

·实践平台·

面向计算机类文献的自动摘要系统的研究与实现

黄水清 李志燕 梁 刚 (南京农业大学信息科技学院 江苏南京 210095)

摘 要：本文首先介绍了自动摘要系统研究的目的、意义，自动摘要系统的发展历史。然后，归纳总结出了面向计算机相关领域文献的自动摘要系统生成过程，并设计实现了面向计算机相关领域的自动文本摘要系统。最后选取了 315 篇相关文献作为测试语料，测试后的结果比较满意。

关键词：自动摘要 自动摘录 计算机

中图分类号：G356.7,TP391

文献标识码：A

文章编号：1003-6938(2006)03-0093-05

On Automatically-abstracting System for the Computer Documents : a study and its accomplishment

Huang Qingshui Li Zhiyan Liang Gang (The College of Information Technology of Nanjing Agricultural University , Nanjing Jiangsu 210095)

Abstract：This article first introduced the aim and significance of studying the automatically-abstracting system with a brief review on its history. Then, the generative process of automatic abstract related with computer documents was summed up to achieve a further design and accomplishment of the system. And the utilizing of this system on 315 selected related articles gained comparatively satisfactory result.

Key words：automatic summarizing ; automatic extracting ; computer

CLC number : G356.7,TP391

Document code : A

Article ID : 1003-6938(2006)03-0093-05

1 引言

摘要是一种通行的解决人的阅读速度跟不上信息量增加速度的办法。人们可以根据摘要了解文章大意，决定是否阅读这篇文章，大大节约了查找时间，提高了阅读效率。摘要有人工编写与计算机自动生成两种，计算机生成摘要称为自动摘要。自动摘要的发展历史并不算久远，是伴随着计算机技术发展成长起来的。1958 年 4 月，卢恩 (Luhn) 发表了一篇关于自动摘要研究的学术论文——“The Automatic Creation of Literature Abstracts (Auto-Abstracts)”^[1]，自此

揭开了计算机自动编制摘要系统研究的序幕。1958 年 10 月，巴森代尔 (P.B.Baxendale) 发表了一篇关于对科技文献进行标引实验的文章——“Machine-made index for technical literature - an experiment”^[2]，分析结论和标引的权重分配，对后来的自动摘要研究有着很大的启迪与推进作用。

二十世纪六十年代及以前，自动摘要的研究可视为探索、试验和理论研究阶段。七十年代以后，自动摘要的研究逐渐走向实用阶段，许多实用系统开始诞生。到目前为止，自动摘要的研究已经有近 50 年的历史了，其价值已充分显露出来。

我国从 1985 年开始介绍国外自动摘要方面的研究情况,从 80 年代末开始研究自动摘要实验系统。在形式特征方面,汉语与西文之间有很大的区别,西文中词汇间存在明显的切分标志(例如“空格”等),但汉语中却没有这样的切分标志,汉语词汇混合在一起,切分比较困难。在其他方面,汉语同样存在一些有别于西文的特点。因此,汉语自动摘要系统的实现也相对困难。

本文介绍了一个针对计算机学科的汉语文献自动摘要系统的实现过程与方法。每个学科都有其独有的特点,各个学科的学术论文也有其各自的结构特征。计算机文章一般篇幅都比较长,字数多,更新快,时效性强,而且计算机类文献中所含的重要句多,大多数文章包含有摘要段,几乎所有文章中都有总结性段落。这些特点既说明了专门设计面向计算机类的自动摘要系统是必须的,也为面向计算机类的自动摘要系统的实现提供了理论依据。

2 面向计算机类文献的自动摘要系统的实现过程

本系统采用基于统计的自动摘要的生成方法,基于统计的自动摘要的过程一般比较简单,大致可以分为待摘文本信息预处理、文本信息分词处理、词频统计、计算句子权重、摘要句提取、生成摘要等几个主要步骤。^[30 40 50 61]

