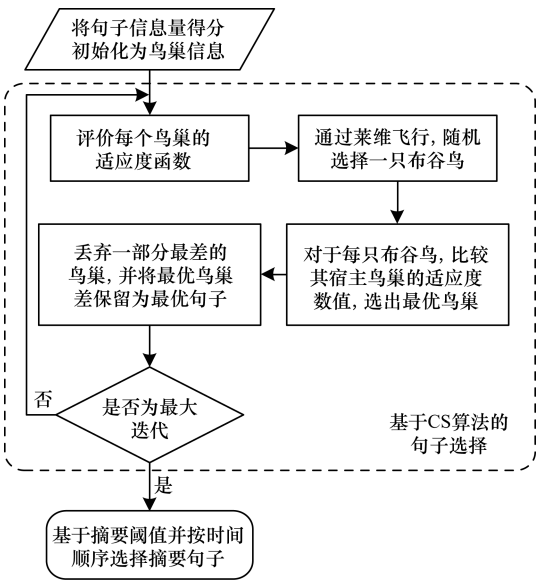

图 4 摘要表示流程

Fig. 4 Procedure of summary representation

2 基于 CS 优化算法的多文档摘要

2.1 CS 优化算法

CS 算法[16]是一种启发式进化算法。布谷鸟是一种具有激进繁育策略的鸟，成年布谷鸟将卵产在其他鸟类的鸟巢中，由宿主鸟代为孵化和育雏。CS 算法将鸟巢视为候选解，每个布谷鸟仅可产一枚卵，表示一个新的候选解。标准 CS 算法包含 3 个理想化规则：1）每个布谷鸟在一个随机鸟巢中产一枚卵，代表一个解集；2）鸟巢中包含的最优卵将传递至下一代；3）可用鸟巢数量固定，一只宿主鸟发现一枚寄生卵的概率为 P_a。当满足条件时，宿主鸟可以丢弃寄生卵或放弃其鸟巢并在其他地方新建一个鸟巢。

在实际应用时可使用 CS 算法的最简单形式，其中每个鸟巢只有一枚卵，在此情况下无需区分鸟巢、卵和布谷鸟，因为鸟巢、卵和布谷鸟均为一一对应关系，而且该算法可扩展至更复杂的情况，其中每个鸟巢中存在多个卵，代表一个解集。

在生成新的解 x_i^{t1} 时，使用局部随机游走和全局随机游走的组合形式，其通过参数 P_a 进行控制。局部随机游走可以表示为：

$$x_i^{t+1}=x_i^t+\alpha\times S\otimes H(P_a-\varepsilon)\otimes(x_j^t-x_k^t) \tag{1}$$

其中，x_j^t 和 x_k^t 为通过随机置换选出的两个不同解，$H(u)$ 表示亥维赛函数，ε 表示从均匀分布中提取的随机数，α 表示步长。

另外，使用莱维飞行执行全局随机游走，其全局收敛性已得到证明[16]。莱维飞行包含连续随机步[17]，其特征为一连串快速跳跃，计算公式为：

$$x_i^{t+1}=x_i^t+\alpha\otimes\mathrm{Levy}(\lambda) \tag{2}$$

其中，α 表示步长，其与优化问题的规模成正比（即 $\alpha>0$），$\otimes$ 表示乘法中的逐项移动，$\mathrm{Levy}(\lambda)$ 表示从莱维分布中提取的随机数。

本文基于 CS 优化算法的多文档摘要方法具体步骤如下：

步骤 1 采集一个多文档集合 M，$M=\{D_1,D_2,\cdots,D_N\}$，其中，每个 D_i 表示集合 M 的单个文档，D_i 的长度表示为句数，不同文档包含不同句数。

步骤 2 使用句子分割、分词、停用词移除和词干化对每个文本文档 D_i 进行预处理。

步骤 3 使用式（3）计算出每个预处理后文档 D_i 的句子 S_j 的信息量得分 IS_{jk}，即通过词频之和推导出的句子权重。

$$\mathrm{IS}_{jk}=\mathrm{tf}_{jk}\times\log_a(n/n_k) \tag{3}$$

其中：IS_{jk} 表示每个句子 S_j 相对于术语 t_k 的信息量得分；tf_{jk} 表示术语频率，即术语 t_k 在句子 S_j 中的出现次数；n_k 表示包含术语 t_k 的句子数量；$\log_a(n/n_k)$ 表示用于句子检索的向量空间模型中的逆句频率。

