# 基于综合的句子特征的文本自动摘要

程 园[1] 吾守尔·斯拉木[1] 买买提依明·哈斯木[1,2]
(新疆大学信息科学与工程学院新疆多语种信息技术实验室 乌鲁木齐 830046)[1]
(和田师范专科学校计算机科学系 和田 848000)[2]

**摘 要** 采用了一种综合的文本自动摘要方法来抽取出涵盖范围广、冗余信息少、最能反映文本中心思想的文本摘要。该方法充分考虑文本中的词频、标题、句子位置、线索词、提示性短语、句子相似度等特征因素,构建了一个综合的特征加权函数,运用数学回归模型对语料进行训练,去除冗余句子信息,提取关键句生成摘要。实验评估表明了该方法的可行性、有效性以及在摘要质量方面的优越性。

**关键词** 自动摘要,特征因素,综合,加权函数
**中图法分类号** TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.046

## Automatic Text Summarization Based on Comprehensive Characteristics of Sentence

CHENG Yuan[1] Wushouer SILAMU[1] Maimaitiyiming HASIMUA[1,2]
(Xinjiang Laboratory of Multi-language Information Technology, School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)[1]
(Department of Information Science, Hetian Teather College, Hetian 848000, China)[2]

**Abstract** To extract the abstract with less redundant information and a wide coverage, which can reflect the main idea of the text, this paper advanced a comprehensive text summarization method. This method takes the frequency of the words, the title, the position of the sentence in the text, cue phrases, similarity of the sentences and other features in the text into consideration, constructs a comprehensive feature weighting function, trains the corpus with mathematical regression model, removes the redundant information, and then gets the abstract. The experiment shows that this method is very effective and feasible, and very superior in the quality of the extraction.

**Keywords** Automatic abstract, Features, Comprehensive, Weighting function

随着 Internet 的迅猛发展和大数据时代的到来,信息的自动化处理问题亟待解决,传统人工抽取信息的方法已远远不能满足用户的需求。因此在信息爆炸的时代,信息自动化处理方向的文本自动摘要受到了人们越来越广泛的关注。

文本自动摘要就是对文本信息内容概括总结,提取出文本主要内容形成摘要,从而反映文本的中心思想。1958 年 Luhn 提出了一种基于高频词打分自动文本摘要方法[1],从此拉开了自动文摘的序幕;1969 年 Edmudson 提出了根据句子位置和线索词等特征提取文摘句的文本自动摘要方法[2];2004 年 Erkan 和 Radev 提出了基于 LexRank 算法[3]的文本自动摘要方法,该算法运用向量空间模型将句子权重通过图形表示,计算相邻节点的句子间的相似度,抽取出与相邻节点相似度最大的句子作为文摘候选句生成摘要;2009 年 Antiqueira、Oliveira 和 Costa 提出了运用图的概念将文本中的每个句子用图中节点表示,将句子间关系用图中的边表示的复杂网络的文本摘要方法[4]。相对国外,国内从 19 世纪 80 年代才开始这方面研究,目前国内自动文摘研究比较突出的有:上海交通大学的基于仿人算法的 OA 中文文献自动摘要系统[5];哈尔滨工业大学的基于语义分析、理解的 HIT-971 英文自动文摘系统[6]等。

文本自动摘要研究有两类主要方法:基于语义理解的自动摘要和基于词频统计的自动摘要,前者需要借助相关领域语料库及自然语言处理中的语义分析等理解原文,生成文摘,它的局限性比较大,技术不够成熟,仅限于一些狭小的应用领域;后者基于文本表层信息,考虑的文本特征信息不全面,语句冗余、不连贯等方面问题比较突出,生成的文摘质量不高。

大数据时代,网络文本信息内容丰富并且不受领域限制,用户对内容简洁精练、覆盖范围广、准确度高的摘要需求急剧上升。为此本文提出了一种基于统计模型、数学回归模型、空间向量模型相结合的综合的文本自动摘要方法;运用文本特征分析、关键词提取、特征加征加权函数、句子抽取和相似度计算等技术,充分地考虑文本的特征、冗余信息,结合实验及分析对文本自动摘要进行深入研究。

