基于主题区域发现的中文自动文摘研究*)

胡 珀[1] 何婷婷[1] 姬东鸿[2]

(华中师范大学计算机科学与技术系 武汉430079)[1] (新加坡国立信息通信研究院 新加坡119613)[2]

摘 要 自动文摘是自然语言处理领域的一项重要的研究课题。文中提出了一种基于主题区域发现的中文自动文摘的方法。该方法的特色在于:产生的文摘能在尽可能全面地覆盖全文多个主题的同时,显著地缩减自身的冗余,从而能有效地平衡两者之间的矛盾。通过采用K-medoids的聚类算法联同新的自定义目标函数的聚类分析方法,实现了段落自适应聚类下的文本潜在主题区域的发现及其在自动文摘领域的应用。此外,一种基于表达熵的新的评价因子被用来评价摘要的冗余。实验结果验证了该方法的可行性,有效性,是对中文自动文摘研究的一种有意义的探索。

关键词 自动文摘,主题区域发现,聚类分析,表达熵

A Study of Chinese Text Summarization Based on Thematic Area Discovery

HU Po[1] HE Ting-Ting[1] JI Dong-Hong[2]

(Department of Computer Science and Technology, Central China Normal University, Wuhan 430079)[1]

(Institute for Infocomm Research, Heng Mui Keng Terrace, 21 Singapore 119613)[2]

Abstract Automatic summarization is an important issue in Natural Language Processing. This paper has proposed a special method that creates text summary by discovering thematic areas from Chinese text. The specificity of the method is that the created summary can both cover as many as different themes and reduce its redundancy obviously at the same time. And the discovery of latent thematic areas under the adaptive clustering of passages is realized by adopting k-medoids clustering method as well as a novel clustering analysis method based on self-defined objective function. In addition, a novel parameter, which is known as representation entropy, is used for summarization redundancy evaluation. Experimental results indicate that this method is effective and efficient in the automatic summarization literature.

Keywords Automatic summarization, Thematic area discovery, Clustering analysis, Representation entropy

1 前言

随着信息过载时代的到来,人们在海量的信息面前开始变得束手无策。鉴于现有的信息检索技术在信息查询的查准率和查全率方面表现出的效果还差强人意,如何才能从众多检索结果,特别是文本检索结果当中有效地寻找到与用户的需求最相关的信息便成为了一个众所关注的问题。

自动文摘技术的重要之处就在于它能在一定程度上解决上述问题。具体表现在以下两个方面:其一,质量良好的文摘能在一定程度上取代原始文档的被检索地位,作为原始文档的一个替代品参与检索,有效地缩减了检索的时间;其二,质量良好的文摘还能用于检索结果的可视化,使得用户无需浏览原始的检索结果集合便能轻松地取舍检索信息,从而有效地节省了浏览时间,提高了相关信息的命中率。

本文提出了一种新的基于主题区域发现的中文自动文摘的研究方法。该方法共分三个阶段进行:首先,通过采用合适的聚类算法(如K-medoids算法[10])与合适的聚类分析方法(如自定义目标函数法、MDL法[11])寻找出文本中的不同主题所覆盖的段落区域,即主题区域;然后,在各个主题区域下,挑选出一个与该主题区域的语义相似性最大的句子作为主题代表句;最后,将选出的代表句按一定的次序组合成文摘。

为了验证该方法的有效性,将该方法连同传统的无主题区域检测的文摘方法分别作用于我们的实验样本,产生出两组不同的文摘集合,然后对集合进行比较。比较结果显示,基于主题区域发现的文摘方法所产生的文摘在自身的信息量(即摘要的主题覆盖程度)、冗余等考察因子上取得了更好的效果,在一定程度上克服了传统文摘方法的不足。

本文第2部分将简单地回顾目前国内外自动文摘研究的基本方法,指出其中一些方法的不足之处以及新方法的提出动机。第3部分,将详细介绍基于主题区域发现的自动文摘方法。第4部分,将介绍实验情况、结果和评价。最后得出结论并指出后续的工作方向。

2 相关工作及新方法的提出动机

关于自动文摘的研究起始于H. P. Luhn(Luhn1958)的工作。到目前为止,已有国内外众多学者参与了此项研究,并取得了不少成果。绝大多数的研究都致力于句子抽取型的文摘方法(即采用Shallower的方法)[1,3,4],而非句子产生型的文摘方法(即采用Deeper的方法)[9]。一方面,这是由理性的自然语言理解技术和知识工程技术的高度复杂性及其应用领域的严重受限性所造成;另一方面,这也与近年来统计学的研究方法、机器学习的方法以及模式识别的方法在自然语言处理的一系列领域所取得的不俗成绩分不开[8]。