步骤 4 使用式（4）计算出预处理后文档 D_i 的句间相似度。

$$\mathrm{Sim}(S_i,S_j)=\frac{\sum_{k=1}^{m}\mathrm{IS}_{ik}\mathrm{IS}_{jk}}{\sqrt{\sum_{k=1}^{m}\mathrm{IS}_{ik}^2\sum_{k=1}^{m}\mathrm{IS}_{jk}^2}},i,j=1,2,\cdots,n \tag{4}$$

步骤 5 基于相似度阈值，选择每个 D_i 中相似度最低的句子。

步骤 6 将选出的每个 D_i 中所有相似度最低的句子合并为单个文档 D_{input}。

步骤 7 初始化 CS 参数，例如种群大小、寄生卵被发现率（P_a）、步长因子（S_f）和莱维指数（λ）。

步骤 8 在指定搜索空间内，将句子信息量得分作为每只布谷鸟的鸟巢信息，每个鸟巢对应给定优化问题的一个候选解。

步骤 9 使用式（3）针对给定问题计算每个鸟巢的适应度函数 f_i。

步骤 10 利用式（2）得到新的鸟巢种群。

步骤 11 计算出与新鸟巢相对应的适应度函数 f_j，并将其与旧鸟巢的适应度函数 f_i 进行比较。

步骤 12 若 f_j 优于 f_i，则使用新候选解替换旧候选解。

步骤 13 在新种群中，选择 P_a 性能较差的一部分鸟巢，将这些鸟巢替换为指定搜索空间内随机生成的新鸟巢。

步骤 14 为新鸟巢计算适应度函数。

步骤 15　基于适应度数值,记录当前种群集合中性能最优的鸟巢。之后,将这些鸟巢与前代最优鸟巢相比,选择其中性能最优的鸟巢。

步骤 16　若未达到最大迭代次数,则返回步骤 9。

步骤 17　基于预定义的句子相似度阈值从文档中按时间顺序选择句子。

CS 算法的时间复杂度计算主要集中于适应度函数 f 的计算[18],在本文中对文档预处理过程为步骤 1 ~步骤 6,适应度函数包括旧鸟巢的适应度函数 f_i(步骤 9)以及新鸟巢相对应的适应度函数 f_j(步骤 10、步骤 14)。当适应度函数 f 的自变量阶数比 n 高时(n 表示文档中的句子数量),算法执行一次的时间复杂度为 $O(f(n))$,当适应度函数 f 的自变量阶数与 n 相同或低于 n 时,时间复杂度为 $O(n)$。在考虑终止条件的情况下,总体时间复杂度为 $O(f(n)+n)$。CS 算法的空间复杂度较低,假设一个句子的存储空间为一个存储单元,对于包含 n 个句子的文档,执行文档摘要操作的空间复杂度为 $O(n)$。

2.2　目标函数建立

文本摘要的目标是最大化生成摘要的信息量,降低冗余并保持可读性。因此,本文基于多个目标,从文档集合中建立摘要,通过内容覆盖度、衔接性和可读性以优化摘要,即创建 3 个子函数 $f_{\text{cov}}(S)$、$f_{\text{coh}}(S)$ 和 $f_{\text{read}}(S)$,构成目标函数 $f(S)$。

$$f(S)=f_{\text{cov}}(S)+f_{\text{coh}}(S)+f_{\text{read}}(S) \tag{5}$$

一个摘要中应包含相关句子集合以覆盖文档集合的主要内容,并通过信息量得分最高的句子反映文档的主要内容,因此摘要的内容覆盖度 $f_{\text{cov}}(S)$ 定义为:

$$f_{\text{cov}}(S)=\text{Sim}(S_i,O),i=1,2,\cdots,n \tag{6}$$

其中,O 表示句子集的主要内容中心,$O=\{O_1,O_2,\cdots,O_n\}$,O_i 为每个文档的句子加权平均数。通过对 S_i 和 O 之间的相似度进行评价,以衡量句子的重要程度。若相似度数值越高,则内容覆盖度越高。

