广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(4):226~228
Vol. 23,No. 4 November 2007

基于局部与全局信息的自动文摘算法*
Research of Automatic Summarization Based on Local and Global Information of Sentences

王 萌,王晓荣,李春贵,唐培和
WANG Meng,WANG Xiao-rong,LI Chun-gui,TANG Pei-he

(广西工学院计算机工程系,广西柳州 545006)
(Department of Computer Engineering,Guangxi University of Technology,Liuzhou,Guangxi,545006,China)

摘要:采用平均特征词频率策略计算特征词权重,用快速 n-grims 算法对各特征词所处的概念体进行加权,用一种改进的 K-means 聚类算法进行段落聚类,提出一种基于局部与全局信息的自动文摘算法并给出算法评估。该算法不仅能够自适应获得 k 值,而且有效防止了初始点的随机选择对聚类结果的影响。评测结果表明该算法对经济类和科技类文章的准确率和召回率都明显高于新闻类和文学类文章,利用机器文摘进行分类的准确率明显高于使用原文本进行分类。该算法所得到的文摘,在各项指标上都优于传统方法生成的文摘。

关键词:K-means n-grims 段落聚类 自然语言理解
中图法分类号:TP301.6 **文献标识码**:A **文章编号**:1002-7378(2007)04-0226-03

Abstract:The idea of our approach is to exploit both the local and global properties of sentences. In order to obtain local property,we use a term weighting scheme that employs average term frequency in a document as the normalization factor. And a fast algorithm for matching N-grams is uesd to optimize term weighting. The method can obtain an improved K-means method to cluster paragraphs,and discovers thematic areas according to clustering results. Furthermore,it integrates local and global property to produce summarization. And experiments do prove that it is feasible to use the method to develop a domain automatic abstracting system,which is valuable for further study.

Key words:K-means,n-grims,paragraph clustering,natural language understanding

作为自然语言处理的一个重要分支的计算机自动文摘已成为 Internet 信息时代的必然需求[1]。自动文摘是利用计算机自动地从原始文献中生成能准确全面地反映文献中心内容、而且语言简洁连贯的摘要[2]。如何获得准确有效的文摘已成为自然语言处理领域的一个重要研究课题。

本文提出一种基于局部与全局信息的自动文摘算法并采用内部评测和外部评测等多种手段来对文摘结果进行评测。该算法包含三个主要部分:局部信息提取,全局信息提取和主题句提取。算法首先采用平均特征词频率策略[3]来进行特征词权重计算,利用快速 n-grims 算法[4]对各特征词所处的概念体(通过 n-grims 获得的词串)进行加权以进一步优化特征词权重;其次利用一种改进的 K-means 聚类算法把若干个描述内容相似的段落进行聚类,形成一个主题区域,将所得到的主题区域作为一个全局信息,在每个主题区域内选择若干主题句;最后根据句子所处主题区域和包含的特征词权重,综合评定得到文本摘要。

1 算法设计

1.1 文本预处理

以段落为基本单位对文本进行划分。将文章的关键词、作者栏、论文出处、作者单位等相关信息删除。对于标题和子标题如果包含的词语个数大于 3,则保留并与文章一起进行计算,否则删除。

在计算词语的 n-gram 值前,删除文本中的停用

收稿日期:2007-09-10

作者简介:王萌(1979-),男,硕士,讲师,主要从事自然语言理解、智能处理研究。

* 国家自然基金项目(60673034),2006 年广西教育厅基金项目(149),广西工学院博士、硕士基金项目资助。

词、介词等无意义的词语,保留形容词、动词、副词等词语作为文本的特征词。对需要处理的特征词进行标记,标记过程为:每一个结点包含一个特征词,每个结点的值分别对应该特征词在文本中出现的频率以及所对应的位置信息(段落、段落中的位置),并根据特征词的频率大小排序,用链表的形式进行存储,并将该结构称为词语存储器。

1.2 特征词权重计算

特征词权重计算公式

$$W_i = \frac{1+\lg(tf(t))}{1+\lg(\text{average}(tf(t)))}/(1-c)\text{average}(L)+c\times L, \quad (1)$$

其中 $tf(t)$ 为特征词 t 在文档 D 中实际出现的频率,average $(tf(t))$ 是 t 在训练文档集合中平均出现的频率,L 是不同特征词的总数,c 是 0 到 1 之间的一个常量。

