自动文摘的方法研究

卫佳君，宋继华
（北京师范大学 信息科学与技术学院，北京 100875）

摘 要：文中总结了自动文摘的主要研究方法和策略并把方法分成了三大类：自动摘录、基于信息抽取的自动文摘和基于理解的自动文摘。自动摘录方法是从文章中抽取重要句子来形成文摘；基于信息抽取的文摘方法是用从文章中抽取的信息填充已经编好的框架，然后用模板将内容输出；基于理解的文摘方法是利用自然语言处理技术生成文摘。文中重点总结了单主题文章和多主题文章的自动摘录方法，在多种算法进行优缺点比较后提出了一种新的多主题划分方法。

关键词：句子权值；相似度；关联网络；词频；聚类；主题划分
中图分类号：TP391 **文献标识码**：A **文章编号**：1673-629X(2011)08-0188-04

Research of Automatic Summarization Methods

WEI Jia-jun, SONG Ji-hua
(Dept. of Information Science and Technology, Beijing Normal University, Beijing 100875, China)

Abstract: It summarizes the main automatic abstracting research methods and strategies and divides the methods into three major categories: automatically extracted summarization, automatic summarization based on information extraction and summarization based on understanding. Automatically extracted method uses that extract important sentences from the article to form a digest; Abstract based on information extraction method uses that extract information from the article to fill framework which has been prepared, and then use the template to output the content; Abstract based on understanding is to use natural language processing technology to generate abstracts. focuses on automatically extracted summarization from single theme articles and multi-topic articles. After comparing advantages and disadvantages of variety of algorithms, a new multi-topic classification method is proposed.

Key words: sentence weights; similarity; association networks; word frequency; cluster; topic segmentation

0 引 言

根据国际标准 ISO214-1979(E)的规定，文摘是“一份文献内容的缩短的精确表达而无须补充解释或评论”。中国国家标准规定，文摘是“对文献内容作实质性描述的文献条目”。即文摘是简明、确切地记述原文献重要内容的语义连贯的短文。文摘是准确全面地反映某一文献中心内容的简洁连贯的短文。自动文摘的概念是由 Luhn[1] 首先提出，是利用计算机自动从原始文献中提取文摘。自动文摘方法概括为：自动摘录、基于理解的自动文摘、基于信息抽取的自动文摘。

1 自动文摘的方法

1.1 自动摘录

1.1.1 自动摘录的阐述

自动摘录方法是从文章中抽取重要的句子形成文摘，因此，确定哪些句子是重要句子是急需解决的问题。一种方法是计算每个句子的权重，权值最高的若干句子确定为重要句子。另一种方法是计算句子的相似度，如果一个句子与其它句子的相似度最高，将被确定为重要句子。计算句子的权值主要从6个方面考虑：词频、标题、位置、句法结构、线索词、指示性短语。

1.1.2 自动摘录的优缺点

自动摘录的优点能够适用于所有领域，不局限于某一领域，是目前用的较多的方法，实现起来相对来说比较简单，缺点是生成的文摘质量很不稳定：

(1)不全面，对于多主题的文章，生成的文摘有时仅包含了原文重点谈论的某个主题，从而影响了文摘的全面性，为了解决这个问题提出了多主题抽取的方法。

(2)不简洁，作者常常为了强调某些内容在文章的不同位置重复阐述，而这些一般都是关键句，很容易同时进入文摘，从而造成文摘内容的冗余。为了解决这个问题一般会在文摘抽取后进行冗余处理。

(3)不连贯，抽取文章中的若干句子组成的摘要往往缺乏连贯性，因为文章是一个有机的整体，每一个

收稿日期：2010-12-13；**修回日期**：2011-03-17
基金项目：国家社科基金项目(05BYY022)
作者简介：卫佳君(1987-)，女，安徽人，硕士研究生，研究方向为中文信息处理；宋继华，博导，教授，研究方向为中文信息处理。