2.1 预处理

预处理是自动摘要生成的第一步,也是非常重要的一步,它直接影响到生成摘要的质量。本系统的测试数据来源于中国期刊网(CAJ 文件)和其它下载的网络文献(PDF 文件及 HTML 文件),预处理首先就是要将这些不同格式的文件转换成文本文件。

将其他格式的文件转换成文本文件后,预处理工作并没有全部完成,还有两部分工作要做,第一部分处理工作主要解决的是在将 PDF、CAJ 等格式的文件转换成文本文件过程中,出现乱码和每行成段(将 PDF 和有些 CAJ 文件转换成文本文件的时候,原文中的每行将成为一段)的现象,此时机器无法完成这些预处理工作,需要借助人的力量来完成。第二部分处理工作主要由以下三项工作组成:(1)去掉在转换过程中段落前后的空格;(2)去掉文中的空行;(3)将全角的英文字符转换成半角的英文字符。并不是所有的待处理文献都要进行这两部分的预处理工作,只有在文件格式转换过程中出现错误的情况下才加入第一部分处理工作,其他情况下是不需要人工处理的,例如从网络上下载的文章一般情况下是不需要这一过程的。

2.2 分词处理

分词方法有多种,比如正向最长匹配算法、逆向最长匹配算法、正向最短匹配算法、逆向最短匹配算法等。在这几

种传统的机械分词方法中,逆向最长分词算法的切分精度最好,其错误切分率只有 1/245,低于其他几种分词算法的错误切分率,所以在系统实现的过程中,选用逆向最长分词算法作为词汇切分的方法。

本系统自行设计了一种改进过的逆向最长匹配汉语分词算法。在给定的分词词表的基础上进行汉语分词时,不但能成功切分出分词词表中已有的词,而且还能自动识别出分词词表中没有的词,即未登录词。该分词算法可以有效地解决大多数未登录词的识别问题,并且能减少分词错误,同时对分词算法的效率基本没有影响。

自动分词的目的是从文中提取出“重要词”,去掉“非重要词”,经过计算机自动分词处理之后,将关键词提取出来,作为“重要词”(同该词出现的段落数/第几个段落)和句子数/段落中的第几个句子)一起存入数据库中备用。

2.3 词频统计

“重要词”能在一定程度上反映文本信息的主题内容,“重要词”出现的次数越多,说明它的重要程度越高。词频信息统计即是对自动分词后得到的“重要词”进行词频统计,统计出每个“重要词”在文本信息中出现的次数。在本系统实现的过程中,词频统计是通过数据库完成的,用到了 SQL 的 group by 语句和 count 等聚集函数。

2.4 句子权重计算

句子权重计算就是根据词频、位置等信息计算出句子的权重。在自动摘要过程中,文本中有些内容非常重要,它们会直接影响到摘要的质量。例如,标题中出现的关键词、重要词的频率、词出现的位置、句子长度、句子结构等,它们在自动生成摘要的过程中,对摘要句的筛选、摘要的组织等发挥着重要作用。句子的权重由两方面的因素决定:句子的具体内容和句子在文本信息中具备的特征信息。^[71]计算句子权重时所涉及的标准主要有以下几点:

(1) 词频信息

词频是指“重要词”在文本信息中出现的频率。一般来讲,具有价值的词汇往往是中频词,高频词一般是反映句子语法结构的虚词,例如“的”、“得”、“在”、“是”等词,而低频词不能反映文献主题。同样,在摘要中发挥重要作用的词汇也是具有较高频率的关键词(重要词),而这些词在整个文章中属于中频词。对这些“中频词”数量的统计可以作为计算句子权重的一个重要标准,这也是基于统计的自动摘要方法的核心思想。

