关键词密度分布法在偏重摘要中的应用研究

闫英杰 林鸿飞 杨志豪

（大连理工大学计算机科学与工程系，大连，116024）；

摘　要：偏重摘要系统的研究和实现是实现智能化信息服务的基础，由于偏重摘要不区分单文档与多文档，根据用户需求为用户提供结果，受到了越来越多的研究关注。本文实现了用关键词密度分布方法获取偏重摘要的一个实验系统。研究了基于关键词密度分布方法生成摘要句的策略，并通过两种不同的实验，对系统进行了评测。最后对实验结果进行了讨论。经实验证明，本系统能够满足用户的查询要求，在基于检索任务和基于问答任务的两项不同的评测中均得到了较好的实验结果。

关键词：密度分布，偏重摘要，自然语言处理

Study and Implementation of Biased Summarization by using Density Distribution of Keywords

YAN Ying-Jie,Lin Hong-Fei,YANG Zhihao

(Department of Computer Science and Engineering, Dalian University of Technology,Dalian，116024,)

Abstract: The study and implementation of biased summarizes system is the basis of implement of intelligent information service. For the reason of that biased summaries does not differentiate between single document and multi-documents, it is noticed by more and more researchers. The paper presents an approach to generate summarizations based on the density distribution, and it also discusses the conceptual expansion and the strategy of the use of density distribution of keywords. Based on the IR experiment and QA experiment, their results show that the summarization system can satisfy user's query requirement and have the better results on two different evaluations.

Keywords: Density Distribution; Biased Summarize; Natural Language Processing;

1　引言

文本摘要最早从 Luhn 的研究工作开始[1]，到今天已有 50 年左右的时间了。对于偏重摘要的研究最近几年才开始见诸于文献。偏重摘要是与一般摘要或通用摘要相对的，它根据用户的兴趣和需要提供相应的有侧重点的摘要，而不是对源文档主题内容的概括浓缩。 提供的摘要是反映读者的兴趣和需求的，而不是反映作者的观点。这就决定了偏重摘要的实现与单文档摘要或多文档摘要的不同。

研究和实现偏重摘要的生成至少有两个方面的意义：一是为用户提供个性化的查询信息服务。在当今网络信息时代，人们为能找到自己有用的信息耗费了许多时间和精力，而偏重摘要的实现有助于人们快速的从网络上、或本地的相关文档中找到自己所需要的信息。二是能为实现智能化的搜索引擎提供基础方法。当今的搜索引擎技术为人们提供了快速找到相关信息的方法，但在定位信息、实现智能化搜索方面还有很大的发展空间，搜索引擎检索出的信息往往鱼龙混杂，还需要经手工进行二次筛选，实现偏重摘要正可以弥补搜索引擎的这一不足。

基金资助：国家自然科学基金（60373095）。

作者简介：闫英杰，男，硕士研究生,主要研究方向为自动文本摘要、中文信息处理；林鸿飞，男，博士，教授，主要研究方向为自然语言理解和文本挖掘，E-mail: hflin@dlut.edu.cn。

当前，对于文本摘要的研究热点集中在多文档摘要方面。因为多文档摘要技术相比较单文档摘要而言，有着更多更复杂的技术难点。而偏重摘要同时处理多文档和单文档，以多文档摘要技术为依托，比较而言就其难度要更大一些。国外的一些学者，近年来对偏重摘要方面有了一些研究，发表了多篇相关论文：如 V. Plachouras 等人研究了运用偏重摘要提高网页中查询精度的方法[2]，S. Sweeney 等人研究了运用偏重摘要为 WAP 手机用户提供信息的意义[3]，Tsutomu HIRAO 等人提出了一个用于提高问答任务结果的偏重摘要方法[4]。

从目前的研究来看，还没有实现一个真正的偏重摘要系统，能为用户直接提供个性化的信息服务。本文意在尝试建立一个适应用户个性化查询需求的偏重摘要系统。本文首先介绍了有关偏重摘要的概述。第 2 部分为密度分布算法的相关论述及偏重摘要系统的结构分析。第 3 部分为实验及实验结果分析。最后的第 4 部分，总结了本文工作，并对应用密度分布算法生成偏重摘要的研究做了一些探讨。

2 关键词密度分布算法

偏重摘要系统与一般的摘要系统有所不同，它需要与用户交互，以获取查询输入，形成偏重。本文采用了关键词密度分布方法来获取偏重摘要。

关键词密度分布算法在文献[5]中被用来实现建立文档间的图表与相关文字内容的链接和寻找文档集中的相关文档。本文在其研究的基础上，实现了运用关键字密度分布算法生成偏重摘要的一个实验系统。生成偏重摘要有不同的其它方法，文献[6]中提到了标题法、位置法、词频法、查询偏重法等方法。关键词密度分布算法是综合了词频法和查询偏重法的一种方法。其理论依据是建立文档中与用户查询词匹配最集中的文档部分一定是用户查询所需求的内容这一假设上的。具体算法过程如下：

