论自动文摘及其分类[1)]

王志琪 王永成 刘传汉

(上海交通大学计算机科学与工程系, 上海 200030)

摘要 自动文摘,即利用计算机自动编制文摘,是信息时代的需要。本文讨论了文摘的不同定义、特点和功能。目前,文摘的分类方法不适用于自动文摘的分类,因此,本文试着从多角度对自动文摘系统进行了分类,这样的分类根据自动文摘的特点进行的划分,是对自动文摘分类的一种总结,可以作为构造自动文摘系统和思考自动文摘发展方向的参考和借鉴。最后,概述了中文自动文摘系统的研究状况,展望了自动文摘的发展趋势。

关键词 自然语言处理 情报科学 自动文摘

On Automatic Summarization and Its Classification

Wang Zhiqi, Wang Yongcheng and Liu Chuanhan

(*Department of Computer Science and Technology, Shanghai Jiao Tong University, Shanghai* 200030)

Abstract Automatic summarization, which makes summaries by means of computers, is the need of the era. The definitions, features and functions of summary are discussed in the paper. As the method of summary classification does not fit with automatic summarization classification, automatic summarization is classified using different standards in the paper. The classification is based on the features of automatic summarization. It will contribute to the construction of automatic summarization and the development of automatic summarization technology. In the end, the research status of Chinese automatic summarization is outlined, and future opportunities and challenges are discussed.

Keywords natural language processing, information science, automatic summarization.

随着科学技术的进步,文献的数量正在迅速的增长,使得专业人员即使不停地阅读也难以覆盖全部文献,因此人们将注意力转向了文摘。文摘是文献的数量聚积到一定程度后的产物,它可以被理解成是文献内容的缩影,文摘的简洁性、准确性和清晰性使得其作用越来越重要。

全世界范围内期刊数量的爆炸性增长,导致二次出版物体积的增加和重要性的提高。纯粹的手工文摘编制速度已经远远落后于文献更新和增长的速度,原始文献出版与文摘出版之间的时差越来越大。为了改变这种状态,人们开始了用计算机进行自动文摘的研究。

Internet 的开通并在全世界范围内的普及,使人们陷入了信息的汪洋大海。高速的通信与交通以及实时的信息使得人们和整个世界的距离缩小了;但大量无用信息的泛滥,几乎淹没了有用的信息。为了发掘出有用的信息而去浏览所有的文章,由于信息量巨大,对任何人,早已不可能。文摘是原始文献的浓缩物,长度只有原文的 10% 或者更少,通过它来了解文献内容,将会节省大量的时间。Internet 上

收稿日期:2004 年 9 月 8 日

作者简介:**王志琪**,女,1978 年生,上海交通大学计算机科学与工程系博士生。主要研究领域为中文信息处理。**王永成**,男,1939 年生,上海交通大学计算机科学与工程系教授、博士生导师。主要研究领域为自然语言理解,网络信息智能化处理。**刘传汉**,男,1970 年生,上海交通大学计算机科学与工程系博士生。主要研究领域为中文信息处理。

1) 基金项目:国家 863 高技术研究发展计划项目(2002AA119050)

的信息数量无法估量,文摘工作只能通过计算机来完成。如果能在自动文摘之后,再进行信息的过滤(去除无用信息和非法信息)和分类,则能按用户要求直接提供其有用信息的文摘,或者直接通过文摘来去除无用和非法信息。

社会各行业都在进行信息库的建设,文摘是信息库的重要组成部分。手工进行文摘的编制,速度跟不上要求,将影响整个信息库建设的进程,而且手工文摘缺乏规范性,不利于统一处理;计算机编制的文摘格式统一,易于建库,易于信息的检索和再处理。

所有这些,都要求文摘的编制必须自动化,文摘自动编制是时代的要求。作为自然语言处理领域的一个重要应用,自动文摘涉及到了大量的理论和应用技术,而且其相关理论方法和技术也可以应用到其他的自然语言处理应用领域,推动整个自然语言处理领域的发展进程。

1 文摘的定义及其分类

1.1 文摘的定义

什么是文摘?不同的研究者给出了不同的定义。

中华人民共和国国家标准《文摘编写规则》(GB 6447 - 86)中,文摘被定义为:"以提供文献内容梗概为目的,不加评论和补充解释、简明、确切地记述文献重要内容的短文。"