(2) 各级标题

文本信息的标题是文本内容的重要体现,文本的各级标题都不同程度地反映了文本所讨论的主要内容。因此,标题中的词汇是摘要的重要素材,其中关键词与原文内容和

讨论主题往往有紧密的联系。剔除了标题中的功能词,余下的关键词可作为抽取摘要句的“重要词”,也就是说出现在标题中的关键词更加重要。

(3) 线索词

文章中会有许多短语(词汇)用于引申出反映文本内容的总结性的句子,这类短语或词汇叫线索词。这类指示词有如下形式:“本文论述了”、“本文的目的”、“综上所述”等等,这些线索词后所接的句子往往高度概括了文献主题。因此,这些句子被选作为摘要候选句的可能性非常大。在本系统实现的过程中,设计了线索词表,当句子中出现线索词时,权重加大。

(4) 废止停用词

与线索词的作用相反,当文中出现某些词汇(如:“例如”,“for example”)时,其所在句子所反映的内容与文章主题一般相差较远,不易用来作为摘要候选句,这些词汇称之为废止停用词。在本系统实现的过程中,设计了废止停用词表,当句子中出现废止停用词时,权重降低。

(5) 位置信息

不同位置的句子对文章及段落的主题贡献是不相同的,出现在首段、末段等位置的句子成为摘要句的可能性很大,在进行自动摘要的过程中,有必要提高句子的权重。

(6) 句法结构

文章中的句子形式多种多样,有陈述句、疑问句、感叹句、省略句等等,但真正反映文章主题内容的主要是陈述句。因此,选择摘要句时,应尽可能地抽取陈述句,避免疑问句、感叹句、省略句等形式的句子进入摘要。对于句法结构的判断比较简单,只要对句子的最后一两个字符进行判断即可,如果最后的字符是问号、感叹号或省略号时,将句子的权重降低。

(7) 句子长度

摘要表现为短而精,即以简短的文字概括文章论述的主要内容。过度冗长的句子通常不宜选入摘要中,但过于短小的句子所能表达的信息量较小,也不能选入摘要。所以在句子的筛选过程中,应该选择长度适中的句子作为摘要句。

从上面的分析可以看出,句子的权重与句子中重要词的数量、重要词的权重、句子位置及句子长度等均有关系,可以用下面的式子计算句子权重:

$$W(S) = \mu_p \mu_s \frac{\sum_{i=1}^n W(T_i)}{l} + \mu_l \quad \text{公式 1}$$

公式 1 中, $W(S)$ 表示句子 S 的权重, $W(T_i)$ 表示句子中所含重要词的权重, l 表示句子的长度, μ_p 和 μ_s 代表句子的位置信息,其中 μ_p 代表句子出现在文中的第几段, μ_s 代表句子在段落中的位置。

2.5 摘要句提取

(1) 一般候选句的选取

选取候选句子就是按照句子的权重高低,根据事先设定的阈值或用户所要求的摘要长度筛选出摘要句。

(2) 摘要段的选取

根据以上的统计,计算机类 63.75% 的文章中都包含有摘要性段落,摘要段具有自己的结构特点,可以综合利用 SQL 语句和正则表达式来提取摘要段。正则表达式提供了一种从字符集合中搜寻特定字符串的机制^[8],它可以让用户通过使用一系列的特殊字符构建匹配模式,然后把匹配模式与数据文件、程序输入等目标对象进行比较,根据目标对象中是否包含匹配模式,执行相应的程序。^[9]

正则表达式主要有四类字符:首先是匹配字符,包含了需要搜索匹配的对象。例如,“.”号匹配任一字符,“[]”号匹配方括弧之间的任一字符,“^”号匹配不在方括弧之间的任一字符。其次是重复操作符,描述了查找一个特定字符的次数。例如,“?”号匹配某一部分一次,“*”号匹配某一部分多次,“+”号匹配某一部分一次或多次,还有“{n}”、“{n,N}”等符号。此外还有锚,它指定了所要匹配的格式,有“^”、“\$”、“\b”、“\d”、“\s”、“\w”等。正则表达式也存在一些保留字符,需要用反斜杠来替换它们,比如“*”号,是个保留字符,如果要在表达式里表示这个特定的字符,就得用“*”这种形式。