## 1 摘要提取算法流程

本文提出的综合的文本自动摘要方法主要是针对单文档的文本自动摘要。该方法考虑文本中的词频、标题、句子位置、线索词、提示性短语、句子相似度等特征因素,首先对文本

到稿日期:2014-04-03 返修日期:2014-08-06 本文受国家 973 重点基础研究发展计划资金项目(2014CB340506)资助。
程 园(1987-),男,硕士生,主要研究方向为信息检索;吾守尔·斯拉木(1942-),男,教授,中国工程院院士,主要研究方向为自然语言处理,E-mail:wushour@xju.edu.cn(通信作者);买买提依明·哈斯木(1980-),男,博士生,讲师,主要研究方向为数据挖掘。

进行预处理、关键词抽取；然后提取句子特征。利用特征加权函数对句子进行加权，提取候选文摘句；最后计算句子相似度，去除冗余信息，生成文本摘要。具体流程如图 1 所示。

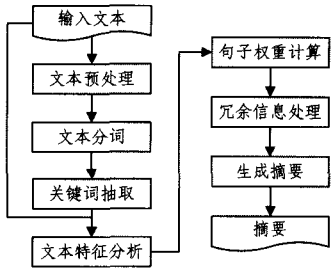


图 1　摘要提取算法程序

## 2　关键技术

### 2.1　文本的预处理

文本预处理对文本结构特征进行分析，识别出文本的标题，提取标题。中文文本中标题一般是对文章主题的高度概括。因此，准确地识别文本的标题能够在一定程度上提高文本自动摘要的准确性。对于标题，我们发现它往往具有以下特征：通常在文本的首部，在一行的中间，字数一般较少，标题中间没有句号、分号等标点符号，但可以出现顿号、引号、破折号、问号等，末尾一般没有标点符号，或者仅出现“!”、“?”、“…”。

文本的预处理还包括对文本进行分段分句处理，我们将“。”、“!”、“?”、“…”作为一个句子的结束符，将文本分成句子集合并将其保存。

### 2.2　分词及关键词提取

#### 2.2.1　分词及分词修正算法

文本自动摘要需要对文本进行分词处理。将文本中的词切分出来是一个关键问题，综合多种分词系统，本文选择了中科院计算机研究所的 ICTCLAS2013 汉语分词系统进行分词，但该系统由于所使的分词词典的限制，无法识别大量由两个或两个以上词组成的组合词[7]。所谓的组合词是客观上表示一个独立、特定语义的词，这些词具有重要的意义，在分词过程中将一些具有独立、特定语义的组合词错误地切分成多个字或词，例如“秦始皇帝陵兵马俑”，ICTCLAS2013 汉语分词系统分词结果为：“秦始皇/nr 帝/ng 陵/ng 兵马俑/n”。分词后组合词已经被切成词碎片，失去了原有的意义。本文开发了基于左右邻接词频统计的分词修正算法，对 ICTCLAS2013 分词结果进行修正，逐步把组合词和关键短语识别出来，还原被切碎的组合词，提高了分词精确度。

基于左右邻接词频统计的分词修正算法如下：

Step1　运用中科院 ICTCLAS2013 对文本进行分词。

Step2　将每个词的词性分别与它左右邻接词的词性进行匹配，如果词性匹配，则将它们合并，组合词的词性不变。

Step3　从第一个词开始，将每个词 A 分别与它左右邻接词 B 进行合并（词性限于 n、v、a、new_word），统计每个组合词的词频，当组合词 AB 或 BA 的词频选择率（词频选择率＝组合词词频/分词 A 与 B 词频之和）大于 0.3 时，则生成新的分词结果，组合词的词性为 new_word。

