一种篇章结构指导的中文 Web 文档自动摘要方法

王继成 武港山 周源远 张福炎
(南京大学软件新技术国家重点实验室 南京 210093); (南京大学计算机科学与技术系 南京 210093)
(gswu@nju.edu.cn)

摘 要 “摘要”、“关键词”是对文档内容提供简要概括的元数据,在 Web 信息检索中起着重要作用. 针对 Web 信息检索的需求和 Web 文档的特点,采用拟人思维,提出了一种以篇章结构为指导的自动摘要方法. 该方法对段落之间的内容语义关系进行分析,进而划分出文档的主题层次,得到文档的篇章结构;在篇章结构的指导下,使用统计方法和启发式规则来提取文档的关键词、关键句,生成文档的摘要. 在实验评估中,该方法取得了令人满意的摘要质量和速度.

关键词 自动摘要;篇章结构;Web;信息检索

中图法分类号 TP391; TP393

Research on Automatic Summarization of Web Document Guided by Discourse

WANG Ji-Cheng, WU Gang-Shan, ZHOU Yuan-Yuan, and ZHANG Fu-Yan
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)
(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract Two important metadata elements, abstracts and keywords, that summarize the content of web document, are of much assistance in web information retrieval. According to the demands of web information retrieval and the characteristics of web document, an anthropopathic method for automatic summarization of Chinese web document is proposed, which is guided by document discourse and based on indicative abstract. This method partitions the document into a hierarchical structure by parsing the semantic distance between each adjacent paragraph, uses statistical approaches and heuristic rules to extract keywords and key-sentences, and finally creates the abstract. Experiments show that this method can generate abstraction effectively and efficiently.

Key words automatic summarization; discourse; web; information retrieval

1 引 言

摘要(abstraction),也称为总结(summarization),是指按照用户的要求以简洁的形式准确地表达文档的主要内容. 从应用的领域来看,摘要分为针对特定领域(例如:金融、军事等)的摘要和非受限领域的摘要;从接受的对象来看,摘要可以是针对特定用户有选择地对文档的某些部分进行,也可以涉及文档的全部重要内容,供不确定的用户群体使用.

在 Web 信息检索中,“摘要”、“关键词”是对文档内容提供简要概括的元数据,对文档的“标题”、“主题类别”起着补充作用. 人们有时仅从检出文档的标题和分类无法判断其是否符合要求. 如果能进一步给出文档的摘要、关键词,那么用户不必浏览全文就可以作出相关性判断,这无疑将提高检索的效果和效率[1]. 此外,用户还能通过关键词和摘要来检索所需文档. 可见,Web 文档的自动摘要在智能

收稿日期:2001-10-24;修回日期:2002-12-31
基金项目:国家自然科学基金(60073030);国家“八六三”高技术研究发展计划项目基金(2001AA110334);富士通研究所“Web 文档清洗”项目基金

化检索系统中占有相当重要的地位.

对自然语言表达的文档进行自动摘要是一个始终没有得到很好解决的问题. 特别是在 Web 环境下, 大规模、多领域、非规范的文档对自动摘要技术提出了新的要求. 本文对如何运用智能化技术, 从 Web 文档中快速、准确地自动提取关键词和摘要信息进行了探索. 文中首先对文档自动摘要的相关工作进行了简要介绍; 然后提出了一种以篇章结构为指导、以机械摘要为基本手段的自动摘要方法, 并对其中涉及的篇章分析、关键词提取、摘要生成等关键技术进行了深入分析; 最后, 给出了相应的实验结果和分析.

2　相关工作

自从 Luhn 在 1958 年发表了第 1 篇关于摘要自动生成方法的文章以来, 国外对文档自动摘要方法开展了研究, 取得了比较丰富的成果. 国内对中文自动摘要的研究起步于 1980 年以后, 从事这方面研究的主要有上海交通大学、复旦大学、山西大学、北京邮电大学等单位. 目前, 自动摘要技术总体上分为两类: 基于统计的机械摘要方法和基于知识的理解摘要方法.