第一步：预处理。将原文档分词后，将文档表示为单词序列，即文档 t 中的一个单词表示为：

$$\text{word} = \text{document} \quad (L) \quad (0 \le L \le L_t) \tag{1}$$

这里 L 是一个位置（单词从文章开头算起的位置），而 Lt 是文档 t 中总共的单词数。

同时，记录每个单词所在的句子序列。

$$\text{word} = \text{sentence} \quad (L_s) \quad (0 \le L_s \le L_{s_t}) \tag{2}$$

第二步：抽取关键词，统计词频。统计文档中名词、动词、形容词、副词及名词短语、常用词的词频 n_k。将其作为关键词集合 K。

第三步：计算权重 $w(k)$。在第二步计算出词频的基础上，进一步计算出单词的权重值，计算公式如下：

$$w(k) = \log n_k + 1 \qquad (w(k) = 0 \quad \text{当 } n_k = 0 \text{时}) \tag{3}$$

第四步：求取用户查询词与相关结果文档的匹配范围，记录匹配值 $b_t(l)$。公式如下：

$$b_t(l) = \begin{cases} w(k) & \text{当 } a_t(l) = k \in K \text{时} \\ 0 & \text{其它情况时} \end{cases} \tag{4}$$

第五步：计算各个单词的密度值 $d_t(l)$，计算公式为：

$$d_t(l) = \sum_{i=-\frac{W}{2}}^{\frac{W}{2}} f(i) * b_t(l-i) \tag{5}$$

其中 $f(i)$为一个窗口函数，这里采用的是汉宁窗口，在文献[5]中对矩形窗口、三角形窗口、汉宁窗口等三种不同的窗口函数进行了比较实验，以汉宁窗口的平滑结果最为理想，所以这里直接选用汉宁窗口函数作为平滑窗口函数。窗口函数对以关键词为中心的窗口大小为 W 的范围内的文档关键词进行密度平，得到平滑后的密度分布值。其窗口大小的取值为 500。汉宁窗口函数公式如下：

$$f_h(i)=\begin{cases}\frac{1}{2}(1+\cos\ 2\pi\frac{i}{w}) & \text{当 } |i|\le\frac{w}{2}\text{时}\\ 0 & \text{其它情况时}\end{cases} \tag{6}$$

第六步：根据阈值，划分密度，从而最后得到符合用户需求的候选文摘句。依据公式(7)，得到符合条件的密度分布值$d_t(l)$。

$$\hat{d}_t(l)=\frac{d_t(l)}{\max_{t,l} d_t(l)}\ge T \tag{7}$$

下图 1 是一个由汉宁窗口函数平滑后的密度分布示例。

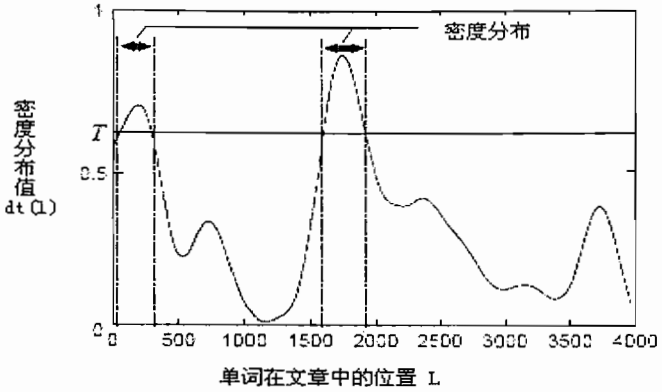

图 1 密度分布示例

Fig 1 An example of a density distribution

其中，T 为阈值，实验取值为 0.6。然后将这些关键词所在的句子序列从原文档中提取出来，成为最后的文摘句。对文摘句的排序方法是，以密度值大小为序，在密度值相同时，以句子在原文档中的顺序为摘要中的顺序。这一排序方法兼顾了密度值的大小和句子的原始顺序。最后根据文摘比率的要求，输出最后的文摘句。在控制文摘大小方面，本文根据实验的不同采用了不同的文摘比率。在基于检索的实验一中采用了 30%的文摘比率；在基于问答任务的实验二中采用了 10%的文摘比率。