Step4　重复 Step3，直到没有新的组合词生成为止。

#### 2.2.2　关键词提取

关键词提取首先过滤掉文本中无意义的词语，如“的、地、得、和、了”，这就是停用词。通过 StopWordTable 过滤掉这些无意义的停用词。分词结果过滤完停用词后，本文运用词频统计算法统计词语的频率，运用 TF-IDF 算法计算词语的权重，然后根据词语权重大小提取出关键词。在关键词提取过程中，文章中的词可以被分为两类[8]：实词和虚词。实词就是在表达文章主题时起主要作用的词，这些实词大多数为名词（n）、动词（v）、形容词（a）、组合词（new_word）；虚词是对句子意义贡献极小的数量词、介词、连词、助词、感叹词、象声词等。因此，在提取特征词的过程中，本文只对名词（n）、动词（v）、形容词（a）和组合词（new_word）进行考虑。词语权重计算公式为：

$$W_{i,j}=tf_{i,j}\times\log\frac{N}{n_j+1} \tag{1}$$

其中，$W_{i,j}$ 表示特征词的权重，$tf_{i,j}$ 表示词语 $t_i$ 在文档 $d_j$ 中出现的频率，$N$ 表示文本集合中所有文本数目，$n_j$ 为文本集合中包含词语 $t_i$ 的文本数目。

根据上述算法计算出词语的权值后，挑选出权重最大的若干实词作为关键词（通常为 5～10 个）。为了充分考虑关键词特征对句子权重的影响，本文抽取 10 个实词作为关键词。

### 2.3　基于向量空间模型的文本表示

向量空间模型 VSM[9] 是 20 世纪 60 年代由 Salton 等人提出来的关于文档表示的一个统计模型，主要用于信息检索、分类、聚类、自动摘要等。其基本思想是把文本的各级对象单元映射成 $n$ 维向量空间中的对应向量，每个向量用义项集合来表示。

在向量空间模型中，每个文本都可以被看成是一组义项的组成集合，本文中我们用上述算法提取的 10 个关键词表示文本中的义项，文本可表示为 $D(T_1,T_2,\cdots,T_{10})$，义项 $T_k$ 被赋予一个权重 $W_k$，句子 $s_i$ 被表示为 $s_i(W(s_{i,1}),W(s_{i,2}),\cdots,W(s_{i,10}))$，$0\leqslant W(s_{i,j})\leqslant 1$，其中 $W(s_{i,j})$ 表示第 $j$ 个义项在第 $i$ 句中的权重。若 $W(s_{i,j})=0$，则表示第 $j$ 义项在第 $i$ 句中不存在。

文本 $D$ 由 $m$ 个句子组成，可用向量表示为：$D(s_1,s_2,\cdots,s_m)^{\mathrm{T}}$。

用矩阵表示如下：

$$\begin{array}{c} D \\ s_1 \\ \vdots \\ s_i \\ \vdots \\ s_m \end{array}\begin{array}{c} \begin{array}{cccccc} T_1 & \cdots & T_5 & \cdots & T_{10} \end{array} \\ \begin{bmatrix} W(s_{1,1}) & \cdots & W(s_{1,5}) & \cdots & W(s_{1,10}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ W(s_{i,1}) & \cdots & W(s_{i,5}) & \cdots & W(s_{i,10}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ W(s_{m,1}) & \cdots & W(s_{m,5}) & \cdots & W(s_{m,10}) \end{bmatrix} \end{array}$$

### 2.4　基于位置的句子权重

美国的 P. E. Baxendale[10] 抽样统计表明，反映主题摘要的“中心思想句”，有 85％出现在段首，7％出现在段尾。我们随机地选取了 100 篇中文文本的人工摘要进行抽样统计，发现中文摘要句出现在首段和其它段落首句、尾句的概率比较大，所以自动摘要的文摘句优先从文本的首段、段首句、段尾句中去寻找。句子 $s_i$ 基于位置权重加权规则为：

$$W_L(s_i)=\begin{cases}1, & s_i \text{ 为第一段的句子}\\ 0.8, & s_i \text{ 为非第一段的第一句}\\ 0.5, & s_i \text{ 为非第一段的最后一句}\\ 0, & s_i \text{ 为其它句子}\end{cases} \tag{2}$$