2.1　机械摘要

机械摘要的基本思想是: 使用统计方法来获取文档的关键词, 并结合提示词(cue phrase) 、位置等启发信息, 从文档中挑选出一些合适的句子, 进行润色后得到文档的摘要. 机械摘要方法具有应用领域不受限制、速度快、摘要长度可调节等优点, 因此, 大多数自动摘要系统均采用了这种方法. 例如: Kupiec 等人开发的“Trainable Document Summarizer”[2], 复旦大学完成的“复旦中文自动文摘系统”[3]、上海交通大学的“中文自动编制文摘实验系统 SJTUCAA”[4] 等. 但是, 机械摘要方法局限于文档的字面表层, 生成的摘要质量较差, 存在反映内容不够全面、语句冗余等问题.

2.2　理解摘要

和机械文摘这种经验主义方法不同, 理解摘要期望利用各种知识和形式化理论, 在理解文档语义内容的基础上生成文摘(对原文的概括或浓缩) . 与机械摘要相比, 理解摘要方法采用了复杂的自然语言理解和生成技术, 因此摘要质量较好, 具有简洁精炼、全面准确、可读性强等优点. 但是, 理解摘要不仅要求计算机具有自然语言理解和生成能力, 还需要表达和组织各种背景、领域知识. 这些工作的难度十分巨大, 迄今为止进展甚微. 因此, 理解摘要方法的使用比较少见, 仅限于非常狭小的应用领域中. 例如: 哈尔滨工业大学实现了一个军事领域的自动文摘实验系统[5], 北京邮电大学研制的文摘系统 LADIES[6].

2.3　Web 信息检索对文档摘要的要求

为了达到辅助信息检索的目的, Web 文档摘要应该不受领域的限制, 供各种不确定的用户群体使用. 从摘要质量上来看, 要求并不是很高, 只要能够准确、全面地指明文档的内容梗概即可; 但是摘要的速度必须很快, 以满足大量 Web 文档的处理需求.

在上面介绍的两种摘要方法中, 理解摘要牺牲了领域宽度, 换取理解深度, 具有很高的理论探索价值, 但在 Web 信息检索中是不必要也是不可行的. 目前, 绝大部分搜索引擎所采用的方法是截取文档的前几行, 这种方法又过于简单、片面. 可以说, 虽然相关的研究工作已经开展了数十年, 但 Web 文档的自动摘要还有很大的发展空间, 可用性更有待于进一步加强. 为此, 本文期望能够寻找一种合适的文档摘要方法, 尽可能地满足 Web 信息检索在领域、用户、速度、质量等各方面的要求.

3　篇章结构指导的文档自动摘要

一篇文档由很多个段落(paragraph) 组成. 通常, 一些连续的段落在内容上是相近的, 形成了一个语义相对内聚的节(section) , 对应于一个小的子主题. 若干个子主题统一在文档的大主题之下. 对人工摘要过程进行观察后发现, 文摘员在做摘要前一般需要通读全文, 把握文档的中心思想和篇章结构, 权衡各个主题的轻重, 从而使文摘能够尽可能地覆盖文档中的有用信息. 因此, 我们认为文档自动摘要应该模拟人类思维, 建立在篇章分析的基础之上, 这样才能够比较全面、准确地反映文档内容.

下面, 我们给出一种对 Web 文档进行自动摘要的方法. 该方法采用拟人思维, 运用智能化技术来分析文档的篇章结构, 然后根据各个主题的轻重来提取关键词、生成摘要. 这样既使得文档的大主题能够得到充分反映, 又不至于忽略掉文档中的一些次要主题, 从而克服摘要的表层性和片面性. 该方法的具体过程如图 1 所示, 其中包含以下 3 个关键的部分: ① 篇章结构分析: 对文档的内容进行分析, 发现其中的主题层次, 形成文档篇章结构表示; ④

关键词提取: 通过计算词条的权重, 从文档中提取能够反映文档主题内容的重要词条; ㈣ 摘要生成: 在篇章结构的指导下, 并利用关键词等信息提取文档中的关键句, 形成最终的文档摘要

