一种基于主题词集的自动文摘方法 *

刘兴林[1,2], 郑启伦[1], 马千里[1]
(1. 华南理工大学 计算机科学与工程学院, 广州 510640; 2. 五邑大学 计算机学院, 广东 江门 529020)

摘 要: 提出一种基于主题词集的文本自动文摘方法,用于自动提取文档文摘。该方法根据提取到的主题词集,由主题词权重进行加权计算各主题词所在的句子权重,从而得出主题词集对应的每个句子的总权重,再根据自动文摘比例选取句子权重较大的几个句子,最后按原文顺序输出文摘。实验在哈工大信息检索研究室单文档自动文摘语料库上进行,使用内部评测自动评估方法对获得的文摘进行评价,总体 F 值达到了 66.07%。实验结果表明,该方法所获得的文摘质量高,较接近于参考文摘,取得了良好的效果。

关键词: 自动文摘; 主题词集; 句子权重; 自然语言处理
中图分类号: TP301;TP391 **文献标志码**: A **文章编号**: 1001-3695(2011)04-1322-03
doi:10.3969/j.issn.1001-3695.2011.04.035

Automatic summarization method based on thematic term set

LIU Xing-lin[1,2], ZHENG Qi-lun[1], MA Qian-li[1]
(1. *School of Computer Science & Engineering, South China University of Techhology, Guangzhou* 510640, *China*; 2. *School of Computer Science, Wuyi University, Jiangmen Guangdong* 529020, *China*)

Abstract: This paper proposed an automatic summarization method based on thematic term set for automatic extracting abstracts from Chinese documents. According to the extracted thematic term set, the method calculated the sentence weights by the weights of the thematic terms, then got the corresponding total weight of each sentence, and selected several sentences with higher weight by percentage, and finally, output the summarization sentences by original order. Experiments were conducted on HIT IR-lab text summarization corpus, and utilized intrinsic automatic evaluation measures to evaluate the performance of the proposed method. Experimental results show that the proposed method achieves 66.07% upon the *F*-measure, which suggests it can generate higher quality summarization, nearly to the reference abstract, achieving very good performance.
Key words: automatic summarization; thematic term set; sentence weight; NLP

0 引言

自动文摘(automatic summarization)的概念最早由 Luhn[1] 提出,是指利用计算机自动地从文章中提取最能表达该文章主要内容的关键句子组成短文的过程。

自动文摘方法可分为两类。一是基于统计的机械文摘[2],通过统计的方法直接从原文中抽取相关句子组成文摘,这类方法不需要进行语法、语义分析,所得文摘通顺,而且易于实现。二是基于语义的理解文摘[3],这类方法需要对文章进行语法和语义分析,在理解的基础上产生文摘句,所得文摘质量高,但实用性较低,实现起来较困难。

本文提出的一种基于主题词集的自动文摘方法,其理论依据是主题词集能贴切地表达文章中心思想。显然,主题词所在的句子亦能很好地归纳文章的主要内容,因此通过主题词权重加权计算句子权重,并按文摘比例选取主题词所在的句子权重最大的几个句子作为文摘是一种可行的方法。

1 相关工作

自动文摘技术结合了自然语言理解和自然语言生成技术[4]。互联网的高速发展,使其成为了全球最大的资源库,网络文本信息的海量式增长,为自动文摘技术的迅速发展和广泛应用提供了重要契机,大量摘要生成技术不断涌现,许多文摘系统投入了实际应用领域。

陶余会等人[5]提出了一种基于文本单元关联网络进行自动文摘的方法,该方法根据文本单元之间的共现关系建立关联网络,计算文本单元的共现信息量,并认为共现信息量越大的文本单元在文本中越重要,通过文本单元的权重计算出句子的权重,并提取权重值较大的一定比例的句子组成文摘。王志琪等人[6]采用基于互增强关系(MRP)的迭代算法模拟句子和词之间的循环加权关系,该方法先对句子中的词进行加权,句子的加权和词的加权互为因果,两者构成一个循环,通过迭代算法得出句子权重,进而提取文摘句子。Ai 等人[7]提出了一种基于潜在语义索引的自动文摘方法,使用向量空间模型来表示文本,通过语义索引来计算句子的相似度,得到句子权重,进而提取文摘句子。Wei 等人[8]提出了一种基于文档敏感图模型的多文档自动文摘方法,在计算句子权重时,考虑大文档集之间句子的相关度,从而实现多文档自动文摘。