### 2.5 基于线索词的句子权重

句子中有些词或短语本身不是关键词，但它们能起到提示作用，告诉读者此句中含有重要信息。这些词或短语就是线索词[11]，如“综上所述”、“总之”、“因此”等。早期的自动文摘系统中线索词和标志短语常用来标识文章中的重要句子。线索词在目前的文摘中仍受到高度重视，尤其能解释文章层次关系的线索词已被公认为是提取文摘句的首选，并取得了不错的效果。句子 $s_i$ 基于线索词的权重加权规则为：

$$W_C(s_i)=\begin{cases}1, & \text{句子中含有线索词}\\ 0, & \text{句子中不含线索词}\end{cases} \tag{3}$$

### 2.6 基于提示性短语的句子权重

文章中常常有一些特殊的短语、字串，它们对文章主题有明显的提示作用，可以用来获取文章的主题，这就是指示性短语，例如“这篇文章的意义是”、“本文的目的是”、“我们认为”等。本文提出根据提示性短语来选择文摘句的方法。句子 $s_i$ 基于指示性短语的权重的加权规则为：

$$W_I(s_i)=\begin{cases}1, & \text{句子中含有指示性短语}\\ 0, & \text{句子中不含指示性短语}\end{cases} \tag{4}$$

### 2.7 基于关键词的句子权重

关键词通常是反映文章主题的词语，一个句子所包含关键词越多，其信息量越大，句子越重要。句子 $s_i$ 基于关键词的权重计算公式为：

$$W_F(s_i)=\sum_{k=1}^{10}W(s_{i,k}) \tag{5}$$

$W_F(s_i)$ 表示句子 $s_i$ 基于关键词的权，$\sum_{k=1}^{10}W(s_{i,k})$ 表示句子 $s_i$ 包含关键词权重之和。

### 2.8 基于句子与标题相似度的句子权重

中文文本中标题往往反映着文本的主题，我们通过计算文本中句子与标题的相似度来评估句子在文本中的重要度。文本中一个句子和文本标题的相似度越大，则该句重要度越大，被抽取的可能性也越大。计算句子和标题相似度的方法中，比较经典的是采用向量空间模型，即将句子 $s_i$ 和标题都表示成一个 $n$ 维的向量，其相似度用 $\cos\theta$ 表示：

$$sim(x,y)=\cos\theta=\frac{(x,y)}{|x|\cdot|y|} \tag{6}$$

式中，$x,y$ 分别为表示文本标题和句子 $s_i$ 的向量，$(x,y)$ 表示向量 $x$ 和向量 $y$ 的数量积，$|x|\cdot|y|$ 表示向量 $x$ 的模与向量 $y$ 的模的乘积。本文中向量 $x$ 表示为 $T(W(T_1),W(T_2),\cdots,W(T_{10}))$，向量 $y$ 表示为 $s_i(W(s_{i,1}),W(s_{i,2}),\cdots,W(s_{i,10}))$。句子 $s_i$ 基于与标题相似度的权重计算公式：

$$W_T(s_i)=\frac{\sum_{k=1}^{10}W(T_k)\times W(s_{i,k})}{\sqrt{\sum_{k=1}^{10}W(T_k)^2}\times\sqrt{\sum_{k=1}^{10}W(s_{i,k})^2}} \tag{7}$$

$W_T(s_i)$ 表示句子 $s_i$ 基于与标题相似度的权重，$W(s_{i,k})$ 表示句子 $s_i$ 中第 $k$ 个关键词权重，$W(T_k)$ 表示标题 $T$ 中第 $k$ 个关键词权重。

## 3 数学回归模型评估句子特征权重

数学回归模型是一种好的评估文本特征权重的模型。数学回归模型是从一组样本数据出发，确定变量之间的数学关系式，对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出影响显著和不显著的变量。利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出预测或控制的精确程度。