句子抽取型的文摘方法按抽取方法的不同可大致分为有指导型和无指导型[3]。有指导型抽取方法的实现依赖于大量人工做的标准摘要,即俗称的"Gold Standards"来帮助训练和确定摘要统计学模型的特征参数。然而,由于人工摘要的置信度问题至今也未获得一个公认的一致说法,因而在一定程度

*)基金项目:中国国家语言文字应用委员会"十五"国家语委应用项目基金(ZDI105-43B);湖北省自然科学基金(2001ABB012)。胡 珀 硕士生,主要研究领域为自然语言处理;何婷婷 博士,教授,主要研究领域为自然语言处理、数据库;姬东鸿 博士,研究员,博士生导师,主要研究领域为自然语言处理。

上促使了研究人员对无指导型文摘方法的研究。无指导型的文摘方法，其最大优势在于无需人工摘要支持，仅从文章自身出发，利用统计学方法和启发式规则来确定句子的权值并选择句子，该方法还可进一步被细分为无篇章结构分析型和基于篇章结构分析型。前一种方法的通常做法是：先给全文的句子打分，然后挑选出权值最高的若干句子，最后按这些句子在原文中的语序关系依次输出它们构成文摘。但人们很快便发现采用这种方法产生的文摘不仅主题覆盖不全而且冗余大，它往往只能提取出文章中分布密度比较大的主题，而忽略了其它主题的存在。针对这一问题，南京大学提出了基于篇章结构分析的文摘方法[1]，通过文章中相邻段落的用词重叠统计来计算相邻段落间的语义距离，从而分析出文章的主题结构，最后从每个主题下抽取句子构成摘要。这种方法在处理篇章结构比较规范的文本时效果比较好，能有效地解决无篇章结构分析型文摘方法所出现的上述问题。然而，值得注意的是，当文章的写作风格比较自由，且主题分布灵活多样，即一个主题可能分布在不相邻的若干个段落中时，采用此方法的效果则较差。

为了能有效地处理目前普遍存在的大量文风自由、主题灵活的中文文本。在 Tadashi Nomoto 等研究人员思路[3]的启发下，尝试了一种基于主题区域发现的句子抽取型的文摘方法，由于文章中主题区域的发现是通过段落的自适应聚类得到的，因而它能一定程度上克服上述方法在处理主题灵活文本时的不足。

3 算法介绍

一篇文章通常包含若干个不同的主题。而一个好的文摘方法产生的文摘应该在尽可能全面地覆盖原文多个不同主题的同时，最大限度地缩减自身的冗余[4]。

本节将详细介绍我们所提出的文摘方法：即通过主题区域的发现来尽可能全面地挑选出最具有各个主题代表性，同时彼此间语义差别大、冗余小的句子集合构成摘要。该方法主要包含以下三个阶段：

阶段1：通过段落聚类与聚类分析寻找出文本中的不同主题区域。

阶段2：从各个主题区域中挑选出一个最合适的句子做为主题代表句。

阶段3：将代表句按一定的要求组合成最终的文摘。

3.1 阶段1：主题区域的发现

主题区域发现的具体过程如图1所示。

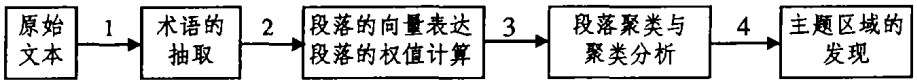

图1 主题区域发现的具体过程(共分4个步骤)

以下是主题区域发现的四个步骤：

Step 1：从原始文本中抽取出术语

与传统的中文自动文摘研究一贯所基于的分词方法不同，我们对原文进行预处理时并没有做通常的分词操作，而是利用了基于开放式语料的汉语术语的抽取方法[2]从原文中抽取出了适量的术语，然后通过术语这种元数据来刻画文档的内容。

采用该方法的好处至少有以下两点：

其一，分词问题一直是中文信息处理领域一个未得到很好解决的问题。而且在深层 NLP 技术和知识工程技术尚难取得实质性进展的今天，分词的质量很难有大的提高。因此，术语抽取技术能在一定程度上缓解中文分词技术的压力。

其二，术语和词虽然同属于语言符号范畴，但术语相对于词而言，表达的语义更加专业、完整，能承载更多的语义内容。术语抽取技术的最大优势就在于它无需固定词典的支持，只需通过对大量真实语料的不断更新统计便能动态地建立和更新术语库，通过不断的修正抽取参数便能提高抽取质量，因此具有相当广阔的 NLP 的实际应用前景。