关键词密度分布算法依据密度值来简单地决定一个抽取部分，实现起来相对简单，实现关键词密度分布算法的偏重摘要系统的一般结构如下图 2 所示。

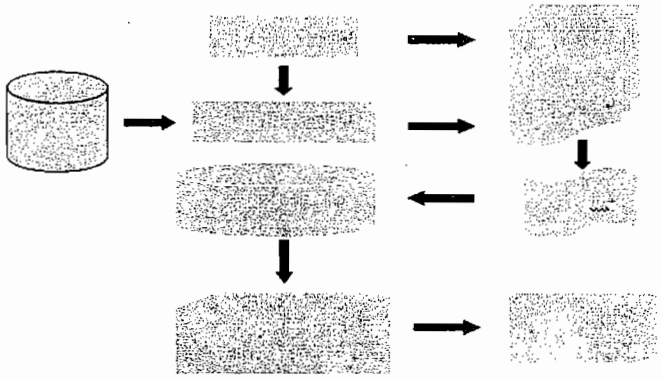

图 2 偏重摘要的一般系统结构

Fig.2 General architecture of Biased Summaries

3 实验结果及其分析

由于偏重摘要与通用摘要的不同，对其评测不能采用通用摘要的方法。而且自动文摘的评测本身就是一个难点问题。在检验的实验中，我们采用的两种不同的方法来对关键字密度分布方法生成偏重摘要进行了评测。一是

基于检索的评测。即以从网上下载的 30 篇政府工作报告为实验语料，根据个人的主观要求输入查询词，进行了文摘生成评测，检验生成了包含查询内容的文章数。二是基于问答任务的评测。即以哈工大的问答任务的问题集为基础，选用其中的 52 个问题句，将以问题句为搜索词由百度搜索引擎搜到的前三篇文章做为实验语料，再以问题句为查询词来求取偏重摘要，检验结果中包含了正确答案的文章比率。因为这两个实验中对每篇文章都生成摘要，在查全率方面区别不大，实验只以正确率作为评测指标。第一个实验评测结果是依赖于人的主观判断，第二个实验基于客观的评测。实验的结果显示，绝大多数文章都能按照用户的偏重要求生成摘要。下面是具体的实验及结果分析。

实验一：基于检索的主观评测

我们从网上下载了 30 篇政府工作报告，主要是 2006 年全国部分省市和中央的政府工作报告。实验的目的中从这些报告中抽取出用户感兴趣的内容。选取这一语料，目的在于便于进行横向比较。输入用户的偏重内容分别是：工业发展计划、农业发展规划、国企改革措施、科教发展计划和政府职能改进等。实验的结果如表：

表 1 实验一的结果

Tab 1 Result of Experiments 1

抽取内容	工业	农业	国企改革	科教发展	政府职能
抽取正确	25 篇	22 篇	15 篇	24 篇	23 篇
正确率	83.3%	73.3%	50%	80%	76%

所有的语料都生成了最后的摘要，其中得到符合用户偏重要求的定为正确的篇数。正确率的计算为：正确率=抽取正确数/总的文章数。

由于报告文体的特殊性，从中抽取出相关片断并不是很难，但实验结果并没有达到 100%正确。在多次实验后，找到一些原因：一是原文中没有包含相关的内容。如在查询国企改革方面的有关内容时，青海省的政府工作报告中就没有提到国企改革方面的内容。二是虽然包含了相关的内容，但是与查询关键词并不完全匹配，以至于未能正确抽取为文摘句。如在查询有关科教发展和国企改革方面的内容时，由于几个省市的政府工作报告对这一方面的内容涉及很少，所以未能完全抽取正确。

实验二：基于问答的客观评测

实验以哈工大的问答实验语料问题集为基础，选用了其中的 52 个问题。首先通过问题关键词用百度搜索从网上下载了相关的文档，（百度找到的前三篇文章），然后，经网页去噪处理，将相关的文件转变为文本格式，以这些文本为实验语料。以问题句为查询输入的关键词，生成了最后的偏重摘要。实验语料根据问题的不同共分为 6 个大类，分别是概念类、时间类、地点类、人物类、事件类及其他类。对这 52 个问题的实验结果如下表所示：

表 2 实验二的结果

Tab 2 Result of Experiments 2

	概念类	人物类	地点类	时间类	事件类	其他类
问题总个数	14	14	9	7	3	5
包含答案数	11	12	8	6	3	3
正确率	78.6%	85.7%	88.9%	85.7%	100%	80%