SQL (Structured Query Language, 结构化查询语言) 是关系数据库的标准语言,它功能强大,是一个通用的关系数据库语言,几乎全部的关系数据库软件都支持它。利用 SQL 语句,可以提高句子匹配效率,简化操作。

结合正则表达式与 SQL 的相关知识,设计摘要段的匹配语句如下所示:

```
select Content from TableName where Content like ' *
第 1 - ] 节 * 第 2 - ] 节 * 第 3 - ] 节 * ' 或:
```

```
select Content from TableName where Content like ' *
本文 * 首先 * 其次 * 最后 * '
```

其中,“Content”代表存放段落的字段,“TableName”代表数据表的名称,“*”属于正则表达式的一个符号,代表其所在位置可以是任何字符,“[1 -]”代表此处可以是字符“1”或“-”,“[2 -]”代表此处可以是字符“2”或“-”,“[3 -]”代表此处可以是字符“3”或“-”。利用上面的语句可以成功匹配绝大多数的摘要段落。

2.6 生成摘要

生成摘要是自动摘要系统的最后一步工作,可以分为两个步骤:首先将提取出来的摘要句按照其在文献中出现

的次序进行排列,此时摘要已经基本完成,但由于文本是一个有机的整体,抽取出来的句子失去了上下文的支持,有可能产生理解歧义或无法理解的现象;第二个步骤是对生成摘要的润色处理,它的作用是弥补前一步生成摘要的不足,使摘要变得更加通顺易于理解,增强了摘要的连贯性。润色处理需要有特定润色规则的支持,就现阶段而言,润色规则的建设还属于初级阶段,需要进一步地改进。

3 系统测评

自动摘要系统内部测试方法和外部测试方法各有优缺点,单纯用一种方法无法体现自动摘要系统的整体性能,采用多种摘要测试方法对自动摘要系统进行测评。测试语料库为315篇计算机领域相关文章,由三部分组成:《软件学报》期刊2004年的165篇文章;其他计算机期刊2004年以后发表的75篇文章;网上75篇计算机相关文章。

3.1 “理想摘要”对比法分析

由于网上的75篇文章没有作者摘要,所以没有做“理想摘要”对比测试。对于《软件学报》,相似程度大于70%的摘要占总数的32.73%,小于40%的摘要占总数的26.66%;而对于其他期刊,效果相对较差,大于70%的摘要占总数的25.33%,小于40%的摘要占总数的34.67%,效果较差的几篇文章中有80%以上来自非核心期刊。

3.2 摘要可接受评价方法分析

摘要可接受评价方法也是一种主观的测试方法,用户将获得的摘要与原文进行对照,参考事先确定的一些定性的指导性评价标准,根据评价者的主观感觉来对摘要进行评价,评价结果为可接受与不可接受。经过三位用户的测试,有222篇机编摘要可以接受,占总数的70.48%。

3.3 自动摘要外部评价方法分析

自动摘要的外部评价方法是一种客观的评价方法,评价结果更加客观可信。一般的学术论文都会配有关键词,这些关键词基本上是作者给出的,能够在一定程度上反映文章的主题内容。笔者设计利用文章中提供的关键词去检索生成的摘要,根据关键词检索成功率公式2,公式3可以评判摘要的质量。

$$\text{检索成功率} = \frac{\text{生成摘要中所包含的关键数量}}{\text{作者提供的关键词数}} \quad \text{公式 2}$$

$$\text{平均检索成功率} = \frac{\text{成功检索出的关键词数量}}{\text{文献集合中关键词总数}} \quad \text{公式 3}$$