美国国家标准学会(ANSI)《文摘编写标准》中给出的文摘定义是:"某一文献内容的简要而准确的表达,不加解释和评论,也不区分这篇文献是由谁写的[1]。"

国际标准《文献工作——出版物的文摘和文献工作》(ISO 214 - 1976(E))中的定义指出:"一份文献内容的缩短的精确的表达而无须补充解释或评论,且对写文摘的人来说没有差别。"

Mani[2]指出,文摘就是从信息源抽取内容,用简练并且用户感兴趣的方式把最主要的内容呈现给用户。

Maybury 认为文摘就是从原文中提取出最重要的信息,并加工成一个简洁的表达提供给用户[3]。

孙春葵、钟义信[4]也给出了一个定义。文摘是对一篇文档中提出的中心主题的概括性的、准确的描述,内容精炼,篇幅短小。它的基本作用是提示性的,帮助读者决定是否有必要对全文进行浏览。除了提示作用外,一篇文摘还可以含有有用信息,如实验结果和主要结论。

本文将文摘定义为:"有关文章主题等内容尽可能简要和方便用户的描述。"

国际和国家标准都规定了"不加评论与解释"之类的要求,也就是对文摘的"客观性"的要求。在实际生活中,任何人对每一样东西的认识都包含了他自己的立场、观点和知识。实验表明,几乎没有两个人对同一篇文章做出相同的文摘来。甚至同一个人,在不同的时间,对同一篇文章所做的文摘都不会完全相同。因此,我们不能要求自动文摘系统绝对排除一切主观因素。我们的定义中的"主题等内容",是指文章的主题或用户所需的偏重信息;"尽可能简要",是由摘录者和用户的水平所决定的;"方便用户"要求文摘的形式多样,例如,纲目型文摘。这里的偏重,是指用户常常要求从不同的角度去看文章。譬如,对于江泽民总书记的一个报告,有些领导分管农业,有些分管工业,他们所要求的文摘就需要有所侧重,而不是笼统的为一般人看的文摘。

实际上,文摘的确难定义。国际著名的模糊数学大师 L.A.Zadeh 对此也有同感。文摘在中文中也可以称为摘要、概要、提要、梗概、简介等,在英文中则有 summary、brief、compendium、precis 等。使用什么术语并不十分重要,只要摘出的内容满足需要即可。

自动文摘就是利用计算机对文献编制的文摘。当然,人们希望自动文摘的结果能够满足人们的需要。国际上对自动文摘的研究可以说是与自然语言处理其他领域的研究同时起步的。由于计算机硬件限制和自动文摘研究缺乏基础性技术,所以,自动文摘在 20 世纪 50—70 年代发展相当缓慢;从 80 年代末期开始,自动文摘技术才进入蓬勃发展、百家争鸣的时代。

1.2 文摘的特点和功能

文献[5]概括了文摘的三大特点和七项功能,具有一定的代表性。三个特点是:

简洁性:文摘比所摘的文献短,长度为原文献的 5% ~ 10%的文摘就能基本上反映文献的主要内容;当文摘的长度达到原文献的 10% ~ 25%时,很多文章的写作风格就可以在文摘中体现出来了。

准确性:无论长短,文摘必须准确无误地报道原文献的基本内容,不能主观改变原文观点,科技文献的文摘应确保正确引用原文中的各项数据。

清晰性:必须使用一种易读的文体把文献内容

清晰地表示出来,最好用完整的句子编写文摘,并尽可能使用著者自己使用的词语。

文摘的这些特点决定了文摘具有以下的功能:

- 促进新资料的快速通报;
- 节省阅读时间;
- 有助于选择文献;
- 有助于克服语言障碍;
- 有利于文献检索;
- 提高标引效率;
- 帮助人们撰写评论文章。

1.3 文摘的分类

在自动文摘近50年来的历史上,专家学者们提出并发展了文摘的分类方法。根据分类标准的不同,我们可以对文摘进行不同的分类[2,4,5]。目前国内外比较常用的文摘分类方法是依据文摘的功能(用途)进行分类的。该分类标准把文摘分为了三类:指示性文摘、报道性文摘、评论性文摘[1]。

(1)指示性文摘(Indicative Summary)