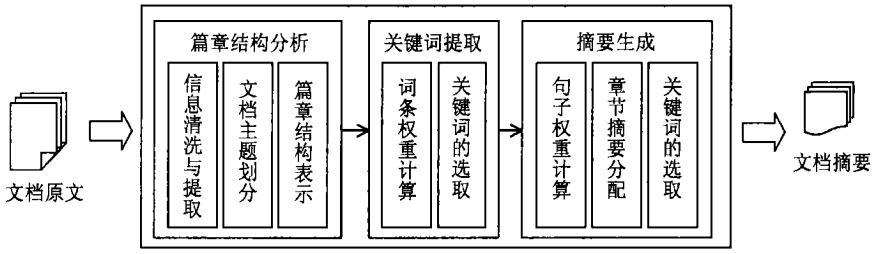

图 1 篇章结构指导的文档自动摘要

3.1 篇章分析

(1) Web 文档的信息清洗与提取

与传统的文本文档相比, Web 文档的一个显著特点在于, 它通过 HTML 标记(TAG) 提供了对文档摘要工作有利的一些辅助信息, 包括: 文档标题(〈*Title*〉), 各级子标题(〈H_1〉, 〈H_2〉, …, 〈H_n〉), 强调词/句(〈*B*〉, 〈*Strong*〉) 等等. 但是, Web 文档中也存在着一些对文档摘要不利的因素, 即 Web 文档不像传统的文本那样整齐、干净, 其中包含了大量噪声, 例如: 为了增强用户交互性而加入的 Script 等; 为了便于用户浏览或出于商业因素所加入的导航链接、广告链接等; 标注版权、日期等信息的文本等等. 如图 2 所示:

图 2 Web 文档的信息清洗

上述因素的存在, 使得在进行 Web 文档自动摘要之前进行信息清洗和提取工作具有特别重要的意义. 这一前处理工作将根据后续摘要工作的需求, 在对 Web 页面进行解析的基础上, 将页面中不需要的部分去除, 得到其中的文档正文, 同时提取出重要的辅助信息. 限于篇幅, 有关这方面的详细情况我们另文介绍[7].

(2) 文档主题的划分

由于段落之间存在自然分割, 因此篇章分析的主要任务是发现文档正文中包含的各个子主题, 即将连续的自然段分割为若干个节. 一种直观的方法是利用 HTML 语法来分析 Web 文档的段落信息, 但是这仅适用于一些结构良好的文档, 同时也忽略了文档片断之间的语义关系. 为此, 我们在对 Web 页面进行解析的基础上结合语义分析的手段来划分文档主题

假定文档 d_i 由 m 个自然段 p_k 构成, 记为 $d_i=\{p_1, \cdots, p_k, \cdots, p_m\}$. 下面, 我们给出一种基于语义内容的主题划分方法. 该方法模仿了人类提炼文档主题的思维机制, 采用智能化方法将内容相近的段落归并到一起, 得到文档的主题层次

首先, 通过计算每两个连续段落之间的语义距

离来判断它们在内容上的相似程度. 文档 d_i 经过分词处理后, 得到文档的每个段落中所包含的词条及相应的词频信息. 这些信息是对段落语义内容的一种近似表示. 任何两个连续段落 p_k 和 p_{k+1} 之间的语义距离定义为

$$distance(p_k, p_{k+1}) = 1 - \| p_k \cap p_{k+1} \|_T / \| p_k \cup p_{k+1} \|_T, \quad (1)$$

其中, $\| p_k \cap p_{k+1} \|_T$ 是 p_k 和 p_{k+1} 所具有的相同的词条数目(考虑同一词条的多次出现); $\| p_k \cup p_{k+1} \|_T$ 是 p_k 和 p_{k+1} 所具有的所有的词条数目. 显然, 段落间语义距离的取值范围位于区间 $[0, 1]$ 之内, 语义距离越大, 说明二者在内容上的差异越大. 图 3 给出了文档段落语义距离的一个示例, 其中, 序号 $k(1 \leqslant k \leqslant 19)$ 处的取值表示的是段落 p_k 和 p_{k+1} 之间的语义距离. 可以看出, 示例文档"浅谈下一代存储器"一共由 20 个自然段组成, 段落间语义距离的最大值为 1.00, 最小值为 0.26.