摘要句子之间的衔接性是指句子层面和段落层面的概念联系,有助于更好地理解全文。若摘要衔接性 $f_{\text{coh}}(S)$ 数值越高,则意味着句子间的联系越紧密,其定义为:

$$f_{\text{coh}}(S)=1-\text{Sim}(s_i,s_j),i,j=1,2,\cdots,n \text{ 且 } i\neq j \tag{7}$$

摘要可读性是指从集合 D 中选出能够最大化句间关系的子集 S。若 $f_{\text{read}}(S)$ 的数值越高,则摘要的可读性越强,其定义为:

$$f_{\text{read}}(S)=\text{Sim}(s_i,s_j),i,j=1,2,\cdots,n \text{ 且 } i\neq j \tag{8}$$

3　实验结果与分析

通过实验对本文多文档摘要方法进行性能测试,并在 DUC 开源基准数据集(DUC2006 和 DUC2007)上比较本文方法与基于双层 K 最近邻(Double-layer K Nearest Neighbor,DKNN)算法[6]和 PSO 算法[9]的多文档摘要方法的性能。实验环境为 MATLAB 2011b,实验平台为 4 核 Intel 酷睿 i5 处理器,3.2 GHz 内存,Windows 7 操作系统,并且使用 ROUGE 工具对摘要结果的 ROUGE 得分进行分析。

3.1　实验数据集

本文使用 DUC 开源基准数据集对文本摘要结果进行评价。DUC 数据集的参数设置如表 1 所示。在数据预处理阶段,通过比较可用的停用词表,从原始文档中移除不重要的词语,并使用词干分析器提取术语词干。

表 1　DUC 数据集参数设置

Table 1　Parameter setting of DUC dataset

参数	DUC2006 数据集	DUC2007 数据集
聚类数量	50	45
每个聚类中的文档数	25	25
文档平均句数	30.12	37.5
文档最大句数	79	125
文档最小句数	5	9
数据源	AQUAINT	AQUAINT
摘要长度(词数)	250	250

3.2　实验控制参数

由于控制参数以应用为导向,因此不存在固定赋值,而是通过大量仿真实验推导出参数值,本文方法、基于 DKNN 和 PSO 的多文档摘要方法的参数设置如表 2 所示,其中:$V_{\min}$、$V_{\max}$ 分别表示粒子群的最小速度和最大速度;C_1、C_2 为加速系数,分别表示粒子向自身极值和全局极值推进的加速权值;W 表示惯性权值;SMP 表示 DKNN 的最优 k 值;CDC 表示步长因子;SRD 表示距离因子;MR 表示两层 KNN 的混合率;w 表示 DKNN 的近邻样本权重;C 表示默认近邻对象个数。

表 2　3 种多文档摘要方法的参数设置

Table 2　Parameter setting of three multiple document summarization methods

基于 PSO 的多文档摘要方法		基于 DKNN 的多文档摘要方法		本文多文档摘要方法	
参数	参数值	参数	参数值	参数	参数值
种群大小	50 个文档	种群大小	50 个文档	种群大小	50 个文档
C_1	[0,2]	SMP	3	P_a	0.75
C_2	[0,2]	CDC	0.2	S_f	0.5
$V_{\min}$、$V_{\max}$	[0,1]	SRD	0.2	λ	0.8
W	0.45	MR	0.5		
		w	0.5		
		C	4		

3.3 摘要评价指标

本文使用敏感度(Sen,又称真阳性率)、阳性预测值(Positive Predictive Value,PPV)和摘要准确度(S_{acc})指标对摘要方法进行评价。指标的定义基于候选摘要 $Cand_{sum}$(使用本文方法生成的摘要)、参考摘要 Ref_{sum}(人工生成的摘要)、真实句子 $True_{sen}$($Cand_{sum}$和 Ref_{sum}中共同出现的句子)和不重要句子 LS_{sen}($Cand_{sum}$或 Ref_{sum}中均未出现的句子)。

敏感度定义为:

$$Sen=\frac{|True_{sen}|}{|True_{sen}|+|Ref_{sum}|} \tag{9}$$

其中,$|True_{sen}|$表示真实句子的长度,$|Ref_{sum}|$表示参考摘要的总长度,参考摘要为人工生成,可信度高,用于评估来源摘要。式(9)还可看作是真阳性样本数与真阳性样本数和假阴性样本数的比值。敏感度反映了摘要方法生成的摘要质量,其值越大表示质量越高。

