基于用户查询的中文自动文摘研究

蒋效宇[1,2],樊孝忠[1],陈 康[1]
JIANG Xiao-yu[1,2],FAN Xiao-zhong[1],CHEN Kang[1]

1.北京理工大学 计算机科学技术学院,北京 100081
2.北京服装学院 商学院,北京 100029
1.School of Computer Science and Technology,Beijing Institute of Technology,Beijing 100081,China
2.Business School,Beijing Institute Clothing Institute,Beijing 100029,China
E-mail:jiangxiaoyu@bit.edu.cn

JIANG Xiao-yu,FAN Xiao-zhong,CHEN Kang.Research on Chinese automatic summarization based on user-query. Computer Engineering and Applications,2008,44(5):48-50.

Abstract: Summarization of text documents is increasingly important with the amount of data available on the Internet.The large majority of current approaches create query-independent summaries.This paper presents a new summary technology which adds the sentence weight based on user-query into the sentence importance.The experimental results show that this method can improve the speed and accuracy of searching information.

Key words: automatic summarization;sentence significance;word segmentation

摘 要:随着网络信息日益增多,文本摘要变得越来越重要。大多数现有的文摘方法采用的是独立于查询的方法来生成文摘。论文提出了一种将基于查询条件的句子权值计算融入句子重要度计算的文摘技术,实验结果表明该方法生成的文摘能有效提高用户搜索信息的速度并提高准确性。

关键词:自动文摘;句子重要度;分词
文章编号:1002-8331(2008)05-0048-03 **文献标识码:**A **中图分类号:**TP391

1 引言

随着互联网的普及和信息获取途径的增加,每天都有不断涌现的海量信息,为了方便用户从这些海量信息中快速准确地获取信息,提供高质量的文档摘要变得越来越重要。通过阅读文摘而不是全文能极大地加快信息过滤速度,帮助用户确定是否需要详读全文。

自动摘要[1]是指利用计算机对文档内容进行处理,从中选出最能代表文章主旨的语句,经过重组修饰以简洁的形式表达出来。传统的自动文摘的方法主要有基于统计的机械式摘要和基于语法语义分析的理解式摘要。机械式文摘在技术上易于实现,应用领域广泛;但生成的文摘不连贯、不简洁、内容不全面。基于语义理解方法生成的文摘可读性好,但由于自然语言处理技术至今还不成熟,若想获得高质量的摘要,必须将待处理的语料限制在某个领域内,而且构建相应的领域知识需要投入大量的精力,因此系统难以移植。

目前,大多数搜索引擎只是简单截取文档的前几行或者将含有查询关键词的语句抽取出来作为摘要返回给用户,内容的可靠性和准确性不高,用户经常需要查看全文来判断该文档是否为所需信息。为此,在中科院计算所汉语词法分析系统ICTCLAS对文档分词标注的基础上,采用识别散串的策略进行高频未登录词和专业术语的识别,在句子权值计算的基础上提出了将查询条件与文档各语句相似度融入句子重要度计算的文摘技术。这样既可以将文档的主要内容显示出来,又可以将含有查询关键词的语句显示出来,更有利于用户做出判断。

2 预处理

2.1 网页清洗[2]

与纯文本文档相比,Web网页是使用了HTML标记的一种半结构化数据,为了增强页面的显示效果或出于某种商业目的,网页作者通常在文档中插入大量的非文本信息(例如:图片、动画等链接信息),而这些非文本对文档主题内容的贡献不大,反而会影响文摘的性能和准确性。同时,HTML标记能够提供一些对自动文摘有用的辅助信息(例如:<H1><Title>等)。因此,在去除噪声数据的同时,也要充分挖掘网页的结构信息对文本处理的价值。

2.2 分词

本文采用的分词算法是中科院计算所的ICTCLAS。首先对文本进行分词,在实验过程中,发现分词后的文本中有很多散串,而这些连续散串在同一文本中出现频率比较高,往往是一

基金项目:教育部高等学校博士学科点专项科研基金(No.20050007023)。
作者简介:蒋效宇(1979-),男,博士生;樊孝忠(1948-),男,教授,博士生导师。
收稿日期:2007-04-16 **修回日期:**2007-11-05

些未登录词或是一些专业术语,因此采用了识别散字的策略进行高频未登录词和专业术语的识别[3]。散串的定义为:文本经过分词后,在文本中出现连续的若干个单字构成散串,散串中不包含"的"等单字高频虚词。具体识别算法为:

(1)分别将每个句子中的散串分解为两字或两字以上的组合。例如:散串"北 理 工"被分解为三个子串:"北理","理工"和"北理工"。

(2)统计每个子串的频度。如:文档中出现"北理工"2 次、"理工"4 次、"北理"3 次。

(3)对每个字串进行加权计算,加权公式:$L*C$,L 是子串长度,C 是子串出现的频度。

(4)权重高于某个阈值的子串作为未登录词,本文采用的阈值为 5。

(5)根据子串的长度和权重对子串降序排列。

(6)依次对原分词结果进行替换,例如:用"北理工"替换"北 理工"。

通过散字识别策略,能够提高未登录词的识别率,提高分词的准确率,更利于统计词频,从而更准确地找出文章的关键词。实验证明这种方法是很有效的。

3　文摘生成

3.1　特征提取及权重计算

对文本进行分词处理后,由于低频词(在文中只出现一次的词)和停用词所含有的信息量很小,故对已经切分的词语过滤这些词后所得的词称为文档的关键词[4],假设共 n 个,分别为 $T_1,T_2,\cdots,T_n$。然后对它们进行权重计算,特征权重计算唯一的准则是要最大限度的区分不同文档, 本文采用著名的 TF·IDF 来计算特征权重。TF (Term Frequency) 是特征频率,IDF(Inverse Document Frequency)是特征倒排文档频率,特征权重 $W_k(1\leqslant k\leqslant n)$定义如下:

$$W_k=\frac{\log(tf_k)\cdot\log(\frac{N}{n_k}+0.01)}{\sqrt{\sum_{i=1}^{n}\log^2(tf_k)\cdot\log^2(\frac{N}{n_k}+0.01)}} \tag{1}$$

其中 tf_k 为 t_k 在文档中出现的次数,N 表示文档中句子总数,n_k 表示含有该词的句子数。

3.2　句子表示

向量空间模型(Vector Space Model,VSM)是 Salton 于 20 世纪 60 年代提出的, 并成功地应用文本分类和信息检索等领域,向量空间模型是简单、有效的文本信息表示模型之一,它以特征项作为文本表示的基本单位, 所有的特征项构成特征项集。每个句子可以表示为一个向量:

$$V(S)=(t_1,W(t_1);\cdots;t_i,W(t_i);\cdots;t_n,W(t_n)) \tag{2}$$

t_i 表示第 i 个特征,$w(t_i)$表示特征 t_i 在句子中的权重。因为每个句子都是用特征集来表示,所以句子的向量表示简化可以为:

$$V(S)=(W(t_1),\cdots,W(t_i),\cdots,W(t_n)) \tag{3}$$

3.3　句子重要度

大多数文摘方法都是针对全文进行文摘,即利用句子中含有关键词权重、句子所在位置和是否含有提示语等因素计算句子权重,然后按照句子的权值大小来选择文摘,而没有考虑到针对某一个或几个关键词对文档进行文摘。

3.3.1　句子权重计算

为了衡量句子的重要性, 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[5]:

(1)句子中包含的关键词的重要性:句子关键词权重之和则说明句子的重要度越大,为了消除句子长度的影响,应该将关键词权重之和除以句子所含的关键词总数,得到句子的平均权重。

(2)句子在文档中的出现位置:处于篇首、篇尾、段首和段尾等位置的句子通常比其他位置的重要度要高。

(3)句子中是否包含有提示语:例如:"综上所述"、"总而言之"等,如果包含,那么句子往往是对文档的主题内容进行了概括,因此该句子重要性相对较高。

(4)句子是否为标题句:标题通常是对下文的一个概括,无论在信息量还是重要性都比较高。

(5)句子是否为"例如"、"比如"等细节性词语开头,这些词语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。

综合上述五个因素,句子的权重计算 $W(S_k)(1\leqslant k\leqslant m)$定义如下:

$$W(S_k)=\lambda_1\times\frac{\sum_{i=1}^{n}w(t_i)}{Len}+\lambda_2\times W_{pos}+\lambda_3\times W_{hint}+\lambda_4\times W_{title}+\lambda_5\times W_{ex} \tag{4}$$