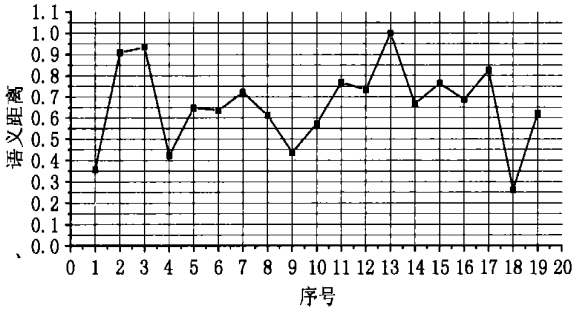

图 3　文档"浅谈下一代存储器"的段落间语义距离

其次, 通过段落之间语义距离来发现主题的改变, 从而得到文档的篇章结构. 每个段落边界都是主题发生变化的潜在点. 为了将内容相近的段落合并在一起, 可以考虑使用聚类方法, 但是此处的聚类仅限于合并彼此相邻的段落. 具体而言, 首先将每个段落作为一个临时的节; 然后依次将具有最小语义距离的两个相邻临时节合并, 形成一个新的临时节; 不断重复这一过程, 直到剩下的节的数目符合需要为止(在本文中设定为不超过 5). 这种方法是一种"自底向上"的合并方法. 与此相对的是一种"自顶向下"的分割方法, 即: 首先将整个文档作为一个临时的节; 然后依次从具有最大语义距离的相邻段落处将临时节分割, 得到两个新的临时节; 不断重复这一过程, 直到生成的节的数目符合需要为止. 这两种方法在结果上是等价的, 二者都会产生一个树状图(dendrogram), 最后得到的每个节一般均对应一个子主题. 但是, 在文档包含段落数目较多时, "自顶向下"分割的速度要快于"自底向上"合并.

图 4 中给出了文档主题生成树的一个示例. 该生成树是在图 3 中给出的段落语义距离的基础上使用"自顶向下"分割方法所得到的结果. 从图中可以看出, 示例文档"浅谈下一代存储器"中包含的 20 个段落, 可以从段落 13, 14 之间粗分为两个部分, 这两个部分又可以分别进一步分割为更小的部分. 查阅示例文档的原文可以看出, 原文的主题是"存储器", 该主题由"快闪存储器"、"铁电存储器"两个子主题构成; 每个子主题又分别包含"技术特点"、"市场状况"等部分.

可见, 采用上述主题划分方法得到的结果与人工理解文档内容的结果是基本吻合的.

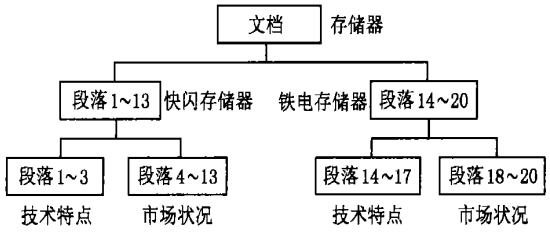

图 4　文档主题生成树

(3) 文档的中间表示

在提取文档的版面信息、划分文档的主题层次以后, 我们采用了一种中间形式来表示文档内容, 如图 5 所示. 这种表示形式将文档的篇章结构按照层次组织起来, 每个层次通过双向链接与它的上、下层(父子结点)建立联系, 同一层次的相邻元素(兄弟结点)之间也建立联系. 这样, 在后续处理工作中, 可以方便地获得任何一个节、子节、段落、句子中所包含的词条、位置以及上下文等多种信息.

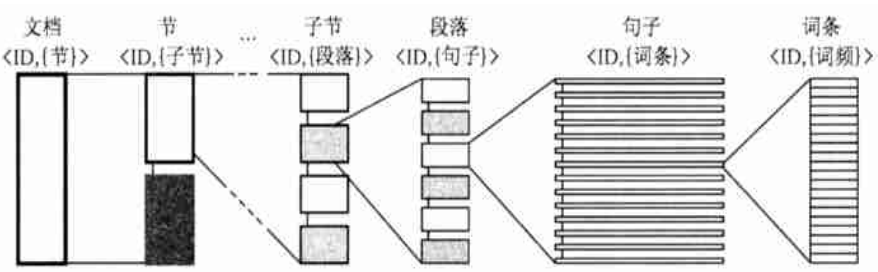

图 5　文档内容的中间表示

由于上述分析、表示方法并不涉及领域知识, 因而可以适用于多个领域中的各种规范或非规范文档. 同时, 篇章结构比文档字面表层深入了一大步, 能够更为全面、准确地探测出文档中包含的主题层次.