1.3 特征词权重修改

考虑 5 个连续特征词间的关系,并采用快速 n-grims 匹配算法[4],得到相应的概念体后,对公式(1)中的特征词权重做如下修改:

$$W_i' = w_i \times \frac{n_i}{n^2}, \quad (2)$$

其中 n_i 为词语 w_i 所在 n-grims 的词串长度,n 为文档 D 中最大 n-grims 词串长度。修改后的特征词权重不仅反应了词语本身的重要性。同时也反映出特征词所在概念体的重要性。

1.4 段落聚类

采用一种改进的 K-means 算法,对文章段落进行聚类。首先,对于聚类个数,也就是 k 的值,采用一种自适应的方法进行,即在一个合理的 k 值范围内进行搜索;其次,随机选择初始点,重复运行 K-means 算法,利用一个失真系数来测量聚类的紧密性,选择紧密性最小的情况作为最优解。

失真系数定义为一个类中的段落与该类中心点的欧拉距离的平方平均和。因此,对于某种聚类情况 $P=\{c_1,c_2,\cdots,c_m\}$ 的失真系数为

$$D(P)=\sum_{i=1}^{m}E(c_i),E(c_i)=\frac{1}{|c_i|}\sum_{j}(s_j-\mu_{(i)})^2,$$

c_i 表示一个聚类,s_j 为 c_i 中的一个段落,$\mu_{(i)}$ 为 c_i 的中心点,$|c_i|$ 为 c_i 包含段落的个数。

假设 N 维向量空间中的每个点分别表示一个段落,待聚类的段落可以表示为 N 维向量空间中的 M 个点,其中 N 表示特征词的个数,M 表示待聚类的段落的个数。

基于改进 K-means 算法的段落聚类算法:

输入:

M,段落的个数;

N,维数,特征词的个数;

D,失真系数,初始值为 F(F 为一个绝对大数);

K_MIN,最小的聚类个数;

K_MAX,最大的聚类个数;

L,聚类的次数。

输出:段落聚类的结果 P,k 聚类数。

做 L 次:

1)随机选择一个 k 值,$k\in[K\_MIN, K\_MAX]$;

2)随机从 M 中选择 k 个初始点;

3)使用(1)、(2)中的参数进行 K-means 聚类,得到一个聚类结果 P;

4)利用失真系数 D(P)测量该聚类结果 P,如果 D(P)小于 F,修改 F;否则,保持原有值;

5)重复 1)~4)。

1.5 主题句提取

根据 1.4 得到的段落聚类结果,假设聚类结果为 $P=\{c_1,c_2,\cdots,c_m\}$,对每个聚类 c_i 计算其相似度,即计算该类中每个句子 S_j 与中心点 $\mu_{(i)}$ 的相似度,其计算公式为

$$Cos(S_j,\mu_{(i)})=\sum_{k=1}^{n}S_{jk}\mu_{(i)k}/\sqrt{|\sum_{k=1}^{n}S_{jk}^2||\sum_{k=1}^{n}\mu_{(i)k}^2|}, \quad (3)$$

并且,聚类 c_i 中心点的相似度设为 1。

利用 VSM 模型对句子进行向量表示,即句子 S_j 表示为 $S_j(W_1,F_{1j};W_2,F_{2j};\cdots;W_M,F_{Mj})$ 其中 N 为特征词的总数,F_{Nj} 表示句子 j 中的第 N 个特征词,W_N 为特征词 F_{Nj} 的权重,其值由公式(1)得出,句子权重计算公式为

$$W(S_j)=\frac{\sum_{i=1}^{N}W_i}{N}*Cos(s_j,\mu_{(i)})。 \quad (4)$$

在得到句子权重后,按照 1.4 所得到的主题区域,将句子在所处主题区域内按照句子权重进行排序,根据需要产生文摘长度,在每类主题区域中选择出前若干权重大的句子为文章主题句。这样,在进行句子抽取的过程中,按照聚类优先抽取的原则,即考虑到句子权重大小和抽取句子的多样性。