有时又称为摘录性文摘。这种文摘指示读者:若查询原始文献,将会发现什么。该类文摘应对文本中的主要内容,特别是其创造性部分利用摘录的办法进行简明的介绍。指示性文摘比较适合多主题的文献。因为它包含的数据比较少,所以往往不能代替原始文献。美国的《工程索引》、日本的《科技文献速报》等都是这一类的指示性文摘工具。

(2)报道性文摘(Informative Summary)

旨在向读者提供原始文献中的定量情报和定性情报。定量情报就是数值性数据,包括数值、范围、公式、平均值等。报道性文摘特别适用于那些描述实验性研究的报告和单主题的文献,是对文本题目的一种补充和说明。报道性文摘往往可以使用户不必查阅原始文献。前苏联的《文摘杂志》、英国德温特公司的《基本专利文摘》均属于这一类。

(3)评论性文摘(Critical Summary)

这是近代发展起来的一种文摘。在这种文摘中,文摘员也是文献的评论者,与各国早期关于文摘的定义要求不同,文摘员应插入自己的看法和分析。评论性文摘的价值很大程度上依赖于文摘员的专业水平,由于自己没有做实验或者查阅相关资料,文摘员很容易产生判断错误,所以编写评论性文摘时,必须格外小心。文摘机构通常不允许采用评论性文摘。但好的评论性文摘对读者的作用最大。美国的《应用力学评论》、前苏联的《力学文摘》都提供评论性文摘。

2 自动文摘的分类

目前情况下,由自动文摘系统产生的文摘都具有指示性和(或)报道性作用。很少区分该文摘究竟是报道性的还是指示性的。显然,上述分类方法不适应自动文摘分类的现实。

文献[6]和[7]将自动文摘系统分成了两种:基于统计的机械文摘系统和基于意义的理解文摘系统,这和自然语言处理的两种指导方法——经验主义和唯理主义相对应。文献[8]把自动文摘系统分为了四种:自动摘录、基于理解的自动文摘、信息抽取和基于结构的自动文摘。本文尝试从多角度对自动文摘系统进行分类,这样的分类根据自动文摘的特点进行的划分,是对自动文摘分类的一种总结,可以作为构造自动文摘系统和思考自动文摘发展方向的参考和借鉴。

2.1 按文摘面向的用户划分

可以划分为通用文摘(Generic Summarization)和偏重文摘(Biased Summarization)。

通用文摘和偏重文摘的区别在于是否考虑了用户的兴趣。通用型文摘就是面向所有用户的、文摘内容不带有任何侧重的、全面反映原文内容的文摘。它是对全文信息的浓缩,是对原文所描述的主题、范围和结果的一种简洁概括。这种文摘是面向原文中心思想的、静态的文摘,不能适应用户的个性化或查询要求。对于一篇长的文章,如果用户只关心某一方面(例如工业),这就涉及到了偏重问题。

偏重文摘也称为用户聚焦文摘(User-focused Summarization)、主题聚焦文摘(Topic-focused Summarization)或查询聚焦文摘(Query-focused Summarization)。它可以根据需要或者用户的兴趣提供相应的有侧重点的文摘。偏重文摘的结果不仅仅决定于原文的主题,也决定于用户的个性化要求。它能够把焦点放在用户关心的部分,而不是把原文的每个部分平等对待。偏重文摘考虑了用户的兴趣,这是实现用户个性化文摘必不可少的技术。

偏重文摘的出现有两个现实意义:首先,在形成偏重文摘的过程中,强调用户的要求,使文摘结果能满足用户特殊要求;其次,在搜索引擎系统中,可以根据查询要求返回一个简短的文摘。用户可以快速浏览这个简短的结果,来判断原文与查询要求的相

关度[5]。在判断相关度方面,由于偏重文摘考虑了原文主题和用户的查询两个方面,偏重文摘比通用文摘和现有搜索引擎所提供的方式更加可靠。

2.2 按文摘处理的文本对象划分

可以划分为单文档文摘(Single Document Summarization,SDS)和多文档文摘(Multiple Documents Summarization,MDS)。

单文档文摘处理的文本对象是单篇文摘,它对每篇文章独立的生成文摘。而多文档文摘处理的文本对象是有多篇文档组成的文档集,它对这个文档集生成一个概括多篇文档内容的综合文摘。