3.2 关键词提取

(1) 词条权重计算

关键词提取是文档摘要的一个重要环节, 同时, 关键词本身也是文档的一种重要的元数据. 为了定量地衡量词条的重要性, 需要给文档 d_i 中的每个词条 t_j 赋予权重 w_{ij}. w_{ij} 的确定通常使用 TFIDF 方法, 即综合考虑 t_j 在文档 d_i 中的词频 tf_{ij} 以及在整个文档集合 D 中的逆文档频率(inverse document frequency) idf_j. idf_j 有多种可行的计算方法, 本文中采用的方法是:

$$idf_j = \lg((N_D + 0.5)/n_j), \quad (2)$$

其中, N_D 为文档集合中包含的所有文档数目, 而 n_j 为文档集合中出现过词条 t_j 的所有文档数目. 这样, 就得到 w_{ij} 的一种可行的计算方法:

$$w_{ij} = tf_{ij} \times idf_{ij} = tf_{ij} \times \lg(N_D + 0.5)/n_j). \quad (3)$$

在 Web 文档中, 一些特殊位置出现的词条具有特殊的重要性, 因此我们对〈$Title$〉, 〈H_n〉等标记中包含的词条权重分别乘上相应的比例因子 λ. 凡是多处出现的词条, 仅考虑具有最大值的比例因子. 在本文中, 经过反复实验比较, 最终将〈$Title$〉标记的权重比例因子设定为 1.4, 其余的〈H_n〉标记的权重比例按照下式依次递减:

$$\lambda = 1 + 0.4 \times 1/(n + 1). \quad (4)$$

采用上述方法计算出文档中所有词条的权重后, 将词条按照权重的大小降序排列. 选取具有最大权重的若干个词条(通常 5 个左右)作为文档的关键词. 例 1 给出了示例文档“浅谈下一代存储器”的词条、相应的词频以及权重. 可以看出, 所得到的关键词很好地反映了文档的主题内容.

例 **1**. 文档关键词提取示例.

Web 文档:“浅谈下一代存储器”

URL: http://168.160.224.37/cw1997/nrxs.php3?lsh=4212

文档词条(按词频降序排列): 公司 60, 快闪存储器 45, 铁电存储器 28, 美国 16, 存储器 15, 动态随机存储器 13, 研制 9, 推出 9, 世界 8, 价格 8, …

文档词条(按权重降序排列): 快闪存储器 182.7, 铁电存储器 113.6, 动态随机存储器 52.8, 存储器 48.3, 存取速度 30.0, 擦除 26.3, 公司 17.9, 研制 17.8, …

文档关键词: 快闪存储器, 铁电存储器, 动态随机存储器, 存储器, 存取速度, …

(2) 关键词的补充

考虑到有些文档中可能包含有多个子主题, 而有些子主题在上述方法得到的前 5 个关键词中无法体现出来, 此时我们采用了一种补充关键词的方法: 对每个子主题所对应的节分别采用上述方法计算该节中各词条的权重, 得到该节中词条的降序排列; 如果排在首位的词条未包含在文档关键词中则将它补充进去. 但是, 当关键词数目超过 10 个时, 也会分散用户的注意力, 因此要注意避免产生过多的关键词.

3.3 摘要生成

经过观察可以发现, 文摘员在进行人工摘要时通常具有一些显著的特点: ① 在手工摘要中, 绝大部分句子都从原文中直接抽取或者只修改了一点. ④包含有提示词的句子通常应该作为候选的文摘句, 例如:“综上所述”、“本文论述了”等. ㈣文档前言、结论、节首、节尾, 以及段首、段尾中的一些句子具有极高的切题性. 基于上述事实, 在篇章结构分析和关键词提取的基础上, 我们根据各个主题的轻重, 采用统计方法和启发规则来提取关键句, 形成文档摘要.