阳性预测值定义如下:

$$PPV=\frac{|True_{sen}|}{|True_{sen}|+|Cand_{sum}|} \tag{10}$$

其中,$|True_{sen}|$表示真实句子的长度,$|Cand_{sum}|$表示候选摘要的总长度。式(10)还可以看作是真阳性样本数与真阳性样本数和假阳性样本数的比值。阳性预测值表示生成的摘要与原始文档集合相同的概率,高相似度和低相似度的摘要均不会被记录在内,因此该指标为针对特定情况。

摘要准确度用于综合评价本文方法生成摘要的准确性,是最重要的评价指标。摘要准确度越高,表明生成的摘要与原始文档集合的相似性越高,其定义如下:

$$S_{acc}=\frac{|True_{sen}|+|LS_{sen}|}{|True_{sen}|+|LS_{sen}|+|Ref_{sum}|+|Cand_{sum}|} \tag{11}$$

3.4 ROUGE 评价

本文使用文献[19]开发的 ROUGE-1.5.5 工具包作为文本摘要的评价度量工具。ROUGE 包括 ROUGE-L、ROUGE-N 和 ROUGE-S 等多种工具,用于衡量生成摘要与人工摘要之间的文法(N-gram)匹配。在工具包中,ROUGE-N 度量与比较两种摘要的 N-gram 并计算匹配数量:

$$ROUGE\text{-}N=\frac{\sum_{S\in Sum}\sum_{N\text{-}gram\in S}Count_{match}(N\text{-}gram)}{\sum_{S\in Sum}\sum_{N\text{-}gram\in S}Count(N\text{-}gram)} \tag{12}$$

其中,N 表示 N-gram 的长度,$Count_{match}$(N-gram)表示候选摘要和参考摘要中同时出现的 N-gram 最大数量,Count(N-gram)表示参考摘要中的 N-gram 数量。

本文使用 ROUGE-N(ROUGE-1 和 ROUGE-2)度量对摘要性能进行评价,其度量与人工判断存在高度相关性。ROUGE-1 衡量系统生成摘要和人工创建摘要之间的一元分词的重叠情况,ROUGE-2 则比较二元分词的重叠情况[20],通过摘要的内容覆盖度、衔接性和文本可读性完成 ROUGE-N 的评价。ROUGE 度量值越高,表明其生成的摘要与原始文档集合的相似度越高。

表 3 给出了基于 PSO 的多文档摘要方法、基于 DKNN 的多文档摘要方法和本文多文档摘要方法在 DUC2006 和 DUC2007 数据集上,ROUGE-N(ROUGE-1 和 ROUGE-2)的 F 度量值的最差值、平均值和最优值统计结果。通过比较 DUC2006 数据集的 F 度量值可以看出,对于这 3 种方法,ROUGE-1 的最优 F 度量值为 0.411 27~0.431 10,ROUGE-2 的最优 F 度量值为 0.078 40~0.139 86。在 DUC2007 数据集上,ROUGE-1 的最优 F 度量值为 0.409 67~0.424 30,ROUGE-2 的最优 F 度量值为 0.076 20~0.103 40。虽然该数值取决于所使用的数据,但是可以明显看出,本文多文档摘要方法在两个数据集的 ROUGE 得分中均得到了较高的 F 度量值,这充分说明了 CS 算法在摘要表示过程中的特征搜索优势,相比于 PSO 算法,可以得到更优的搜索解。

表 3 3 种多文档摘要方法基于 ROUGE-N 的 F 度量值比较
Table 3 Comparison of F measure value based on ROUGE-N of three multiple document summarization methods

数据集	度量方法	摘要方法	F 度量值		
			最差值	平均值	最优值
DUC2006	ROUGE-1	PSO	0.390 87	0.400 9	0.411 27
		DKNN	0.400 30	0.407 0	0.422 90
		本文	0.404 22	0.411 5	0.431 10
	ROUGE-2	PSO	0.058 48	0.065 1	0.078 40
		DKNN	0.071 40	0.083 1	0.090 33
		本文	0.076 77	0.086 4	0.139 86
DUC2007	ROUGE-1	PSO	0.391 60	0.399 1	0.409 67
		DKNN	0.390 80	0.409 8	0.420 70
		本文	0.400 00	0.411 6	0.424 30
	ROUGE-2	PSO	0.074 30	0.075 8	0.076 20
		DKNN	0.080 90	0.088 1	0.089 03
		本文	0.081 70	0.089 2	0.103 40