随着在线信息的快速增长,提供一些有效查找和合理描述文本内容的机制正变得越来越重要。传统的信息检索系统包括现代搜索引擎都是基于用户查询的最大相关性来查找和排序文档,这样用户为了找到目标信息仍然需要阅读大量的文档从而获得最终的目标文档的相关部分的内容。对用户的查询,搜索的结果可能有几百个相关文档,有大量的信息重复,同时在某些部分又有区别,这样就需要一种特殊的技术来处理。

多文档文摘就是从一个文档集中去除冗余,考虑文档相互的关联性及各自的差异,产生一个浓缩的信息集[9]。多文档文摘实际上是对单文档文摘的一个扩展,它与单文档相比较需要一些新的技术和方法来处理,它主要考虑以下几个方面的问题[10]:

(1)需要一个高效地去除冗余的方法。

(2)系列文档可能包含时间及空间的变化。

(3)文摘结果压缩比很大,通常 1% ~ 10%,而单文档可以在30%左右。

(4)发生在不同文档中的事件及实体,它们的关联、融合处理是一个难题。

对于用户对目标信息的不同需求,有的需要一个主要包含文档集的共同(或交叉)内容的文摘,有的需要对一组文档内容的总览,多文档文摘可以有不同的类型,比如:

(1)文档集的共同部分,找到一组文档的共同的最重要的相关部分,然后使用它们生成一个文摘。

(2)文档集的共同部分加各自的独特部分。

(3)中心文档文摘,计算出文档集最重要的文档,然后做单文档文摘。

(4)中心文档加外围文档描述组成的文摘。

(5)最新的文档加外围文档描述组成的文摘。

2.3 按文摘的制作方法划分

可以划分为摘录型文摘(Summarization Based on Extraction, SBE)、基于理解的文摘(Summarization Based on Understanding, SBU)、模板型文摘(Summarization Based on Template,SBT)和基于结构的文摘(Summarization Based on Discourse Structure, SBS)。

摘录型文摘中大部分的句子都是直接或间接的选自原文,只有少数句子经过加工整理而成。这种方法充分利用计算机的计算能力,采用统计的方法绕过文章意义的理解问题,它将文本视为句子的线性序列,将句子视为词的线性序列。在进行文摘时,首先计算词的权重,然后计算句子的权重,再从文章中挑选出权重大的句子,按照句子在原文中的自然序列进行排列,加以修饰最终生成文摘输出。在摘录型文摘中,主要的依据有关键词、题名、位置、线索词、段首段尾等文章的特征部分。通过对这一系列的文本的形式特征的分析找出文章最重要的部分。这种方法在实际使用中,处理速度快,对于一般的文章以及结构规范的文章处理效果较好。经验测试表明,对文本的位置和线索词特征处理效果明显,可适合处理大部分任意文章。同时,它的处理不需要非常复杂的语言学知识,比较容易移植到多种语言处理。然而,具体地讲,一篇文章常常在某些形式特征上符合常规,而在另一些形式特征上违反常规,或者是在文章的某一部分符合常规,而在另一部分违反常规,摘录的结果能否抓住原文的中心内容要看文章在多大程度上符合常规;因此,自动摘录的质量不是很稳定。但是由于处理中过分依赖文章中规范的结构进行分析,而对句子或段落没有进行意义分析,因此,这种文摘存在明显的不足。尤其是对于包含有多个主题的文章进行文摘工作时常常发生遗漏主题以及文摘不连贯的问题。另外,这种自动摘录对于一些结构不规范的文章的处理效果比较不理想也是它一大弱点。

基于理解的文摘方法是建立在人工智能、自然语言处理的基础上的,它利用语言学知识对文章进行复杂的语法分析、语义分析和语用分析,最后进行文摘的生成。基于理解的自动文摘通常有以下步骤:

(1)文本预处理。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。

由于这种方法实现了对文章内容的理解,并且,文章的文摘句子是另外生成的,因此,从文摘的质量角度来看,文摘与文章的内容符合较好,且语句精炼,连贯性好。但是,由于目前在人工智能与自然语言处理方面还存在许多难以解决的问题,这种方法难以得到快速的发展。另外,由于给予理解的自动文摘与文章涉及的领域之间有密切联系,如果将某个领域的理解文摘推广到另一领域,则需要做较大的修改,使得文摘系统难以移植。

模版型文摘有预先定义好的框架,文摘的生成过程其实就是从原文中检索出文摘模版所要求的内容,填到文摘模板中即可。模版型文摘通常有以下步骤:

(1)特征提取。检索原文,提取出模板需要的文本特征。

(2)特征规范。将抽取出来的特征进行规范,可采用句子到短语的压缩、同义词替换等方法。

(3)填充模板。将规范后的特征填充到模板的相应位置,生成文摘输出。

基于结构的文摘采用自上而下分析方法,首先对文章的结构进行分析,再逐渐细化到段落、句子和概念,整个的分析过程是一个自上而下的过程,即由上层分析逐渐细化到底层分析。一般说来,文章中的不同部分承担着不同的功能,各部分之间在逻辑上是有一定的关联的。文章的这种关联找到了,其核心部分也就自然能够找到。这也就是基于结构的文摘方法的思想方法。应该说这种方法更利于从全局的观点把握原文作者的意图。但是,目前说来,语言学对于文章结构的研究还很不够,可用的形式规则就更少了,这使得基于结构的自动文摘方法到目前为止还没有形成一套完整的理论方法。

2.4 按照文摘是否需要学习样本划分

可以分为有监督学习文摘和无监督学习文摘。

有监督学习的文摘分为学习和文摘两个过程。学习过程主要利用人工文摘进行学习,从中找出进行自动文摘的特点或者参数。然后,在文摘过程中利用之前学习到的知识或参数进行文摘。而无监督学习无需对人工文摘的学习过程。一般说来,有监督学习的文摘系统面向特定的领域,文摘质量和训练的样本质量有关系。

3 自动文摘的发展现状

我国在自然语言理解领域较早地开展了研究,机器翻译走在了世界的前列。中文自动文摘的研究起步较晚,1985年才有人正式撰文介绍国外的自动文摘的研究情况[11]。从20世纪80年代末,我国才开始研究自动文摘实验系统,至今也只有10余年的历史。

自动文摘的关键技术主要是以自然语言处理技术作为基础的,包括分词、词性标注、句法分析和自动语义分析等,在很多方面采用了和西文相类似的方法和技术。但是,汉语作为一种特殊的语言又有其许多特殊的方面,主要表现在:

(1)西文语言为拼音文字,而汉语为表意文字;

(2)西文的书面语言,词与词之间有空格,而汉语的词与词之间无空格;

(3)西方语言的同音词相对较少,而汉语的同音词很多;

(4)西方语言多有形态变化,而汉语缺少形态变化;

(5)汉语的语法尚未形成规范化,而且人们习惯于非规范化的语法。

这些特点,给中文自动文摘,乃至中文信息的计算机处理带来了一定的困难。

在实际自动文摘的应用系统的研发上,近年来,国内外先后有多所大学和一些研究机构开展了研究,建立了一批实验系统,取得了许多重要成果。表1中列出了主要的中文自动文摘系统研究发展情况。

4 自动文摘的发展展望

想让计算机像人一样能够阅读各类文章,并做出令人满意的文摘,还有很长的路要走。其实这也是计算机领域的一个宏伟目标。根据自动文摘这些年来的的研究与实践,本文认为应该在以下几个方面进一步研究,使自动文摘的发展上升到一个新的台阶。