其中,$\sum_{i=1}^{n}w(t_i)$是句子 S_k 中关键词的权值和;Len 是 S_k 中包含关键词总数;W_{pos} 表示位置权值;W_{hint} 表示提示语权值;W_{title} 表示标题句权值;W_{ex};λ 是加权系数,$\lambda_1\geqslant 0,\lambda_2\geqslant 0,\lambda_3\geqslant 0,\lambda_4\geqslant 0,\lambda_5\leqslant 0$,$\sum_{i=1}^{5}\lambda_i=1$。

3.3.2　基于查询条件的句子权值计算

如果返回给用户的文摘的句子中包含用户查询关键词的话, 这更有利于用户判断搜索引擎返回的文档是否是他们所需,基于这样的假设,在计算句子重要度时增加了"查询权值",所谓查询权值[6]是每个句子针对查询关键字统计出来的权值,具体查询权值计算定义如下:

$$W(q,S_k)=\frac{C^2(q,S_k)}{Len(q)} \tag{5}$$

其中,$C(q,S_k)$表示句子 S_k 中包含查询条件 q 中的关键词的总数;$Len(q)$表示查询条件中关键字个数。

3.3.3　句子重要度计算

句子重要度由两部分组成:一个是大多数基于统计的文摘方法得到的句子权值,另一个是每个句子的查询权值,句子 S_k 重要度的定义如下:

$$I(S_k)=\mu\times W(S_k)+(1-\mu)\times W(q,S_k) \tag{6}$$

其中:$W(S_k)$是句子 S_k 权重,由式(4)计算所得;$W(q,S_k)$是基于查询条件的句子权重,由式(5)计算所得;μ 是加权系数,当 $\mu=1$ 时就是目前大多数文摘采用未考虑查询的方法生成的文摘,当 $\mu=0$ 时就是现在大多数搜索引擎采用的将出现关键词的语句提取出来作为文摘。

3.4　粗文摘生成

各句重要度计算出来后, 依据其重要度将各句降序排列。摘要构造方法是依次将重要度最大的句子抽取出来,直到摘要达到指定长度,摘要长度一般由用户确定,通常是原文的 5%~25%,接着将这些从原来抽取的文摘句重新组织,按其在原文中的顺序排列。这样文档的粗文摘就生成了。

3.5　减少文摘冗余度

生成的粗文摘中往往会出现文摘冗余度大的问题,因为抽取的文摘句子都是很重要的句子,但文档中经常会有关于某一

方面重复描述的一些句子,所以,要通过句子相似度计算减少文摘中这样的句子,但是出现查询关键词的语句不予考虑。

用 $Sim(S_i,S_j)$来表示句子之间的相似度,又记向量空间的原点为 O,则利用向量间的夹角余弦公式表示为:

$$Sim(S_i,S_j)=\cos\angle S_iOS_j=\frac{\sum_{k=1}^{n}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{n}W_{ik}^2)\times(\sum_{k=1}^{n}W_{jk}^2)}} \quad (7)$$

式中:S_i,S_j 为 VSM 表示的句子向量,n 为向量维数,W_{ik} 和 W_{jk} 分别为第 k 个特征在两个句子中的权值。设定一个阈值,相似度高于该阈值的两个句子认为是意义重复的,只保留重要度高的一句,丢弃另一句。

4 实验结果与评价

为了检验基于查询条件的句子权值计算融入句子重要度计算的文摘技术是否能够节省用户搜索信息时间和提高检索的准确率,从百度上搜集了 5 个相近主题,每个主题分别 5 篇文章,平均每篇文章 26.2 句,在不同压缩比下分别使用抽取包含关键词的语句法(方法 1)、基于统计的文摘方法(方法 2)和使用基于查询条件的句子权值计算融入句子重要度计算的文摘技术(方法 3)在不浏览整个文档情况下,将每篇文章进行归类。实验结果如图 1 所示。

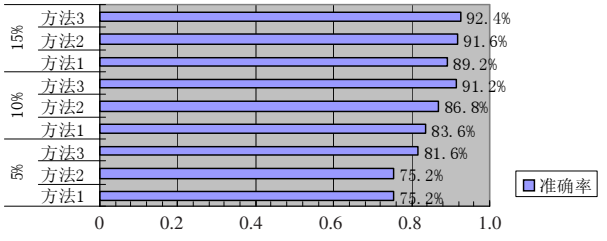


图 1 实验结果

从实验结果可以看出,方法 3 在判断查询条件与文档相关性的准确性和所用时间均优于方法 1 和方法 2,虽然测试所用语料范围较小,数量较少,但还是能说明使用基于查询条件的句子权值计算融入句子重要度计算的文摘方法能够提高在使用搜索引擎判断返回文档与查询条件相关性的准确率并节省时间。同时句子的权值计算可以提前完成,搜索时只需计算基于查询条件的句子权值,不会降低搜索引擎返回结果的速度。