表 4 给出了基于 PSO 的多文档摘要方法、基于 DKNN 的多文档摘要方法和本文多文档摘要方法在 DUC2006 和 DUC2007 数据集上,ROUGE-N(ROUGE-1 和 ROUGE-2)度量的召回率、精度和 F 度量值比较。

表 5 给出了基于 PSO 的多文档摘要方法、基于 DKNN 的多文档摘要方法和本文多文档摘要方法在 DUC2006 和 DUC2007 数据集上的敏感度、PPV 和摘要准确度比较。通过对 ROUGE-N 的文档摘要度量和指标进行分析得出,与基于 PSO 的多文档摘要方法和基于 DKNN 的多文档摘要方法相比,本文多文档摘要方法在两个数据集上均得到了更好的 ROUGE-1 和 ROUGE-2 结果,其原因为本文所采用的 CS 算法更加适用于多文档摘要的特征搜索,ROUGE-N 的 F 度量值不受召回率和精度的影响,并且摘要准确度也不受敏感度和 PPV 的影响。

表 4　3 种多文档摘要方法的精度、召回率和 F 度量值比较

Table 4　Comparison of precision, recall and F measure value of three multiple document summarization methods

数据集	度量方法	摘要方法	精度	召回率	F 度量值
DUC2006	ROUGE-1	PSO	0.384 91	0.441 51	0.411 27
		DKNN	0.415 20	0.430 98	0.422 90
		本文	0.425 80	0.436 55	0.431 10
	ROUGE-2	PSO	0.074 69	0.082 55	0.078 40
		DKNN	0.082 71	0.099 50	0.090 33
		本文	0.161 29	0.123 46	0.139 86
DUC2007	ROUGE-1	PSO	0.378 25	0.446 79	0.409 67
		DKNN	0.386 62	0.461 58	0.420 70
		本文	0.395 10	0.458 30	0.424 30
	ROUGE-2	PSO	0.069 70	0.084 10	0.076 20
		DKNN	0.085 90	0.092 40	0.089 03
		本文	0.098 24	0.109 30	0.103 40

表 5　3 种多文档摘要方法的敏感度、PPV 和准确度比较

Table 5　Comparison of sensitivity, PPV and accuracy of three multiple document summarization methods

数据集	摘要方法	评价度量		
		敏感度	PPV	准确度
DUC2006	PSO	0.500 0	0.400 0	0.973 4
	DKNN	0.560 0	0.529 4	0.980 0
	本文	0.600 0	0.570 8	0.990 0
DUC2007	PSO	0.500 0	0.352 9	0.980 8
	DKNN	0.583 3	0.500 0	0.990 4
	本文	0.620 0	0.540 0	0.995 1

3.5　衔接性分析

衔接性是自动摘要的重要属性,一般是指句子之间(甚至一个句子的不同部分)的联系程度能够支持读者理解其中的含义。摘要中的衔接性并不仅指句子的语法正确性,还包括句子层面和段落层面的联系。由此可知,前后句的衔接有助于更好地理解完整文本,通常使用余弦相似度计算摘要衔接性得分。图 5 给出了 3 种多文档摘要方法在 DUC2006 和 DUC2007 数据集上的衔接性得分。可以看出,在两个数据集上,本文多文档摘要方法的衔接性得分均高于基于 DKNN[6] 和 PSO[9] 的多文档摘要方法。

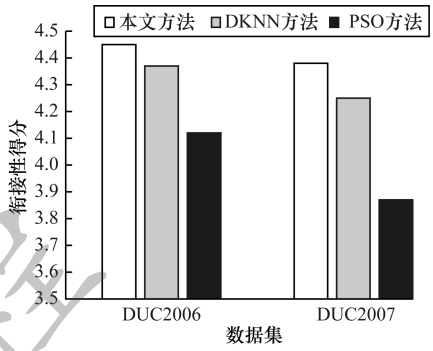

图 5　3 种多文档摘要方法的衔接性得分比较

Fig. 5　Comparison of the cohesiveness scores of three multiple document summarization methods