(1) 句子权重计算

为了定量地衡量句子的重要性, 需要给文档中的每个句子 s_k 赋予权重 $w(s_k)$. $w(s_k)$ 主要由以下几个因素决定:

① 句子中包含的词条的重要性. 句中词条权重之和越大, 则说明句子的重要性可能越大. 由于文档中相当一部分词条对文档内容的影响不大, 因此可以只考虑词条按权重降序排列的前面部分(例如前 10 个), 称之为扩展关键词集. 同时, 词条权重之和应该除以句子所包含的词条总数, 得到句子的平均词条权重, 从而消除句子长度的影响.

④句子在文档中的出现位置. 前言、结论等处的句子通常比其他位置的重要性要高.

㈣句子中是否包含有提示词, 例如:“综上所述”、“本文论述了”等. 如果包含, 那么句子往往对文章的主题内容进行了概述.

¼ 句子是否以“例如”等细节性词条开头, 这些词条的出现意味着句子包含举例成分, 并非概要性语句, 因此重要性相对较低.

综合考虑上述因素, 定义句子权重的计算公式

如下:

$$w(s_k) = LC(s_k) \times CC(s_k) \times EC(s_k) \times \sum t_{kj} / \| s_j \|_T, \quad (5)$$

其中, $\sum t_{kj}$ 是 s_k 中包含的扩展关键词词条权重之和, $\| s_k \|_T$ 是 s_k 中包含的所有词条数目, $\sum t_{kj} / \| s_k \|_T$ 为 s_k 的平均词条权重; $LC(s_k)$ 是 s_k 的位置权重比例因子, 各个位置所对应的因子被设定为: LC(前言句) = LC(结论句) = 1.5, LC(节首句) = LC(节尾句) = 1.3, LC(段首句) = LC(段尾句) = 1.1; $CC(s_k)$ 是 s_k 的提示词权重比例因子, 本文中设定为 1.5; $EC(s_k)$ 是 s_k 的细节词权重比例因子, 本文中设定为 0.5.

(2) 关键句选取

在生成文档摘要之前, 需要先确定摘要的大小. 该值可以是摘要占原文档大小的比例, 也可以用摘要包含的句子数目或者字数来表示. 同时, 所生成的摘要的大小应该是可以动态调整的, 以满足用户的各种需要.

为了能够准确、全面地反映文档的主题, 我们没有像一般的机械摘要方法那样, 直接从全文中按照权重选取关键句, 而是以篇章结构为指导, 将摘要大小分配到文档的各个主题中去. 考虑到首节(fc)和尾节(lc)往往是对全文的概括, 因此这两个部分的摘要比例应该相对高于其他部分, 式(6)给出了从首节和尾节中摘取的句子数目 $\| Abst(c) \|_S$:

$$\| Abst(c) \|_S = \begin{cases} [1.5 \times R \times \| c \|_S] + 1, & \text{给定摘要比例 } R, \\ [1.5 \times SN \times \| c \|_S / \| d_i \|_S] + 1, & \text{给定摘要句数 } SN, \end{cases} \quad (6)$$

其中, $\| c \|_S$ 是首节或尾节中包含的句子数目, $\| d_i \|_S$ 是文档 d_i 中包含的句子总数, $[\,]$ 为取整函数. 剩下的摘要比例或句数分配到文档的其他节中. 对于节 c, 从中所要摘取的句子数目为 $\| Abst(c) \|_S$ 为:

$$\| Abst(c) \|_S = \begin{cases} [(R \times \| d_i \|_S - \| Abst(fc + lc) \|_S) \times \| c \|_S / \| d_i - fc - lc \|_S] + 1, & \text{给定 } R, \\ [(SN - \| Abst(fc + lc) \|_S) \times \| c \|_S / \| d_i - fc - lc \|_S)] + 1, & \text{给定 } SN, \end{cases} \quad (7)$$

其中, $\| Abst(fc + lc) \|_S$ 是首节和尾节摘要所包含的关键句数目, $\| d_i - fc - lc \|_S$ 是文档 d_i 中去掉首节和尾节后所包含的句子总数.

在每节中, 使用式(5)计算各个句子的权重, 并将句子按照权重降序排列; 使用式(6)或式(7)确定从该节中摘取的关键句数, 从句子序列中选择权重最大的、相应数目的句子作为关键句. 为了保证良好的可读性, 并维持与原文一致的逻辑性, 将从各节中选出的关键句按照它们在原文中的出现顺序进行排列、合并, 或采用与原文相对应的段落格式输出, 这样就得到了最终的文档摘要.