数学回归模型可以利用数学关系式把输入和输出关联起来，得到一个输入输出的关系模型。本文把人工总结的文本特征作为独立的输入变量，句子权重作为输出变量，得到一个综合特征加权函数，通过加权函数对语料库进行训练，得到一组最优的比例因子。通过矩阵概念表示数学回归模型为：

$$\begin{pmatrix}W(s_1)\\ \vdots\\ W(s_i)\\ \vdots\\ W(s_m)\end{pmatrix}=\begin{pmatrix}W_F(s_1) & W_L(s_1) & W_C(s_1) & W_I(s_1) & W_T(s_1)\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ W_F(s_i) & W_L(s_i) & W_C(s_i) & W_I(s_i) & W_T(s_i)\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ W_F(s_m) & W_L(s_m) & W_C(s_m) & W_I(s_m) & W_T(s_m)\end{pmatrix}\cdot\begin{pmatrix}\alpha\\ \beta\\ \gamma\\ \kappa\\ \lambda\end{pmatrix} \tag{8}$$

$[W(s_i)]$ 是输出向量，表示句子 $s_i$ 的综合权重，$(W_F(s_i),W_L(s_i),W_C(s_i),W_I(s_i),W_T(s_i))$ 是输入向量，表示句子 $s_i$ 的特征向量；$\alpha,\beta,\gamma,\kappa,\lambda$ 是比例因子。综上，综合的句子权重计算函数为：

$$W(s_i)=\alpha W_F(s_i)+\beta W_L(s_i)+\gamma W_C(s_i)+\kappa W_I(s_i)+\lambda W_T(s_i) \tag{9}$$

$W(s_i)$ 表示句子 $s_i$ 基于综合特征的权重。

## 4 冗余信息去除与摘要生成

通过上述方法我们从文本中抽取了粗略的文摘候选句，但可能存在多个候选句描述文章的同一内容主题的情况，即这些文摘候选句含有大量冗余信息，因此选择文摘句时，不仅仅要考虑句子综合权重的高低，还要考虑选取的候选句是否含有大量冗余信息，即避免重复。本文通过计算句子间的相似度来去除冗余信息，以提高文摘的质量。消除句子冗余与摘要生成的算法如下：

Step1 设定一个相似度阈值 $a$（$a$ 的取值根据实验文本语料题材不同而不同，0 表示两句完全不相同，1 表示两句相同，经过实验表明，$a$ 为 0.6 时效果较好）。

Step2 从粗文摘候选句 $i=1$ 开始，依次对 $i$ 从 2 循环到 $m$ 计算 $sim(s_1,s_i)$，如果 $sim(s_1,s_i)>a$，则删除 $s_i$，提取 $s_1$ 加入文摘句，然后删除第一句。

Step3 对粗文摘候选句按权重由大到小进行排序，$m$ 为排序后的粗文摘候选句的句子总数。

Step4 重复 Step2、Step3，直到选取的文摘句数目满足用户要求的文摘长度为止。

## 5 实验评测

### 5.1 实验语料

目前中文自动摘要没有一个公认的评估语料和评估标准，本文的训练以及测试的语料来自人民网新闻。我们使用爬虫程序从人民网上爬取了题材为两会报告、新闻、体育、经济、军事的 5 个类别的 800 篇中文文本，其中 500 篇作为训练语料，300 篇作为测试语料。训练与测试语料库特征如表 1 所列。

表1 训练与测试语料库特征

| 文档 | 训练语料 | 测试语料 |
|---|---|---|
| 文档包含最少句子数 | 9 | 12 |
| 文档包含最多句子数 | 48 | 45 |
| 句子总数 | 12655 | 7643 |
| 大小(kb) | 1965.8 | 1153.7 |

3名语言专业的研究生对训练语料和测试语料进行人工摘要提取，每个摘要员各自独立地从文档中准确抽取出10个摘要句子，最后综合3人摘要结果，取结果相同的5个句子作为人工摘要句。