Step 2：段落的向量表达与段落的权值计算

向量空间模型(VSM)的伟大之处就在于：它很好地解决了将非结构化的文档结构化的问题，从而使得利用已有的成熟的数学工具对大规模真实文本进行处理成为了可能。这里，我们把从文档中抽取出来的所有术语当作文档的特征，对术语的某种统计加权值当做特征的权值。从而可以建立起段落的术语向量空间模型，即将每个段落 $P_i(i:1\sim M$，M 是文档中的段落总数)映射成术语的权值向量 VP_i，$VP_i=(WP_{i1}, WP_{i2}, \cdots, WP_{iN})$；其中，$N$ 是文档的特征总数(即术语总数)，WP_{ij} 表示第 j 个术语在第 i 个段落中的权值。关于 WP_{ij} 的计算方法有很多，如 tf 法、tf * id 方法、mutual information 法[5]等。本文采用了如下的计算方法[4]：

$$WP_{ij}=\log(1+TF(T_{ij}))*\log(M/M_j)$$

其中，$TF(T_{ij})$ 是第 j 个术语在第 i 个段落中出现的频次，M/M_j 是术语 j 的倒排段落频率，它是为了克服单纯频次统计在描述术语的段落重要性时的不足而添加的，这里，M_j 表示包含术语 j 的段落个数。

相应的，在定义了 WP_{ij} 的基础上还可以进一步定义段落 i 的权值 $W(P_i)$ 的计算方法：

$$W(P_i)=\frac{\sum_{j=1}^{n} WP_{ij}}{n}$$

上述公式清楚的表明：我们将段落中不同术语的权值总和与段落中包含的不同术语的总数之比值定义成了该段落的权值。公式中，n 表示第 i 个段落 P_i 中包含的不同术语的总数。

Step 3：段落聚类与聚类分析

1) 段落聚类

现有的聚类算法从大体上可以分为两类[5]：层次聚类算法(如 Agglomerative 算法、Divisive 算法等)和扁平分割聚类算法(如 K-means、K-medoids 等)。

层次聚类算法的计算复杂度为 $O(n^2\mathrm{Log}(n))$，n 为参与聚类的元素个数，它通常高于扁平分割聚类算法的复杂度，如 K-means 算法的复杂度是 n 的线性关系，因而出于对算法效率的考虑，我们决定选用后一种算法对段落聚类。

K-means 聚类算法在很多情况下都是一种不错的选择，因为它简单易行，且效果不错。然而，我们发现在采用 K-means 算法进行聚类的过程中，由于聚类质心的选择方式是取类中所有元素的均值，因而聚类质量受类边界元素的影响严重，且类的质心起不到代表类中实际元素的作用。因而，在段落聚类算法的选择上，我们采用了对边界元素的影响不那么敏感的 K-medoids[10]算法。

假设 N 维样本空间的每个样本点分别对应于文本中的一个段落向量。这样，段落的聚类问题便可直观化为 N 维样本空间中的 M 个样本点的聚类问题(N：文本中的术语总数，

M：文本中的段落总数）。表1将形式化描述基于 K-medoids 的段落聚类算法：

表1　基于 K-medoids 的段落聚类算法

输　入：二元组(a,b)，它们分别对应于由文本中所有的段落向量组成的段落矩阵、聚类数 K(这里，将 K 的取值范围设为2～M)。
第一步：随机选择 K 个段落向量做为初始 K 个类的 medoids(即代表段落)。
第二步：分派每个段落向量到离它最近的 medoid X 所在的类中。
第三步：计算所有的段落向量到离它们最近的 medoid 的欧拉距离之和。
第四步：随机选择一个段落向量 Y。
第五步：对于所有的 X，如果交换 X 和 Y 可以减小所有的段落向量到离它们最近的 medoid 的欧拉距离之和，则交换 X、Y，否则保持不变。
第六步：重复第二步到第五步，直到没有变化为止。
输　出：三元组(A,B,C)，它们分别对应于聚类数 K 下得到的各个类的标号、类中的代表段落向量以及此类所包含的所有段落向量。

2）聚类分析

采用 K-medoids 聚类算法以及其它众多聚类算法的一个经典难题就是聚类数 K 的捕获问题。在传统的 K-medoids 算法当中，聚类数 K 必须事先由用户来提供，然而在很多情况下，这种要求是不切实际的。就段落聚类而言，由于用户无法事先预知文档中可能存在的潜在主题数，因此也就无法正确的提供所需的 K 值。

针对这个问题，本文提出了一种根据自定义目标函数的取值情况来自适应确定 K 值的新的聚类分析方法。该方法的基本思路是：如果聚类数 K 确定得比较合适，那么对应的聚类结果就能比较好地分辨出文本中的不同主题，相应的，所有主题下最具有代表性的段落的权值平均将会趋向最大化，我们把这个性质称为目标函数的最大值属性。

相应的，我们定义如下的目标函数 $Objf(K)$ 来反映聚类质量并确定聚类数 K：

$$Objf(K)=\frac{\sum_{j=1}^{K}W(P_j)}{K}$$

这里 $W(P_j)$ 是第 j 个类下选定的某个代表段落的权值（代表段落根据上个环节中介绍的 K-medoids 的方法选定，代表段落的权值计算方法见阶段1的 Step2）。将目标函数分别作用于采用 K-medoids 聚类算法得到的对应于各个 K 值的 K 个聚类结果当中，再由目标函数的最大值属性来自适应地确定最终的聚类数 K，完成聚类分析。