2 算法评估

采用内部评测和外部评测两种方法对算法得到

的文本摘要进行评估。内部评测是测试文摘本身是否与文章要点一致以及是否包含文章的所有基本要点;外部评测是通过文摘与文章的相似度计算或者文摘在信息检索中所起作用的大小来评估文摘。

2.1 内部评测

从新浪网上分别下载关于经济、文学、新闻、科技等4个方面的文章各10篇。首先用本文所设计的算法对40篇文章做出文摘，然后请两位中文系的专家为40篇文章分别做出相应的人工文摘。将机器自动生成的40篇文摘与两位教师做出的80篇文摘分为经济、文学、新闻、科技4类，另外请一位语言教授对机器文摘和所对应的2位专家的人工文摘进行综合评价，按召回率和准确率进行比较，其测试结果如图1所示。由图1可以看出，本文所采用的自动文摘算法对经济类和科技类论文，无论是准确率还是召回率都明显高于新闻类和文学类文章。主要是由于经济类和科技类论文的用词和论文结构比较确定严谨，相反文学类则比较随意。因此，本算法对论文结构和用词相对固定的经济类和科技类论文更为适用。

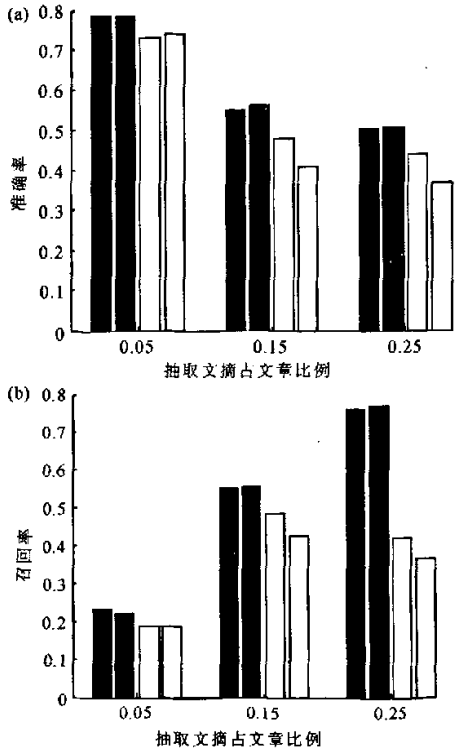

图1 综合评价比较

(a)文摘准确率;(b)文摘召回率

■:经济类;■:科技类;□:新闻类;□:文学类。

2.2 外部评测

运用一个事先已经做好的分类器，选择650篇文章作为训练语料，然后分别运用原文本和机器产生的文摘句作为一个测试语料进行分类，对60篇文章进行分类后的平均结果是：采用原文本的分类准确率为61.3%，利用机器文摘进行的分类准确率为66.8%，利用机器文摘分类明显高于使用原文本进行分类的准确率。

3 结束语

本文利用快速n-grims算法和改进的K-means算法，提出一种基于局部与全局的自动文摘算法，并采用内部评测和外部评测两种方法对算法得到的文摘进行评估。该算法不仅能够自适应获得k值，而且有效防止了随机选择对聚类结果的影响。内部评测表明本文采用的算法对经济类及科技类论文在准确率和召回率方面都明显高于新闻类和文学类文章，而外部评测表明利用机器文摘分类明显好于使用原文本进行分类。综合分析表明该算法得到的文摘，在各项指标上都优于传统的方法生成的文摘。

参考文献：

[1] CALIFF M E, MOONEY R J. Relational learning of pattern match rules for information extraction: Agrawal R Proceedings of the 19th National Conference on Artificial Intelligence[C]. New York: Holy Publishing Company, 2003: 87-90.

[2] 李蕾，钟义信. 全信息理论在自动文摘系统中的应用[J]. 计算机学报，2000，23(1)：4-7.

[3] SINGHAL A, BUCKLEY C, MITRA M. Pivoted document length normalization: proceedings of the 19th annual international ACM-SIGIR conference on research and development in information retrieval SIGIR'96, ACM New York[C]. New York: [s. n.], 1996: 21-29.

[4] HILDA HARDY, NOBUYUKI SHIMIZU, TOMEK STRZALKOWSKI, et al. Cross-document summarization by concept classification: proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR'02, ACM[C]. New York: [s. n.], 2002: 121-128.

(责任编辑：尹 闯 邓大玉)