表1 中文自动文摘系统研究成果简表

单位	主要方法	时间(年)
上海交通大学 王永成[12,13,14,15,16]	最初为面向科技领域的自动文摘系统,随后扩展至可处理新闻 领域文本。目前,系统版本不断升级,性能也逐渐提升,可处理领域基本不受限制。 系统设计主要采用仿人算法,它融合了标题法、位置法、关键字串、词频分析、章法分析、主题敏感辞分析等多种方法为一体,综合分析文本主题生成文摘。	1988
哈尔滨工业大学 王开铸[17,18,19] MATAS 系统	MATAS 系统为基于意义的理解文摘系统。首先对文本分析生成篇章意义的机内表示 TMR,对 TMR 进行句子级、上下文级压缩并加权,选择权重较大句子生成文摘。	1992
HIT－863 I 系统	HIT－863 I 系统为基于统计文摘系统。首先将输入原文转换成系统定义的机内表示,然后采用特征词提取,利用加权函数对句子加权,选取权重较大句子构成文摘。	1992
HIT－97 I 系统	HIT－97 I 英文自动文摘系统采用词频统计、词类标注、意义段划分、关键句提取、关键句压缩等多种方法,将统计式文摘与理解式文摘的优点较好地结合在一起。	1997
HIT－863 II 系统	HIT－863 II 系统采用基于篇章多级依存结构(TMDS)的自动文摘方法,利用 TMDS 分析获取文本的主题,选取文摘。	1999
北京邮电大学 钟义信[20,21,22,23] GLANCE 系统	GLANCE 系统为非受限领域自动文摘系统。它主要是基于对各种主题信息的统计计算,以此筛选出文摘。	1993
NEWS 系统	NEWS 系统针对新华社外事新闻进行分析,提出了基于言语行为理论的话语分析方法,实现了对该类新闻的篇章理解和报道性文摘生成。	1997
LADIES 系统	LADIES 系统基本属于摘录型文摘,其最大特点是采用全信息词典来对文本进行分析处理。	1997
LADIES-NEW 系统	LADIES-NEW 系统在原有系统基础上增加了文摘句式的自动学习模块和生成规则的测试模块部分,使得系统具有一定学习功能。	2000
复旦大学 吴立德[24,25] FDASCT 系统	首先对文本进行分词,进行文本特征信息提取,然后进行词性与语义标注,随后进行词、句子、段落加权处理,最后根据权重进行选取句子构成文摘。	1996
文本自动综述系统	该系统是多文档文摘的一种推广。采用基于统计的文本自动综述方法,利用文档内和文档之间段落的语义相关性,实现多文档的自动综述。首先对文本进行分段实现信息分割;再对文本段进行聚类实现信息凝聚;最后抽取代表段产生综述结果实现信息压缩。	2003
东北大学 姚天顺[26]	采用脚本知识表示,通过与用户交互获取文摘。	1989
中国科学院软件研究所 李小滨[27] EAAS 系统	局限于"就业机会介绍"领域。首先通过与用户交互获得信息焦点集,然后对文章进行语法语义分析,生成文章意义框架,最后按照信息焦点集从框架中获取信息,生成文摘。	1990
清华大学 罗振声[28,29]	基于主题概念的自动文摘方法。以概念统计和层次分析为基础,利用 Word Net 以概念统计代替传统的词频统计,基于主题概念构建向量空间模型,计算句子重要度。并且根据主题概念在概念层次树上的分布进行文本结构分析划分意义块,以意义块为单元抽取文摘。	2002
南京大学 李明[30]	对文本中的汉字进行字频统计得出关键字,以此为基础从原文中选取候选文摘句,经适当加权后摘出合适的文摘句输出。	1996
山西大学 郭炳炎[31]	基于统计的方法分析文本结构,然后依据结构信息,辅助词分析方法抽取文摘。	1996
杭州大学 姜贤塔[32]	基于语料库的方法,利用"后邻字符树"的方法在领域语料库中生成字符树库,用于自动文摘候选句子选取时提高选取精度。	1998

(1)基于理解的文摘系统大多受限于特定领域,难以移植,发展空间不大。因此,目前的文摘系统大多属于摘录型文摘系统,文摘中的语句基本上都来自于原文,只是进行一些简单的修饰和润色,这些语句让我们感觉仍然是那么的冗长。今后的发展方向应该充分考虑自然语言的特点,研究句子的特性,发现一些信息浓缩的基本规则,并应用于摘录型自动文摘系统中,使得语句在主题不变的情况下能够进行压缩,从而使得这些文摘系统生成的文摘更接近于人工文摘。

(2)需要加强对概念的研究,形成一套以概念为核心的自然语言分析体系。这个体系的最终目标可能是绕过目前以词为核心的一些滩点问题(例如,词的定义、分词、语法分析、语义分析等)。

(3)多文档文摘是亟待研究的一个课题。在和搜索引擎集成时,这种文摘尤其重要。多文档文摘可以对连续报道做一个评述型文摘,这比单文档文摘更能节约用户的时间。

(4)便携式设备和移动电话的发展为自动文摘的应用提供了新的机遇和挑战。在这种环境中,要求自动文摘系统提供更高的压缩能力(比如将一篇新闻报道压缩成一句话甚至一、两个词),还可能要求自动文摘系统和声控系统集成起来,能对日常的对话进行文摘。

(5)文摘及自动文摘系统的评价问题尚需进一步研究。内部评价方法通常采用主观性判定方法,其评价结果的稳定性、可重复性都较差;而外部评价方法又具有局限性大、难以标准化等缺陷。因此,迫切需要研究出一套客观性较强、标准化程度较高的测试标准,以促进自动文摘领域的研究。