可以看出, 该方法得到的摘要具有与一般的机械摘要结果相类似的性质: ① 摘要过程简化、速度快; ②摘要比例可调整, 小比例摘要完全包含在大比例摘要之中; ㈣ 摘要效果与文档所属的领域无关; ¼ 摘要内容由从原文中抽取的句子组成并保持了原文的分段、位置信息. 同时, 由于采用了篇章结构作为指导, 所得到的摘要结果能够比较全面、准确地反映文档的主题内容.

4 实验结果与评价

4.1 实验设计

针对自然语言处理技术面向真实语料、面向实用化的趋势, 我们从"新浪"、"计算机世界报"等 Web 站点上收集了 11440 篇(约 100MB) IT 领域的中文 Web 文档作为实验样本. 实验中采用篇章结构指导的文档摘要方法对上述所有文档进行了自动摘要. 对文档摘要结果的测试主要包括: 文档主题划分以及关键词提取的准确度, 文档摘要的速度与质量. 实验平台为赛扬 500MHz, 128MB 内存.

由于文档摘要所具有的不确定性, 因此在摘要质量的评估上缺乏比较理想的定量评估方法. 目前常用的评估方法[8]主要依靠人工主观地从以下几个方面进行考虑:

① 完备性(exhaustivity)或覆盖率(coverage): 即摘要是否能反映出文档的主要内容, 是否发生遗漏主题的现象;

② 文摘的概括性(generalization)和紧凑性(compact): 即文摘对源文本的概括程度或压缩程度, 是否有语句冗余;

㈣ 可读性(readability): 即摘要在内容、语句顺序上的连贯性, 还有语气的流畅程度等.

4.2 主题划分与关键词提取的评价

我们从实验样本中随机选取了 228 篇文档, 采

用人工对文档进行理解和分析, 得到文档包含的主题及子主题、关键词(由于人工分析工作量大, 因此仅能选取少量文档). 将人工分析结果与计算机得到的结果进行比较, 可以得到主题划分与关键词提取的准确度.

实验结果表明, 主题自动划分的准确度达到了 81. 6%, 这一准确度已经基本上能够满足为摘要提供指导信息的需求. 主题划分错误会影响文档中各个部分所分配到的摘要关键句比例, 但不会对 Web 信息检索工作造成直接影响. 关键词自动提取的准确度达到了 90. 3%, 无论是作为文档元数据直接提供给用户, 还是作为摘要的基础都能够较好地满足 Web 信息检索的需求. 此外, 由于本文的方法不涉及领域知识, 因而主题划分与关键词提取的效果与文档内容所属的领域基本无关. 但是, 我们也观察到, 实验效果与文章的体裁有一定的关系, 科技论文的效果要好于其他体裁.

4. 3 摘要结果的评价

我们对实验结果中的 453 篇文档在摘要质量上进行了人工评估. 评估时, 将每篇摘要与文档原文相对照, 在综合考虑完备性、概括性、可读性等因素之后, 对文档摘要的可接受性(acceptability) 进行打分. 表 1 给出了可接受性的分级、具体要求与相应的评价结果. 从中可以看出, 绝大部分摘要均能够满足完备性和概括性的要求, 反映了文档的主要内容. 经过仔细分析后发现, 由于本文提出的摘要方法以篇章结构为指导、以机械摘要为基本手段, 因此摘要的质量与原文档的写作方法有一定的关系. 当文档的写作思路比较清晰、开头结尾概括了文档的主要内容时, 摘要效果较好; 而在原文中内容混杂, 使用了一些文学性手法, 开头和结尾没有直接切入主题, 描述性语句居多时, 摘要的效果较差. 要能够处理好后一种类型的文档, 需要系统具备很强的理解能力和概括能力.