句子通过省略、指代、同义词、相同词以及内在的逻辑关系与其上下文融为一体。当把文章中不同位置上出现的若干关键句连接成一个段落时,这些关键句由于脱离了上下文而难以准确地理解。甚至有时用户可能得出与原文不符的观点。

(4)自动摘录大部分只考虑到了词频和词义这个层次,对文章的理解度几乎为零,所以生成的文章性能不佳。

1.1.3 单主题自动摘录的算法

只有一个主题的文章叫单主题文章,即所有内容是围绕一个主题来阐述和讨论的,只要把与这个主题密切联系的句子抽取出来就可以形成文摘,但是主题不是已知的,这就无法根据主题抽取句子。简单起见,通常是抽取文章中重要的句子,因为重要的句子一般包含体现文章主旨的重要信息,因此,确定哪些句子是重要的句子是急需解决的问题。文中将把自动摘录算法分为五类,计算句子的权重、自动聚类、句子相似度、词义和文章结构,前三个算法只涉及到词频、位置等等表层的信息,没有涉及到词义的层次,词义这个类别最大的特点就是利用了词语的意思这个层次,一般后台会有一个强大的概念库或者知识库,文章结构这个类别最大的特点是利用了文章的结构层次,使得摘录性能较好。

(1)基于句子的权重的算法。

算法是先通过词频来计算每个句子的权重,然后按照权值对句子降序排列,依次选取权重大的句子并按原文顺序输出形成文摘。

a. 基于文本单元关联网络的自动文摘方法[2]提出了单词和句子权重计算方案,分别构建了基于单词和基于句子的关联网络。基于单词的关联网络是首先由词频计算词权值,然后通过计算句中所有单词的权值之和的平均值来计算句子权值。基于句子的关联网络是以句子为单元来计算句子的权重。

b. 如果一个词出现在很多权重较大的句子中,那么这个词的权重应该较大;如果一个句子中包含很多权重较大的词,那么这个句子的权重应该较大。基于此,基于互增强关系的自动文摘句子加权方法[3]提出了互增强关系的迭代算法来计算句子权重,实验证明比传统计算权值方法具有更好的性能。

上述算法在计算句子权值的时候仅仅考虑到了词频,这样得出的句子权值不能准确反应句子的重要性。

(2)基于自动聚类思想的算法。

一种使用自动聚类思想的自动文摘方法[4]采用了自动聚类的思想,首先进行词频统计确定文章的关键词,将段落分成几部分然后计算每个部分与关键词的相关度,选出候选部分,最后计算候选部分中每个句子与候选部分的相关度,选出候选句子,形成文摘。该算法最大的特点是提出了自动聚类的思想,也提出了新的一种抽取重要句子的算法。缺点是没有涉及词义这个层次,性能有待考证。

(3)基于句子相似度的算法。

一种基于 LexRank 算法的改进的自动文摘系统[5]是通过计算两个句子之间相似度来选取句子,计算句子相似度是从句子中所有词的词频、句子长度、句法结构(肯定句或否定句)、句子位置这些方面来考虑的

该算法的优点是在计算句子相似度的过程中,不仅考虑到词频,还从句子长度、句法结构、句子位置等多个方面考虑,因为这些因素很大程度上影响到句子权值,所以是很可取的做法,缺点依然是没有涉及词义这一层次。下面的算法解决了这个问题。

(4)基于词义的算法。

词频统计忽略了词语语义这个层次,容易漏掉那些词频统计次数较少而表达文章的重要概念的词语,所以应该在计算句子权值时考虑到词语语义层次。

a. 基于 HowNet 概念获取的中文自动文摘系统[6]提出用概念统计代替词形频率统计方法,建立概念向量空间模型,计算句子重要度,并对句子进行冗余度计算抽取文摘句。通过内部评测和外部评测两种方法得出文摘效果较好的结论。

b. 基于 HowNet 自动文摘的研究[7]提出了一种基于知网的词义相似度计算方法。

c. 文本聚类在自动文摘中的应用研究[8]先通过构造词义距离表和句子加权(标题、位置、词频、指示词)来计算句子相似度,开始每个句子自成一类,然后把句子相似度较大的句子归为一类,再计算新类与其它类之间的距离,一直到所有句子聚成所需要的类的个数。