5 结论

文章提出了将基于查询条件的句子权值计算融入句子重要度计算的文摘技术,通过此方法能够提高用户搜索信息的速度和准确率。当然,系统还存在一些不足之处:文摘缺乏句间的连贯性和指代词悬挂等问题,这些在今后的工作中进一步改善。本文只考虑了 A and B 类型的查询条件的文摘生成,A and(.not.B)类型的查询条件的文摘生成技术是下一步研究重点。

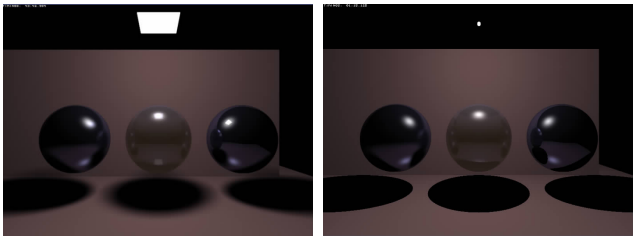
参考文献:

[1] Luhn H P.The automatic creation of literature abstract[J].IBM Journal of Research and Development,1958,2(2):159-165.
[2] 陈志敏,沈洁.基于主题划分的网页自动摘要[J].计算机应用,2006,26(3):641.
[3] 于海滨,秦兵.中文单文档自动文摘技术研究[D].哈尔滨工业大学,2005:8-9.
[4] 傅间莲,陈群秀.基于规则和统计的中文自动文摘系统[J].中文信息学报,2006,20(5):10-16.
[5] 王继成,武港山.一种篇章结构指导的中文 Web 文档自动摘要方法[J].计算机研究与发展,2003,40(3).
[6] Tombros A,Sanderson M.Advantages of query biased summaries in information retrieval [C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,1998.

(上接 47 页)

3 算法的结果及分析

图 4 是本算法的效果示意图,(a)图为有软阴影的效果,(b)图为没有软阴影的效果,从图中可以明显地看出,(a)图相对(b)图,场景中阴影边界更加柔和过渡,阴影更具真实感。程序实现的平台为 Inter Pentium4 CPU 2.4 GHz, 显卡为 GeForce4 MX440。



(a)面光源产生的软阴影效果图 (b)点光源生成的硬阴影效果图

图 4 基于光线跟踪生成的软阴影与硬阴影效果图对比

4 结论以及下一步工作

论文提出的基于光线跟踪的软阴影算法与其他软边界阴影算法相比,生成的阴影更加准确,克服了本影区域过估计问题,并且解决了阴影映射算法产生的锯齿现象的问题。虽然提出的算法可以在当前的图形硬件上实现,但是由于该算法的计算量相当大,因此还难以用于大型的虚拟现实系统当中。下一步的工作就是针对新推出的图形硬件,对算法进行优化,使之能适用于大型的虚拟现实系统。

参考文献:

[1] Woo A,Poulin P,Fournier A.A survey of shadow algorithms[J].IEEE Computer Graphics and Applications,1990,10(6):13-32.
[2] Brabec S,Seidel H P.Single sample soft shadows using depth maps[C]//Graphics Interface,2002.
[3] Gray K.DirectX 9 programmable graphics pipeline[M].Washington:Microsoft Press,2003:67-69.
[4] Ying Zhengming,Tang Min,Dong Jinxiang.Soft shadow maps for area light by area approximation[C]//10th Pacific Conference on Computer Graphics and Applications,IEEE,2002:442-443.
[5] Herf M.Efficient generation of soft shadow textures,Technical Report CMU-CS-97-138[R].Carnegie Mellon University,1997.
[6] Agrawala M,Ramamoorthi R,Heirich A,et al.Efficient image-based methods for rendering soft shadows[C]//Computer Graphics(SIGGRAPH 2000),Annual Conference Series,ACM SIGGRAPH,2000:375-384.
[7] 彭群生,鲍虎军,金小刚.计算机真实感图形的算法基础[M].北京:科学出版社,2002.