### 5.2 实验分析

为了合理评估摘要的质量，本文采用准确率 $P$、召回率 $R$ 和 $F$ 值作为衡量标准，计算每篇文本每个句子的 $W_F(s_i)$、$W_L(s_i)$、$W_C(s_i)$、$W_I(s_i)$、$W_T(s_i)$，统计句子关键词特征的权重衡量因子 $W_{AF}(s_i)$、位置特征的权重衡量因子 $W_{AL}(s_i)$、线索词特征的权重衡量因子 $W_{AC}(s_i)$、提示性短语特征的权重衡量因子 $W_{AI}(s_i)$、与标题相似度特征的衡量因子 $W_{AT}(s_i)$，其中

$$W_{AF}(s_i)=\frac{\sum_{i=1}^{m}W_F(s_i)}{m} \tag{10}$$

式中，$m$ 为文本句子总数，$W_{AL}(s_i)$、$W_{AC}(s_i)$、$W_{AI}(s_i)$、$W_{AT}(s_i)$公式与式(10)同理。通过实验表明，$W_F(s_i)$的值对综合权重$W(s_i)$值影响最大，$W_L(s_i)$、$W_T(s_i)$其次，$W_C(s_i)$、$W_I(s_i)$最小。综合特征句子加权函数中的比例因子 $\alpha,\beta,\gamma,\kappa,\lambda$(其中 $\alpha+\beta+\gamma+\kappa+\lambda=1$)，我们对 $\alpha,\beta,\gamma,\kappa,\lambda$ 取值进行训练，在大量的实验研究后发现，当 $\alpha,\beta,\gamma,\kappa,\lambda$ 以小于0.1级增加时，该权重增加值不明显，对综合权重 $W(s_i)$ 贡献很小，对实验结果影响可以忽略，所以比例因子每次取0.1级别或0.1的整数级别增加。训练实验结果研究表明，选取50组比例因子组合进行实验，利用综合加权函数对训练语料进行训练，用训练抽取出来的摘要句与人工摘要句进行匹配，计算不同比例因子组合下的 $P$、$R$、$F$ 值，实验结果如表2所列(选取了12组有代表性的比例因子组合)。

表2 不同比例因子组合下的 $P$、$R$、$F$ 值

| 序号 | α | β | γ | κ | λ | P(%) | R(%) | F(%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | 0 | 0 | 0 | 0 | 57.2 | 52.2 | 53.6 |
| 2 | 0 | 1.0 | 0 | 0 | 0 | 46.7 | 41.3 | 43.8 |
| 3 | 0 | 0 | 1.0 | 0 | 0 | 30.8 | 25.5 | 27.9 |
| 4 | 0 | 0 | 0 | 1.0 | 0 | 29.4 | 24.3 | 26.3 |
| 5 | 0 | 0 | 0 | 0 | 1.0 | 40.9 | 35.7 | 38.1 |
| 6 | 0.8 | 0.1 | 0 | 0 | 0.1 | 58.5 | 52.3 | 55.2 |
| 7 | 0.7 | 0.1 | 0.1 | 0 | 0.1 | 60.5 | 55.4 | 57.8 |
| 8 | 0.6 | 0.1 | 0.1 | 0.1 | 0.1 | 66.8 | 61.6 | 64.1 |
| 9 | 0.5 | 0.2 | 0.1 | 0.1 | 0.1 | 71.3 | 65.2 | 68.1 |
| 10 | 0.4 | 0.2 | 0.1 | 0.1 | 0.2 | 73.2 | 67.6 | 70.3 |
| 11 | 0.3 | 0.3 | 0.1 | 0.1 | 0.2 | 59.4 | 55 | 57.1 |
| 12 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 48.6 | 43.2 | 45.7 |

实验结果表明，当 $\alpha,\beta,\gamma,\kappa,\lambda$ 分别为0.4、0.2、0.1、0.1、0.2时，综合加权函数性能最佳。