该类算法的优点是利用了词语语义,在概念库的基础上利用了文章中词语语义的层次,这样会使得文摘性能很好,c 算法在计算相似度时不仅考虑到词义还考虑到句子中词的特点和句子本身的特性(标题、位置、词频、指示词),是当前较好的算法,但是没有涉及到文章的结构信息。下面算法考虑了文章结构信息。

(5)基于文章结构的算法。

美国 Cornell 大学的 Salton 等人则将文章视为段落的关联网络,文章中的每个段落被赋予一个特征向量,两个段落特征向量的内积作为这两个段落的关联强度。如果两个段落的关联强度超过给定阈值,则认为两个段落有语义联系。和很多段落都有联系的中心段被提取出来组成一篇文献摘要[9~11],这样就考虑了文章结构方面,有利于生成高质量的文摘。这就是段落间的关联网络,有专家将文章视为句子间的关联网

络,是通过计算句子间相似度来实现。它们的区别是句子间的关联网络时空开销大,难以承受;而段落之间的关联网络要小得多,由段落拼接起来的文摘连贯性有显著提高,但是它可能包含一些无关紧要的句子,不够精炼,为了解决这个问题一般会在抽取句子后做冗余处理,冗余处理目的是去除句子意思类似的句子,算法是先将权重最大的句子作为候选类,然后依次从余下的文摘中选取句子,如果该句子与候选类所有句子的相似度小于一定阈值,则把该句添加到候选类,否则舍弃,重复直到文摘长度达到要求。

1.1.4 多主题自动摘录

(1)多主题划分阐述。

一篇文摘应该将原始文章的主要信息全面地反映给读者,使读者不需查阅原文就可以全面了解主要信息。而很多文章是多主题的,从很多不同方面来阐述。若抽取信息只从句子重要度从高到低抽取,则很容易造成次主题相关信息的遗漏和缺少,破坏了文摘的完整性,所以提出了多主题文摘。一般来说,先按照主题划分把文章分成几部分,然后从每个部分中抽取重要句子形成文摘。多主题文摘最重要的内容是多主题划分问题。作者在阐述一个主题时,其所用重点词汇通常局限在能代表该主题内容的一个较小范围,具有一定的重复性[11]。若两个段落所含词语,特别是关键词,在一定程度上发生重复,即可初步认为这两段谈的是同一主题,若位置合适的话,即应划在同一个语义段中。目前主题划分方法有四种:基于标题的主题划分、相邻段落相似度的主题划分、多个段落相似度主题划分、连续段落相似度的主题划分,文中还提出了一个新的算法。

(2)多主题划分方法。

a. 基于标题的主题划分方法是以子标题作为主题划分的标准,很容易实现但是适用范围较小,只适用于含有子标题的文章。

b. 基于相邻段落相似度的主题划分,如果文章中有话题的转移,则相邻两段间的相似度会变小[12]。其算法是:假设全文共有N段,分别计算各相邻段间的相似度,得到$N-1$个相似度,从中选取最小的一个作为主题划分边界,再在剩下的相似度里找次小的作为主题划分边界,一直重复,直至主题数达到要求或者剩下的相似度都大于设定的阈值。该算法只考虑到了相邻段落间的相关性,不够准确。

c. 多个段落相似度主题划分的算法是首先把文章平均分成几个部分(每一个部分包含相同段落数),然后计算相邻两个部分的相似度,形成相似度数曲线,然后选出所有的低拐点作为候选点,最后从这些候选点中按照一定原则选出主题划分处。该算法考虑到多个段落的内部情况所以比相邻段落相似度算法要更准确。但是一开始把文章分成几个部分时没有考虑到每个部分内部也许会是主题划分处,结果会有失偏差。