表 1 摘要质量的可接受性评价

可接受性	具体要求	符合要求的摘要数目	所占比例/ %
好	摘要的完备性、概括性和可读性均较好 能够全面、准确、精炼地反映原文主要内容, 语句通顺	91	20. 1
一般	摘要的完备性和概括性较好, 能够反映原文的主要内容 但可读性较差, 语句欠通顺	318	70. 2
差	摘要的完备性、概括性和可读性均较差 遗漏原文的重要内容, 细节描述过多, 语句不通顺	44	9. 7

实验中, 我们还对文档自动摘要的时间进行了记录. 结果表明, 文档摘要时间与文档长度基本成正比. 一篇大小为 10 KB 文档的自动摘要时间为 110 ms. 测试集中所有的 11440 篇文档的摘要时间总计为 1143 s, 平均摘要速度约为 10 篇/ 秒. 因此, 从速度上来看, 本文提出的自动摘要方法完全能够满足对大规模 Web 文档进行快速处理的要求.

5 结束语

本文根据 Web 环境对文档自动摘要技术在质量、速度、领域、用户等方面的要求, 提出了一种以篇章结构为指导、以机械摘要为基本手段的 Web 文档自动摘要方法. 理论分析和实验结果表明, 该方法具有不受领域限制、摘要速度快、摘要比例可调等优点. 同时, 由于采用了智能化技术对文档的主题层次进行分析而不是仅仅停留在文档的字面表层, 因而能够在一定程度上避免简单机械摘要方法在完备性等方面的不足, 提高摘要质量. 在今后进一步的工作中, 我们将针对摘要可读性较差这一缺陷, 重点研究如何在不影响处理速度的前提下, 适当地利用一些自然语言理解和生成技术来提高摘要质量, 尤其是改善摘要的可读性.

参 考 文 献

1 王继成, 张福炎等. Web 信息检索研究进展. 计算机研究与发展, 2001, 38(2): 187~ 193
(Wang Jicheng, Zhang Fuyan *et al*. State of the art of information retrieval on the web. Journal of Computer Research and Development(in Chinese), 2001, 38(2): 187~ 193)

2 J Kupiec. J Pedersen *et al*. A trainable document summarizer. In: Proc of the 18th Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR' 95). Seattle, Washington, USA: ACM Press, 1995. 68~ 73

3 王文欣, 黄萱菁等. 基于统计方法的汉语自动文摘系统研究. 计算机应用与软件, 2000, 17(9): 28~ 33
(Wang Wenxin, Huang Xuanjing *et al*. Research on automatic

summarization system of Chinese documents based on statistic methods. Journal of Computer Application and Software(in Chinese), 2000, 17(9): 28~ 33)

4 王永成等. 中文信息处理技术及其基础. 上海: 上海交通大学出版社, 1991
(Wang Yongcheng *et al*. The Fundamentals of Chinese Information Processing(in Chinese). Shanghai: Shanghai Jiaotong University Press, 1991)

5 吴岩, 刘挺等. 中文自动文摘原理与方法探索. 中文信息学报, 1998, 12(2): 8~ 16
(Wu Yan, Liu Ting *et al*. The fundamentals and methods of automatic summarization for Chinese documents. Journal of Chinese Information(in Chinese), 1998, 12(2): 8~ 16)

6 孙春葵, 李蕾等. 基于知识的文本摘要系统研究与实现. 计算机研究与发展, 2000, 37(7): 874~ 881
(Sun Chunkui, Li Lei *et al*. Research and implementation of knowledge-based text summarization systems. Journal of Computer Research and Development(in Chinese), 2000, 37(7): 874~ 881)

7 张波, 王继成等. Web 文档清洗技术研究. 计算机科学, 2002, 29(6): 52~ 54
(Zhang Bo, Wang Jicheng *et al*. Research on web document cleaning. Journal of Computer Science(in Chinese), 2002, 29(6): 52~ 54)

8 R Brandow, K Mitze, L F Rau. Automatic condensation of electronic publication by sentence selection. Information Processing and Management, 1995, 34(5): 575~ 685

王继成　男, 1973 年生, 博士, 副教授, 主要研究方向为 Web 信息检索与挖掘、中文信息处理

武港山　男, 1967 年生, 博士, 副教授, 主要研究方向为 Web 信息检索、多媒体信息处理

周源远　男, 1980 年生, 硕士研究生, 主要研究方向为 Web 信息检索、中文信息处理

张福炎　男, 1939 年生, 教授, 博士生导师, 主要研究方向为数字化图书馆、多媒体技术