3.6　可读性分析

可读性是内容的可辨识和理解程度,取决于句子平均长度、句中包含的新词数量以及文章中使用语言的语法复杂度等因素。常用的可读性评价指标有 Flesh Kincaid 难度级数(FKGL)、Gunning fog(FOG)得分、SMOG 指数、Coleman Liau(CL)指数和自动易读性指数(Automatic Readability Index, ARI)等[21]。以上评价指标都是数值越高,可读性得分越高,即生成的摘要更易于阅读和理解。图 6 和图 7 分别给出了在 DUC2006 和 DUC2007 数据集上 3 种多文档摘要方法的可读性得分。可以看出,在 DUC2006 数据集上,本文多文档摘要方法在 CL、FOG、SMOG 和 ARI 指标上的得分均优于基于 PSO 和 DKNN 的多文档摘要方法,在 FKGL 指标上 3 种多文档摘要方法得到了几乎相同的结果。在 DUC2007 数据集上,本文多文档摘要方法的可读性得分均高于其他两种方法。

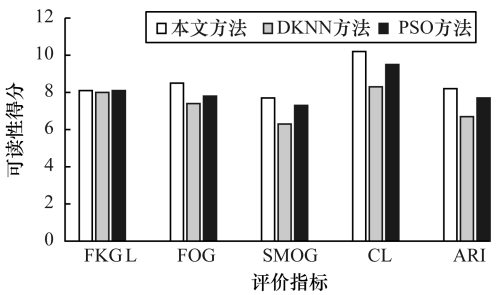

图 6　3 种多文档摘要方法在 DUC2006 数据集上的可读性得分比较

Fig. 6　Comparison of the readability scores of three multiple document summarization methods on DUC2006 dataset

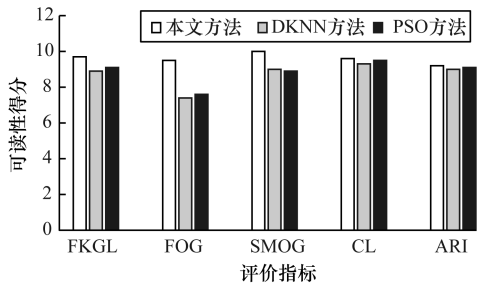

图 7 3 种多文档摘要方法在 DUC2007 数据集上的可读性得分比较

Fig. 7 Comparison of the readability scores of three multiple document summarization methods on DUC2007 dataset

4 结束语

本文提出一种基于布谷鸟搜索优化算法的多文档摘要方法,先对输入的多个文档进行预处理,再通过输入表示和摘要表示提取文档摘要,并对摘要的内容覆盖度、衔接性和可读性进行优化。实验结果表明,与基于 PSO 和 DKNN 的多文档摘要方法相比,本文多文档摘要方法能最大化摘要信息量,降低冗余并保持可读性。后续将优化布谷鸟搜索算法的参数设置,并使用模拟退火算法、蚁群算法等启发式算法进一步提升本文多文档摘要方法的整体性能。

参考文献

[1] GONG Yihong, LIU Xin. Generic text summarization using relevance measure and latent semantic analysis[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2001: 19-25.

[2] AZHARI M, KUMAR Y J, GOH O S, et al. Automatic text summarization: soft computing based approaches[J]. Advanced Science Letters, 2018, 24(2): 1206-1209.

[3] LIU Haiyan, ZHANG Yu. Chinese single document summarization based on LexRank[J]. Journal of Ordnance Equipment Engineering, 2017, 38(6): 85-89. (in Chinese)
刘海燕, 张钰. 基于 LexRank 的中文单文档摘要方法[J]. 兵器装备工程学报, 2017, 38(6): 85-89.

[4] GAMBHIR M, GUPTA V. Recent automatic text summarization techniques: a survey[J]. Artificial Intelligence Review, 2017, 47(1): 1-66.

[5] LIU Ziping. Study on multi-document summarization algorithm based on fusing topic sentences[D]. Chongqing: Chongqing University, 2016. (in Chinese)
刘子平. 基于主题句语义融合的多文档摘要算法研究[D]. 重庆: 重庆大学, 2016.

[6] GUO Hairong, ZHANG Hui, ZHAO Xujian, et al. Update summarization using incremental graph clustering[J]. Application Research of Computers, 2016, 33(7): 2034-2038. (in Chinese)
郭海蓉, 张晖, 赵旭剑, 等. 基于增量图聚类的动态多文档摘要算法[J]. 计算机应用研究, 2016, 33(7): 2034-2038.