### 5.3 对比实验

通过实验分析我们得到了一组最佳的比例因子组合，分别采用本文所提出方法、基于LexRank改进算法[12]的文本自动摘要方法和基于潜在语义分析(LSA)[13]的文本自动摘要方法对300篇测试语料文本进行实验测试，并对500篇训练语料文本进行了交叉实验测试。根据实际项目需求，抽取文本30%的句子作为摘要，将这3种方法抽取出来的摘要句子与人工摘要句子进行匹配，计算它们的准确率 $P$、召回率 $R$ 和 $F$ 值，评估3种自动摘要方法的性能。实验结果如表3所列。

表3 3种自动摘要方法的性能评估

| 方法 | P(%) | R(%) | F(%) |
|---|---|---|---|
| 本文方法 | 72.9 | 67.5 | 70.1 |
| LexRank改进算法 | 56.3 | 53.2 | 54.7 |
| LSA方法 | 67.5 | 63.4 | 65.4 |

表3表明本文所提出的自动摘要方法优于其它两种方法。实验分析发现，基于LexRank改进算法和基于潜在语义分析的文本自动摘要方法没有充分考虑文本的特征，如句子的位置特征和文本的标题特征等，不能反映句子的全局重要度，因而不能准确计算出句子的权重，因此文本的特征对实验效果有着重要的影响。本文方法充分考虑了文本句子的特征，运用综合的加权函数评估句子的重要度，利用句子相似度去除冗余信息。实验表明：本文提出的方法能够实现多领域的文本自动摘要，提取出来的摘要更接近人工摘要，满足用户需求，并且摘要的速度比较快，摘要长度可以根据用户需求进行调节，生成的摘要涵盖范围广、冗余信息小，准确率比较高。

**结束语** 本文自动摘要的方法是针对单文档的文本自动摘要。该方法考虑文本中的词频、标题、句子位置、线索词、提示性短语、句子相似度等特征因素，抽取关键词，构建了一个综合的句子加权公式，计算句子权重，去除冗余信息，生成文本摘要。实验证明，本方法生成的摘要具有不受领域限制、涵盖内容全面、冗余信息少、摘要准确性比较高等优点。在今后进一步工作中，将针对摘要的可读性进行重点研究，运用自然语言理解和语法语义分析技术，改善摘要句的连贯性，提高摘要质量。

## 参考文献


[1] Luhn H P. The automatic creation of literature abstract[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165

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

[3] Erkan G, Radev D R. LexRank: Graph-based lexical centrality as salience in text summarization[J]. J. Artif. Intell. Res. (JAIR), 2004, 22(1): 457-479

[4] Antiqueira L, Oliveira Jr O N, Costa L F, et al. A complex network approach to text summarization[J]. Information Sciences, 2009, 179(5): 584-599

[5] 王永成，许慧敏. OA中文文献自动摘要系统[J]. 情报学报，1997，16(2)：128-132

[6] 吴岩，李秀坤，王开铸. HIT-971型英文自动文摘系统[J]. 情报学报，1998，17(5)：358-364

[7] 蒋昌金，彭宏，王开铸. 基于主题词权重和句子特征的自动文摘术[J]. 华南理工大学学报，2010，38(7)：50-54

[8] 刘功中，李建华，李生红. 基于类信息的特征选择和加权方法[C]//第一届全国信息检索与内容安全学术会议. 2004

[9] Salton G, Lesk M E. Computer evaluation of indexing and text processing [J]. Journal of the ACM, 1968, 15(1): 8-36

[10] Machine B E. Made index for technical literature an experiment [J]. IBM Journal of Research and Development, 1958, 12(4): 354-361

[11] 张志昌，张宇，刘挺，等. 基于线索词识别和训练集扩展的中文问题分类[J]. 高技术通讯，2009，19(2)：111-118

[12] 纪文倩，李舟军，巢文涵，等. 一种基于LexRank算法的改进的自动文摘系统[J]. 计算机科学，2010，37(5)：151-154

[13] Ozsoy M G, Alpaslan F N, Cicekli I. Text summarization using latent semantic analysis [J]. Journal of Information Science, 2011, 37(4): 405-417