这里正确率的计算同实验一：正确率=抽取出的包含有正确答案的文章数/总的文章数。实验结果显示：偏重摘要系统能够为问答任务更快、更准确的找到答案。在未能包含正确答案的文摘中，经分析找到以下原因：一是百度搜索引擎搜到的前三篇文章中没有包含正确答案。如对人物类问题“1997 年英华早逝的英国王妃是谁？”，百度搜到的前三篇文章是：【分享】[9/08]电影更新下载 霏凡论坛 - powe...，【分享】[9/08]电影更新下载 口- 影视交流版 ...，[9/08]电影更新下载 Xproe 论坛 资料下载 学习交..。三篇文章都没有包含问题句的答案，而且第一篇和第二篇是重复的。二是在分词过程中，没有识别出专有名词，导致最后关键词匹配不准确，如对概念类问题“贝斤是一种什么遗迹？”对专有名词“贝斤”没有识别出来，导致最后没有将正确答案抽取出来；三是由于查询关键词太少，不能定位包含正确答案的文章片断。如时间类问题中“花生在几月份种植比较合适”在用原问题句进行摘要生成时，没有生成包含正确答案的摘要，而用“花生、栽培、时间”做关键词是就生成了包含正确答案的摘要。

答案的摘要。

针对以上问题，考虑一些可行的解决方法：一是扩大语料集。本文的实验采用了搜索引擎的前三篇文章，看来是不够的，还需要扩大范围，进一步进行实验。二是有针对性的进行查询扩展。实验表明：单纯依靠概念库或同义词词林，还不能获得满意的结果，查询扩展需要有一定的针对性。建立不同领域的专门的知识库可以解决这个问题。三是依据用户意向选择查询关键词。偏重摘要从用户的查询开始，查询关键词的选用就尤为重要。如能依据用户的背景知识进行查询关键词的选用，必定能够事半功倍。四是进一步提高中文分词的精度，也将在一定程度上提高偏重摘要的精度。

4 结束语

本文通过实验检验了应用密度分布算法生成偏重摘要的可行性，实验结果显示密度分布算法可以生成满足用户偏重要求的文摘。通过对实验结果的分析，可以得到以下一些结论：

一是依靠关键词密度分布方法综合了文章的词频信息和用户的查询要求，在一定程度上还能反映出文章的主旨内容。如在查询时间类问题“父亲节是哪一天”时，在文摘比率为 30%时，可以获得正确答案。而文摘率为 10%时，得到的只是文章的主旨内容，即有关父亲节的风俗。

二是依据密度分布值求取文摘句，控制文摘大小方面非常灵活，只需简单地调整密度的阈值就可以得到不同长度的候选文摘句。在文献[5]中，对不同的窗口大小 W、不同的阈值 T 求取文摘进行了实验，其对比结果以 W=500 和 T=0.6 最为理想。

三是避免了使用概念扩展产生的副作用，而采用同义词词林进行关键词扩展。在文献[7]中提出了概念库知识进行偏重摘要生成的方法，但此方法易造成查询结果的泛滥和运行时的时间花销增大。由于没有获得概念库，未能做一个对比实验。但就目前的系统在运行时间开销方面已经较大了。如处理 80K 左右的政府工作报告时，运行时间过长，超到了 20 分钟。

总得来说，本文实现了以密度分布算法为基础的能够满足用户偏重摘要需求的自动摘要系统。但仍然需要在以下方面进行改进：

一是查询关键词的扩展方面。在采用不同的查询关键词得到不同的文摘，如对于“花生在几月份种植比较合适”这一问题，哈工大的同义词词林(扩展版)中没有将“种植”和“栽培”列为同义词，以致于不能通过“种植”查询出正确答案。

二是在对查询词分词方面仍需进一步提高精度。由于分词程序中“贝丘”这个专有名词未能识别，以致于生成的结果文摘中没有包含正确答案。

三是对窗口函数的窗口大小和密度分布阈值调整方面需要有一个测试，以获取更高的正确率。

四是增加对文本内容进行概念理解方面的内容。本文实现的系统单纯地以机械统计方法获取文摘，在文摘的连贯性方面和对多文档文摘重复句过滤方面缺乏处理方法，需要借助于文本理解的手段进行处理。

参考文献：

[1] H. P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 1958, 2(2): 159~165

[2] V. Plachouras and I. Ounis. Query-Biased Combination of Evidence on the Web. Workshop on Mathematical/Formal Methods in Information Retrieval, ACM SIGIR Conference, 2002

[3] S. Sweeney, F. Crestani, and A. Tombros. Mobile delivery of news using hierarchical query-biased summaries. In Proceedings of ACM SAC 2002

[4] Hirao, T., Sasaki, Y., Isozaki, H.: An extrinsic evaluation for question-biased text summarization on qa tasks. NAACL-2001 Workshop on Automatic Summarization(2001)

[5] Kise, K. Mizuno, H. Yamaguchi, M. Matsumoto, K. On the use of density distribution of keywords for automatedgeneration of hypertext links from arbitrary parts of documents.Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on

[6] Anastasios Tombros, Mark Sanderson，Advantages of Query Biased Summaries in Information Retrieval，In Proceedings of SIGIR-98

[7] 刘功申、胡佩华、岳奕、王永成，《偏重摘要技术及其实现》，第一届学生计算语言学研讨会(SWCL2002)。