收稿日期: 2010-10-11; **修回日期**: 2010-11-28 **基金项目**: 广东省自然科学基金资助项目(9451064101003233);华南理工大学中央高校基本科研业务费专项资金资助项目(2009ZM0125,2009ZM0189,2009ZM0255)

作者简介:刘兴林(1976-),男,博士研究生,CCF 会员,主要研究方向为数据挖掘、文本知识获取(jmxlliu@163.com);郑启伦(1938-),男,教授,博导,博士,主要研究方向为智能信息处理及其应用;马千里(1980-),男,讲师,博士,主要研究方向为智能计算、混沌时间序列、数据挖掘等.

文献[5,6]通过词权重来计算句子权重,文献[7~9]直接计算句子权重,这些研究成果表明,根据句子权重高低来选取文摘句是一种行之有效的方法,也取得了较好效果。本文提出的基于主题词集的自动文摘,原理与上述研究工作类似,但计算复杂度方面相对要低些,且当将主题词集的提取扩展到多文档时,该方法也适用于多文档自动文摘。

2　基于主题词集的自动文摘

基于主题词集的自动文摘主要分为主题词集提取和文摘生成两个部分。

2.1　主题词集提取

主题词集的提取采用基于词位置权重算法获得。认为同一个词在文章的不同位置出现,对该词是否成为主题词的影响是不一样的,因此根据词出现的位置不同而赋予该词不同的权重。将词在文章的出现位置分为三类:段序、句序、词序,下面给出相关的定义及其取值。

定义 1　段序(po)表示词出现在文章的不同段落,段序 = {首段,末段,其他}。

定义 2　句序(so)表示词出现在段落中的不同句子,句序 = {首句,末句,其他}。

定义 3　词序(wo)表示词在句子里出现的顺序,词序 = {首词,末词,其他}。

依上述分析,一个词可能出现的位置共有|段序| × |句序| × |词序| =27 种。为便于量化计算,结合中文行文习惯及同一篇文章中段落、句子、词序对表达文章中心思想的影响强弱,设置段序、句序和词序的各选项位置值如表 1 所示。

表 1　段序、句序、词序值

位置	首段	首句	首词	其他	末段	末句	末词
值	2^6	2^5	2^4	2^3	2^2	2^1	2^0

词 t 的位置值 pv_t 的计算公式如下:

$$pv_t = po + so + wo \tag{1}$$

词 t 单次出现的权重 w_{ti} 计算公式如下:

$$w_{ti} = \frac{pv_{ti}}{\sum_{i=1}^{27} pv_i} \tag{2}$$

式(2)中,$\sum_{i=1}^{27} pv_i$ 是指 27 种位置值的总和。

词 t 的总权重 w_t 计算公式如下:

$$w_t = \sum_{i=1}^{|t|} t_{wi} \tag{3}$$

式(3)中,$|t|$ 是指词 t 的出现次数。

由此可得到一个按总权重从高至低排序的候选主题词集。考虑到侯选主题词集中可能出现同义词的现象,则使用哈工大信息检索研究室同义词词林扩展版进行同义词合并,然后根据需要输出前 N 个候选主题词构成主题词集 TS,为满足后续自动文摘的需求,在输出主题词集时,同时输出每个主题词的总权重及所在句子编号。

主题词集 TS 的数据结构定义如下:

```
type TS
    cTerm    as    string
    fWeight    as    single
    cSentenceID    as    string
end type
```

其中,cSentenceID 是由分号间隔的句子编号字符串。

2.2　文摘生成

文摘生成的基本思想是按一定文摘比例(一般为 20%)选取句子权重最大的几个句子,再按原文顺序输出。从 2.1 节可知,主题词权重的大小反映了该词所在位置的重要性,因此由主题词权重进行加权计算而得到的句子权重也必然反映了该句子在整篇文章的重要性;而一篇文章的主题词能贴切地表达文章的中心思想,显然,相应地选取到的文摘句子也能很好地归纳文章主要内容。

2.2.1　句子权重的计算

计算句子权重时遵循以下规则:

规则 1　句子 s_j 包含 m 个不同的主题词,则将 m 个不同主题词的权重和乘以 $\sqrt{m}$ 作为句子的权重。

规则 1 反映了这样一个事实:包含多个权重较低的主题词的句子要比仅含一个高权重主题词的句子成为文摘句的可能性要大。

如上所述,设 $w_{t_i}(1 \leqslant i \leqslant N)$ 表示第 i 个主题词的权重,$w_{s_j}(1 \leqslant j \leqslant |S|)$ 表示第 j 个句子的权重,其中 $|S|$ 表示包含主题词的不重复句子总数。句子权重计算公式如下:

$$w_{s_j} = \sum w_{t_i} \times \sqrt{|w_{t_i}|} \tag{4}$$

其中:$\sum w_{t_i}$ 表示出现在句子 s_j 中的所有不同主题词的权重和,$|w_{t_i}|$ 表示出现在句子 s_j 中不同主题词的个数。

2.2.2　文摘句的选取

文摘长度的确定有两种方法:a)固定长度,文献[6,8]等采用这种方法;b)按比例选取,文献[5,7]等采用这种方法。本文采用按比例选取,在选取文摘句时,考虑到最后一个句子加入文摘后使得文摘长度出现小于或大于按比例计算出的文摘长度的可能,允许文摘长度有 ±10% 的误差。

设 L_{sys} 表示算法提取到文摘的长度,L_{ref} 为参考文摘长度,文摘句的选取步骤如下:

a)将主题词集所在的句子按权重从大到小排序;

b)若 $L_{sys} < L_{ref} \times 90\%$ 重复执行步骤 c),否则转步骤 d);

c)将第 $j(1 \leqslant j \leqslant |S|)$ 个句子加入文摘,$j++$;

d)若 $L_{sys} > L_{ref} \times 110\%$,将超出部分删除;

e)按原文顺序输出文摘。

3　实验与评价

实验在哈工大信息检索研究室单文档自动文摘语料库(HIT IR-lab text summarization corpus)上进行,包括“奥运”57 篇、记叙文 40 篇、说明文 40 篇、议论文 46 篇、应用文 18 篇、03 年“863”评测语料 10 篇,共计 211 篇。语料由五人分别进行人工标注,按照原文 20% 标注文摘句,形成五组参考文摘。

自动文摘评测方法广义上分为两类:内部评测和外部评测[10]。内部评测直接对摘要的质量进行评估,分为两种方法:手动评估和自动评估。手动评估由专家对摘要质量进行主观性打分;自动评估由算法进行,将算法文摘(由算法获得的文摘,以下同)与参考文摘进行比较,使用召回率(recall)、精度(precision)和 F 度量来反映自动文摘的优劣。

本文方法按原文 20% 选取文摘句,采用内部评测,对算法

文摘和参考文摘进行自动评估。召回率、精度和 F 度量的计算公式如下：

$$\text{recall} = \frac{\text{len(算法文摘中正确句子)}}{\text{len(参考文摘)}} \tag{5}$$

$$\text{precision} = \frac{\text{len(算法文摘中正确句子)}}{\text{len(算法文摘)}} \tag{6}$$

$$F\text{-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{7}$$

3.1 主题词集规模的确定

为考察主题词集规模对文摘质量的影响，选取其中一组（命名为 group-1）参考文摘，取 $N=6,7,8,9,10$ 进行文摘提取，实验结果如表 2 所示。

表 2 不同主题词集规模算法文摘质量评估结果

N		6	7	8	9	10
recall	max	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
	min	0.406 1	0.393 4	0.413 6	0.413 6	0.340 5
	avg	0.716 6	0.705 1	0.682 2	0.712 2	0.695 4
precision	max	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
	min	0.282 1	0.282 1	0.282 1	0.282 1	0.282 1
	avg	0.689 8	0.700 2	0.712 9	0.720 4	0.704 6
F-measure	max	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
	min	0.397 9	0.397 8	0.397 8	0.397 8	0.397 8
	avg	0.689 2	0.684 5	0.682 0	0.700 2	0.685 6

从表 2 实验结果可知，主题词集规模对算法文摘质量仅有细微影响，各项指标都非常接近，最大值都达到了 100%，当 $N=9$ 取得了最佳结果。召回率、精度和 F 度量分别为 71.22%、72.04% 和 70.02%。