参 考 文 献

1 Weil B H. Standards for writing abstracts. Journal of the American Society for Information Science, 1970, 21(5): 351 ~ 357

2 Inderjeet Mani. Automatic Summarization. John Benjamins Publishing Company, 2001

3 Mayburym M. Generating Summaries from Event Data. Information Processing and Management, 1995, 31(5): 735 ~ 751

4 孙春葵. 自动文摘及其知识获取技术研究[博士论文]. 北京邮电大学, 2000

5 Borko H, Bernier C L. Abstracting Concepts and Methods. Academic Press, 1975

6 吴岩,李秀坤,王开铸. HIT-971型英文自动摘要系统. 情报学报, 1998, 17(5): 358 ~ 364

7 刘伟权. 自然语言理解与汉语文本信息处理理论研究. 北京邮电大学[博士论文], 1997

8 刘挺,王开铸. 自动文摘的四种主要方法. 情报学报, 1999, 18(1): 10 ~ 19

9 Inderjeet Mani. Automatic Summarization. John Benjamins Publishing Company, 2001

10 I. Mani, E. Bloedorn. Summarizing Similarities and Differences Among Related Documents. Information Retrieval, 1999, 1(1): 35 ~ 67

11 王兵. 美国机编文摘概况. 情报学报, 1985, 4(2): 166 ~ 171

12 苏海菊,王永成. 中文科技文献摘要的自动编写. 情报学报, 1989, 8(6): 433 ~ 439

13 莫燕,王永成. 中文文摘摘要的自动编制. 现代图书情报技术, 1993(3): 10 ~ 13

14 陈桂林,王永成. Internet网络信息自动摘要的研究. 高技术通讯, 1999(2): 33 ~ 36

15 沈洲,王永成,韩客松. 一种基于主题敏感辞分析的新闻文献自动摘要系统的研究与实践. 高技术通讯, 2001(9): 28 ~ 32

16 Wang Yongcheng, Liu Gongshen, Shen Zhou, Bao Zhengrong, Hu Peihua. On Automatic Summarization, In Proceeding of First International Conference of Chinese Information Processing (ICCIP2001), Shanghai, 2001: 6 ~ 8

17 王建波,唐正伟,杜春玲,王开铸. 篇章物理结构和意义结构的一种形式化表达方法. 情报学报, 1996, 15(4): 291 ~ 299

18 刘挺,王开铸. 基于篇章多级依存结构的自动文摘研究. 计算机研究与发展, 1999, 36(4): 479 ~ 488

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

20 杨晓兰,钟义信. 基于全信息词典的自动文摘系统研究与实现. 情报学报, 1997(5): 408 ~ 414

21 孙春葵. 自动文摘及其知识获取技术研究.[博士论文].北京邮电大学, 2000

22 刘伟权. 自然语言理解与汉语文本信息处理理论研究.[博士论文].北京邮电大学, 1997

23 郭祥昊. 语言信息处理理论及自动文摘关键技术的研究.[博士论文].北京邮电大学, 1998

24 吴立德. 大规模中文文本处理. 复旦大学出版社, 1997

25 郑义,黄萱菁,吴立德. 文本自动综述系统的研究与实现. 计算机研究与发展, 2003, 40(11): 1606 ~ 1611

26 姚天顺. 自然语言理解——一种让机器懂得人类语言的研究. 清华大学出版社, 1995

27 李小滨,徐越. 自动文摘系统EAAS. 软件学报, 1991, 3(4): 12 ~ 18

28 季姮,罗振声,万敏,高小云. 基于概念统计和语义层次

分析的英文自动文摘研究. 中文信息学报,2003,17(2):14 ~ 20
29 万敏,罗振声,季姮,高小云. 基于概念统计的英文自动文摘研究. 计算机工程与应用,2002(24):7 ~ 16
30 李明. 从字频统计出发的中文文摘自动编写. 现代图书情报技术,1996(3):42 ~ 45
31 薛翠芳,郭炳炎. 中文自动文摘系统. 第五届全国 AI 联合会会议论文集,1998:200 ~ 206
32 姜贤塔,陈根才. 利用语料库技术的中文自动文摘系统. 中文信息学报,1999,13(2):16 ~ 23

(责任编辑　马兰)