d. 连续段落相似度主题划分的思想是对全文所有段落与其它若干个连续的段落相似度进行比较,若某个段落与前面连续的若干个段落相似度小而与后面连续的若干个段落相似度大,这认为该段是主题划分段。如果都较小,那么看该段下一段与后面连续若干个段落的相似度是否较大,如果是,则该段是主题划分段。因为该段很可能是标题段,所含词汇少,所以与其它段落相似度都小[13]。

e. 鉴于以上方法的不完善,文中提出一个新的算法,假设全文共有N段,分别计算各相邻段间的相似度,得到$N-1$个相似度$C_1,C_2\cdots,C_{n-1}$,从中选择相似度大的若干个段落组成为几部分(例如C_1,C_4,C_7较大,则把1、2段作为一部分,4、5段作为一部分,7、8段作为一部分,剩下的每一段作为一部分,形成$N-3$部分),再对这几个部分计算相邻部分间的相似度,再取相似度大的若干个部分组成一部分,一直重复直到部分数(主题数)达到要求或者相似度小于设定的阈值。这个方法既不会产生偏差也考虑到段落的内部情况,是很好的一个算法。

1.2 基于信息抽取的自动文摘

基于信息抽取方法是先对文本进行主题识别,再选择已编好的该领域的文摘框架,对文中有用的片段进行有限深度的分析,利用特征词提取相关短语或句子填充文摘框架,再利用文摘模板将文摘框架中的内容转换为文摘输出。基于信息抽取和文本生成的自动文摘系统设计[14]提出了一个设计方案,包括信息抽取和文本生成两个过程。信息抽取过程是对原文进行词语频率、词语分布和修辞结构的分析,并在此基础上参考用户对摘要的需求,抽取原文的部分内容来填充文摘框架;文本生成过程对文摘框架中的句子进行加工、组织,生成连贯的段落。它的优点是避免了从文中抽取句子顺序输出的不连贯性,而且文摘框架比理解文摘中的脚本等要简单得多,更易于编写。但是它必须先编写一个该主题领域的框架,而文摘框架的编写完全依赖于领域知识,要想应用于多个领域,就必须为每个领域都编写一个文摘框架。并且,由于使用模板生成文摘,使得文摘的语言千篇一律,十分呆板。

1.3 基于理解的自动文摘

基于理解的文摘方法[15]是利用自然语言处理技术生成摘要,其步骤是:

(1)语法分析,借助词典中的语言学知识对原文中的句子进行语法分析,获得语法结构树。

(2)语义分析,运用知识库中的语义知识将语法

结构描述转换成以逻辑和意义为基础的语义表示。

(3)语用分析和信息提取,根据知识库中预先存放的领域知识在上下文中进行推理,并将提取出来的关键内容存入一张信息表。

(4)文本生成,将信息表中的内容转换为一段完整连贯的文字输出。

用这种方法生成的文摘性能最好,但是它很大程度上依托于自然语言处理技术,自然语言处理中句法分析、语法语义分析技术尚未完全成熟,因此如果想获得高质量的语言分析结果,就必须将待处理的语料限制在某个范围之内。一旦限制在某个领域内,文摘的可移植性就差。随着自然语言处理技术中句法、语法语义分析技术的成熟,自动文摘的性能会有相当大的提高。

2　结束语

综上所述,在自动文摘的三种方法中,最主要的方法还是自动摘录方法,用计算句子权值、计算句子相似度和聚类的方法来抽取句子形成文摘,但是这种文摘往往不连贯,就产生了基于信息抽取的方法,虽然这种方法生成的文摘连贯性好但是它完全依赖于领域知识而且生成的文摘千篇一律,这样的文摘性能也不够好。基于理解的方法能生成质量相当高的文摘,它的发展很大程度上依赖于中文信息处理技术,随着中文信息处理技术的发展,生成自动文摘性能会很好。

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] 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):441-443.