3.2 实验结果评价

取 $N=9$，将算法文摘与其他四组参考文摘进行对比评价，得到表 3 所示的结果。

表 3 $N=9$ 时与其他四组参考文摘对比评价结果

corpus		group-2	group-3	group-4	group-5
recall	max	0.984 8	1.000 0	1.000 0	1.000 0
	min	0.448 5	0.447 7	0.424 0	0.444 4
	avg	0.627 3	0.649 0	0.691 2	0.673 0
precision	max	1.000 0	1.000 0	1.000 0	1.000 0
	min	0.444 4	0.225 3	0.148 1	0.282 1
	avg	0.717 5	0.687 3	0.669 4	0.636 5
F-measure	max	0.992 4	1.000 0	1.000 0	1.000 0
	min	0.531 4	0.331 4	0.243 9	0.407 4
	avg	0.658 6	0.651 3	0.654 3	0.639 3

表 3 实验结果表明，尽管人工标注主观性较强，所得参考文摘不尽相同，但是都表达了文章的主题思想，因此与各组参考文摘评测结果大致相同。随机选取语料库 10 篇文章，将各篇文章五组参考文摘进行比较，相似度最高的达到了 100%，平均相似度达到了 78.91%。这个发现证实了表 3 结果的可信度和有效性。

综合五组参考文摘对比评估结果，召回率、精度和 F 度量算术平均值如表 4 所示。

表 4 五组实验召回率、精度和 F 度量算术平均值

recall	precision	F-measure
0.670 5	0.686 2	0.660 7

从表 4 可见本文方法获得的文摘质量较高，较贴近于人工标注文摘，且文摘句直接来源于原文，不会出现语法问题，内容连惯性好、可读性强。

3.3 与其他算法比较

文献[6]与本文都采用了哈工大信息检索研究室单文档自动文摘语料库中 2003 年国家“863”自动文摘系统评价测试集，该测试样本集共有 10 篇文档，文档的平均字数为 3 017 个，平均句子数为 56.2 句。由于其他算法实现起来具有一定的困难，其他算法的实验结果未能通过实验直接给出，因此其他算法的实验结果引用文献[6]中的比较数据，比较结果如表 5 所示。

表 5 与其他算法比较结果

算法	recall	precision	F-measure
MS Word Summarizer	0.28	0.33	0.30
传统自动文摘	0.47	0.48	0.47
MRP[6]	0.59	0.60	0.59
本文算法	0.626 9	0.569 1	0.596 6

在这几种自动文摘方法中，本文方法取得了最佳成绩，尽管仅比 MRP 略好，但也反映了本文方法获取文摘是否更贴近于人工标注文摘方面要优于 MRP。

4 结束语

本文提出的基于主题词集的自动文摘方法，从实验结果来看，获得了较高的文摘质量。本文方法中文摘质量的高低依赖于主题词提取的准确性，准确性越高文摘质量也就越高。由于在进行自动文摘前需要先提取文章主题词，本方法运行效率方面有所欠缺。

后续的主要工作有：a）改进主题词提取算法，提高主题词提取准确性；b）采用内容相似度来评测自动文摘的质量；c）优化算法，提高运行效率。

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

[3] 杨晓兰，钟义信. 基于文本理解的自动文摘系统研究与实现[J]. 电子学报, 1998, 26(7): 155-158.

[4] RADEV D R, TEUFEL S, SAGGION H, *et al*. Evaluation challenges in large-scale document summarization[C]//Proc of the 41st Annual Meeting of the Association for Computational Linguistics. 2003: 375-382.

[5] 陶余会，周水庚，关佶红. 一种基于文本单元关联网络的自动文摘方法[J]. 模式识别与人工智能, 2009, 22(3): 440-444.

[6] 王志琪，王永成，刘传汉. 互增强关系的自动文摘句子加权方法[J]. 上海交通大学学报, 2007, 41(8): 1297-1300.

[7] AI Dong-mei, ZHENG Yu-chao, ZHANG De-zheng. Automatic text summarization based on latent semantic indexing[J]. Artificial Life and Robotics, 2010, 15(1): 25-29.

[8] WEI Fu-ru, LI Wen-jie, LU Qin, *et al*. A document-sensitive graph model for multi-document summarization[J]. Knowledge and Information Systems, 2010, 22(2): 245-259.

[9] 刘茵，李弼程. 自动文摘系统评测方法的回顾与展望[J]. 情报学报, 2008, 27(2): 235-243.