[7] RAUTRAY R, BALABANTARAY R C. Comparative study of DE and PSO over document summarization[M]//JAINL C, PATNAIK S, ICHALKARANJE N. Intelligent computing, communication and devices. Berlin, Germany: Springer, 2014: 371-377.

[8] YING Wenhao, LI Sujian, SUI Zhifang. A topic-sensitive extractive method for multi-document summarization[J]. Journal of Chinese Information Processing, 2017, 31(6): 155-161. (in Chinese)
应文豪, 李素建, 穗志方. 一种话题敏感的抽取式多文档摘要方法[J]. 中文信息学报, 2017, 31(6): 155-161.

[9] GAO Ling, SHEN Yuan, GAO Ni, et al. Clustering analysis of vulnerability information based on text mining[J]. Journal of Southeast University (Natural Science Edition), 2015, 45(5): 845-850. (in Chinese)
高岭, 申元, 高妮, 等. 基于文本挖掘的漏洞信息聚类分析[J]. 东南大学学报(自然科学版), 2015, 45(5): 845-850.

[10] KARWA S, CHATTERJEE N. Discrete differential evolution for text summarization[C]//Proceedings of International Conference on Information Technology. Washington D. C., USA: IEEE Press, 2015: 1201-1209.

[11] XIONG Shufeng, JI Donghong. Query-focused multi-document summarization using hypergraph-based ranking[J]. Information Processing and Management, 2016, 52(4): 670-681.

[12] LIU Na, LU Ying, TANG Xiaojun, et al. Multi-document summarization algorithm based on significance topic of LDA[J]. Journal of Frontiers of Computer Science and Technology, 2015, 9(2): 242-248. (in Chinese)
刘娜, 路莹, 唐晓君, 等. 基于 LDA 重要主题的多文档自动摘要算法[J]. 计算机科学与探索, 2015, 9(2): 242-248.

[13] LUO Senlin, BAI Jianmin, PAN Limin, et al. Research on multi-document summarization merging the sentential semantic features[J]. Transactions of Beijing Institute of Technology, 2016, 36(10): 1059-1064. (in Chinese)
罗森林, 白建敏, 潘丽敏, 等. 融合句义特征的多文档自动摘要算法研究[J]. 北京理工大学学报, 2016, 36(10): 1059-1064.

[14] WANG Lijin, YIN Yilong, ZHONG Yiwen. Cuckoo search algorithm with dimension by dimension improvement[J]. Journal of Software, 2013, 24(11): 2687-2698. (in Chinese)
王李进, 尹义龙, 钟一文. 逐维改进的布谷鸟搜索算法[J]. 软件学报, 2013, 24(11): 2687-2698.

[15] DASH P, SAIKIA L C, SINHA N. Comparison of performances of several Cuckoo search algorithm based 2DOF controllers in AGC of multi-area thermal system[J]. International Journal of Electrical Power and Energy Systems, 2014, 55: 429-436.

[16] RUDOLPH G. Convergence analysis of canonical genetic algorithm[J]. IEEE Transactions on Neural Networks, 1994, 5(1): 96-101.

(下转第 71 页)

(上接第 64 页)

[17] KANAGARAJ G,PONNAMBALAM S G,JAWAHAR N. Reliability-based total cost of ownership approach for supplier selection using Cuckoo-inspired hybrid algorithm [J]. The International Journal of Advanced Manufacturing Technology,2016:84:801-806.

[18] ZHANG Yongwei, WANG Lei, WU Qidi. Dynamic adaptation Cuckoo search algorithm [J]. Control and Decision,2014,29(4):617-622. (in Chinese)
张永韡,汪镭,吴启迪. 动态适应布谷鸟搜索算法[J]. 控制与决策,2014,29(4):617-622.

[19] LIN C Y,HOVY E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Washington D. C., USA: IEEE Press,2003:71-78.

[20] WANG Xuefei. Research of Chinese automatic summarization based on word2vec [D]. Harbin: Harbin Institute of Technology,2017. (in Chinese)
王雪霏. 基于 word2vec 的中文自动摘要方法研究[D]. 哈尔滨:哈尔滨工业大学,2017.

[21] SAINI J R. Estimation of comprehension ease of policy guides of matrimonial Websites using gunning fog, Coleman-Liau and automated readability indices [J]. Social Science Electronic Publishing, 2015, 10 (4): 19-33.

编辑　陆燕菲