[3] 王志琪. 基于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):1298-1299.

[4] 杨建林. 一种使用自动聚类思想的自动文摘方法[J]. 情报学报,2001,20(5):534-535.

[5] 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):152-153.

[6] 王　萌,何婷婷,姬东鸿,等. 基于 HowNet 概念获取的中文自动文摘系统[J]. 中文信息学报,2004,19(3):90-91.

[7] 柴晓丽,张丽伟,管玉玲. 基于 HowNet 自动文摘的研究[J]. 电脑编程技巧与维护,2003,12(3):164-165.

[8] 郭庆琳,樊孝忠,柳长安. 文本聚类在自动文摘中的应用研究[J]. 计算机应用,2005,25(5):1037-1038.

[9] Salton G, Allan J, Buckleym C, et al. Automatic Analysis, Theme Genaration, and Summarization of Machine Readable Texts[J]. Science,1994,264(3):1421-1426.

[10] Salton G, Allan J, Singhal A. Automatic Text Decomposition and Structuring[J]. Information Processing & Management, 1996,32(2):127-138.

[11] Salton G, Singhal A, Mitra M, et al. Automatic Text Structuring and Summarization[J]. Information Processing & Management,1997,33 (2):193-207.

[12] 万　敏. 基于统计和语义分析的中英文自动文摘研究[D]. 北京:清华大学,2003.

[13] 傅间莲,陈群秀. 自动文摘系统中的主题划分问题研究[J]. 中文信息学报,2005,19(6):29-32.

[14] 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]. 情报学报,1997,16(增刊):25-28.

[15] 刘　挺,王开涛. 自动文摘的四种主要方法[J]. 情报学报,1999,18(1):11-16.

(上接第 187 页)

畅,且系统的实时性较高。实践证明,满足直升机计算机辅助训练系统的仿真需求。

参考文献:

[1] 陈东帆. 航空 CBT 中协同训练系统的设计与实现[J]. 计算机工程与设计,2007,28(15): 3727-3730.

[2] 刘丽娇. 基于 GL-Studio 的飞行模拟机虚拟座舱开发[D]. 哈尔滨:哈尔滨工业大学机电工程学院,2009.

[3] 孙　鑫,余安萍. VC++深入详解[M]. 北京:电子工业出版社,2008.

[4] 王大勇. 基于 VAPS 下虚拟仪表开发[D]. 哈尔滨:哈尔滨工业大学机电工程学院,2006.

[5] 朱　敏,陈立奎,王宏伟,等. 基于 GL Studio 的分布式虚拟训练系统关键技术[J]. 兵工自动化,2010,29(8):46-48.

[6] 罗正卫,刘建群. Windows(客户端)和 MS-DOS(服务器)网络通信的实现[J]. 计算机技术与发展,2010, 20(4): 191-194.

[7] 王恩涛,李　祥. 基于 Socket 的手机与数据库服务器通信的研究[J]. 计算机技术与发展,2007,17(2):81-84.

[8] 王俊鸣,张智军,张安旭. 基于 LabWindows/CVI 的多线程技术的电磁兼容预测试系统设计与实现[J]. 弹箭与制导学报,2008, 28(2): 311-314.

[9] 雷振山, 赵晨光. 虚拟仪器系统的网络技术研究与应用[J]. 国外电子测量技术,2006, 25(5): 59-61.

[10] 杨志菊,李　洋. GL Studio 在武器系统仿真模拟中的应用[J]. 电子测试,2010(8):80-86.

[11] 邱岳恒,卢京潮,刘　秉. 直升机视景仿真及座舱仪表显示系统实现[J]. 测控技术,2010,29(7):13-15.