测试语料库中有两部分可以做外部测试,即《软件学报》的165篇文章和其他期刊的72篇文章(另3篇文章中没有提供关键词)。经过测试,处理《软件学报》后生成的机

编摘要中有38篇检索成功率为100%,占总数的23%,有4篇摘要无法检索到任何一个关键词,占2%,平均检索成功率为65.97%;处理其他期刊后生成的机编摘要中有17篇检索成功率为100%,占总数的24%,有2篇摘要无法检索到任何一个关键词,占3%,平均检索成功率为64.54%。表1中显示了检索情况,检索成功率大于80%的摘要分别占总数的38%和26%,检索成功率大于60%的摘要分别占总数的67%和70%。总体而言,检索成功率大于60%的摘要为162篇,占总数的68.35%。

表1 检索成功率测试结果

	软件学报	所占百分比	其他期刊	所占百分比
检索成功率 $\geq 80\%$	63	38%	19	26%
检索成功率 $\geq 60\%$ 且 $< 80\%$	48	29%	32	44%
检索成功率 $\geq 40\%$ 且 $< 60\%$	31	19%	9	13%
检索成功率 $\geq 20\%$ 且 $< 40\%$	19	12%	9	13%
检索成功率 $< 20\%$	4	2%	3	4%

除此之外,也可以用另一个指标来评价摘要的好坏,即用“作者提供的关键词数量”与“成功检出的关键词数量”之间的差值来评判,二者的差值越小,说明摘要的质量越高。在处理软件学报后生成的机编摘要中,差值小于等于1的摘要数为86篇,占总数的52%,小于等于2的摘要数为123篇,占总数的75%;在处理其他期刊后生成的机编摘要中,差值小于等于1的摘要数为47篇,占总数的66%,小于等于2的摘要数为60篇,占总数的84%。总体而言,差值小于等于2的摘要数为183篇,占总数的77.22%。

4 总结

自动摘要技术属于自然语言学研究的范畴,其研究意义非常深远。本文对基于统计的自动摘要技术进行了研究探讨,并在计算机相关领域进行了实践,最后选取了315篇相关文献作为测试语料,测试后的结果比较满意,但相关技术理论还有待进一步提高。

参考文献:

- [1] Luhn H. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development. 1958(2): 159-165.
- [2] Baxendale P. B. Machine-made index for technical literature—an experiment[J]. IBM Journal. 1958(10): 354-361.
- [3] Paice C. D. Constructing literature abstracts by computer: Techniques and Prospects[J]. Information Processing

- & Management. 1990, 26(1): 171-186.
- [4] Zechner K. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentence[C]. In Proc. Of the 16th International Conference on Computational Linguistics. 1996: 986-989.
- [5] 王知津. 基于句子选择的自动文本摘要方法及其评价[J]. 现代图书情报技术, 1998(1): 46-52.
- [6] 苏新宁主编. 信息检索理论与技术[M]. 北京: 科学技术文献出版社, 2004.
- [7] 刘挺, 王开铸. 自动文摘的四种主要方法[J]. 情报学报, 1998, 18: 10-19.
- [8] The Single UNIX Specification, Version 2[EB/OL]. <http://www.opengroup.org/onlinepubs/007908799/xbd/re.html>.
- [9] 吕晓波. 正则表达式使用详解[EB/OL]. <http://dev.csdn.net/article/8/8254.shtm>.
- 作者简介: 黄水清(1964-), 男, 1988年毕业于北京大学图书馆情报学系, 获理学硕士学位, 南京农业大学信息科技学院教授, 研究方向为计算机信息检索; 李志燕(1982-), 女, 南京农业大学信息科技学院硕士研究生; 梁刚(1979-), 男, 南京农业大学信息科技学院硕士研究生。

(上接第58页)

7 结语

如前所述, 对网站进行科学评价的社会需求促使了很多评价机构的诞生, 也促进了网站评价研究的发展。从总体上说, 国外的网站评价研究和实践要比国内的成熟, 这一点值得我们借鉴、学习。

无论是国外的研究还是国内的研究, 无论是采用何种方法对网站进行评价, 如何合理地选择样本网站都是研究人员必须直面的问题。本文在网站评价中的样本网站选取与确定方面做了试探性的研究, 总结出了若干经验。首先, 运用搜索引擎、推荐网站、网站导航进行“滚雪球”式的收集, 力争做到全面收录某类网站; 其次, 利用相关指标(如 Alexa 的“每百万用户的访问数(Reach per million users)”)对网站进行从高到低的排序; 最后, 运用统计软件对指标进行分析处理, 根据相关原则最终确定样本网站。最后, 利用 Google 的各种检索式来测度样本网站的内外链接数及网页数, 计算出网络影响因子和外部网络影响因子。

本文所采用的指标数据是从国外的数据仓库 Alexa 中获得的, 而其所提供的数据也存在一定的局限性。例如, Alexa 通过 alexa 工具栏来收集数据, 而 alexa 工具栏目前只支持 IE 浏览器, 对于一些使用非 IE 浏览器的站点和用户会减少统计。从目前国内的研究来看, 尚未有专门的机构提供可靠的数据, Alexa 所提供的数据也不失为一个折中的选择。面对如此庞大的市场, 我们应该在网站的实时监测和数据仓库的建设方面加强建设, 全面促进我国网络化的发展。

参考文献:

- [1] [8] 中国互联网信息中心[EB/OL]. <http://www.cnnic.net.cn/> 2006-2-25.
- [2] ALEXA 网站[EB/OL]. <http://www.alexa.com/> 2005-3-28.
- [3] Forrester 调研公司网站[EB/OL]. <http://www.forrester.com/> 2005-04-10.
- [4] 商务部信息化司. 2005 年中国政府网站绩效评估报告[EB/OL]. wzpx/index.shtml/2006-04-24.
- [5] 刘雷鸣, 王艳. 关于网站评估模式的比较研究[J]. 情报学报, 2004(2): 198-203.
- [6] T. C. Almind, P. Ingwersen. Informetric analyses on the world wide web: methodological approaches to webmetrics[J]. Journal of Documentation. 1997(4).
- [7] P. Berthon. Positioning in Cyberspace: Evaluation of Telecom Websites using Correspondence Analysis[J]. Information Resources Management Journal. 2001(14): 13-21.
- [9] [10] 社会科学检索词表[S]. 北京: 社会科学文献出版社, 1996: 5-20.
- [11] 新浪网[EB/OL]. <http://www.sina.com.cn/> 2006-04-30.
- [12] 搜狐网[EB/OL]. <http://www.sohu.com/> 2006-04-30.
- [13] Google 搜索引擎[EB/OL]. <http://www.google.com/> 2006-04-30.
- [14] 百度搜索引擎[EB/OL]. <http://www.baidu.com/> 2006-04-30.
- [15] 互联网观察推荐网站[EB/OL]. <http://www.internetobserver.com/> 2004-12-28.
- [16] 中文电信黄页[EB/OL]. <http://www.china114.com/> 2004-12-25.
- [17] 中文酷站大全[EB/OL]. <http://www.chinacool.com/> 2004-12-21.
- [18] 一目了然网[EB/OL]. <http://www.1yml.com/> 2004-12-20.
- [19] 邱均平, 陈敬全. 中国大学网站链接分析及网络影响因子探讨[J]. 中国软科学. 2003(6): 151-155.
- [20] 沙勇忠, 欧阳霞. 中国省级政府网站的影响力评价——网站链接分析及网络影响因子测度[J]. 情报资料工作, 2004(6): 17-22.

作者简介: 王知津(1947-), 男, 南开大学商学院信息资源管理系教授、博士生导师; 郑红军(1981-), 男, 2003年安徽大学信息管理与信息系统专业本科毕业, 现为南开大学国际商学院情报学硕士研究生。