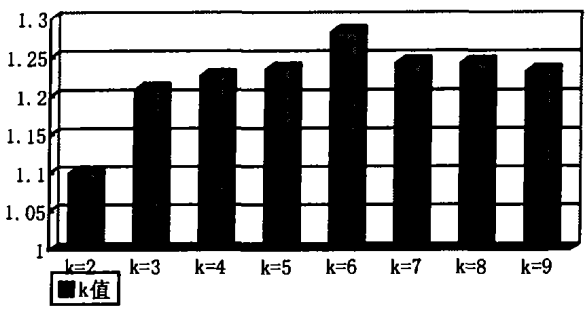

图2　当 K 取不同值时得到的目标函数值的分布情况

图2给出了示例文档“浅析大连市海洋捕捞业面临的形势及对策”在采用此聚类分析方法的过程中得到的目标函数值的具体分布情况。根据目标函数的最大值属性（把使目标函数取最大值时所对应的 K 值确定为最终的聚类数 K）并结合下图可知，$K=6$，即我们采用此方法从该文9个自然段落中共发现出了6个潜在的主题区域。

图3相应给出上述示例文档在采用 K-medoids 的聚类过程中，对应于 $K=6$ 时的聚类结果图。

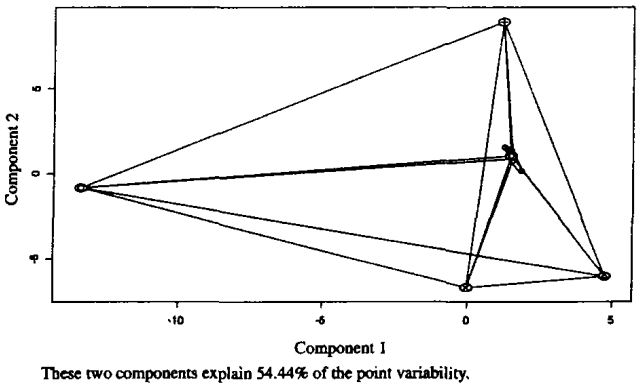

图3　采用 K-medoids 的聚类算法（当 K＝6）时得到的聚类结果图

Step 4：主题区域的发现

以主题区域、主题区域中的代表段落、主题区域所覆盖的全部段落和全部句子的形式输出各个主题区域的完整信息表。

3.2　阶段2：主题区域中代表句的挑选

为了从各个主题区域中挑选出一个最合适的代表句，本文提出了如下的挑选方法：

方法：选取与各个主题区域的语义相似性最大的句子做为代表句

在具体实现该方法之前，有两个问题必须解决：

① 句子和主题区域的向量表达问题

句子和主题区域的向量表达类似于前面所介绍的段落的向量表达，只需将其中术语的权值考察范围由段落内部改为句子或主题区域内部即可。相应地，有句子向量 $VS_j=(WS_{j1}, WS_{j2}, \cdots, WS_{jN})$，主题区域向量 $VA_k=(WA_{k1}, WA_{k2}, \cdots, WA_{kN})$。

② 句子与主题区域的语义相似性计算问题

句子与主题区域的语义相似性计算可以通过计算句子向量与主题区域向量间的向量距离来实现，我们这里采用的是传统的 cosine 向量距离计算方式。相应地，定义句子向量 VS_j 与主题区域向量 VA_k 之间的距离为：

$$Cos(VS_j,VA_k)=\sum_{i=1}^{N}(WS_{ji}\times WA_{ki})/\sqrt{(\sum_{i=1}^{N}WS_{ji}^2)\times(\sum_{i=1}^{N}WA_{ki}^2)},$$

3.3　阶段3：摘要的生成

将从各个主题区域当中挑选出来的主题代表句按照它们在原始文档中的位置先后次序依次输出，即构成了最终的文摘。

4　实验结果及评价

公正客观的评价不同文摘方法的优劣是个颇具挑战性的课题。目前，两种主流的评价方式分别是针对不同方法所产生的文摘做内在式或外延式的评价[7]。我们采用的是前一种评价方式，即通过定义如下的评价因子来进行评价。

1) 评价因子介绍

① 主题覆盖度 TC(对文摘主题完备性的评价)

主题覆盖度(Theme Coverage)的定义:原文中的主题内容被文摘句所覆盖的百分比。该项因子可通过多个人工专家分别打分,取得分的平均值来确定。

② 表达熵 RE(对文摘冗余的定量评价)

为了能客观、有效地评价文摘的冗余量,文[6]所提出用来计算特征选择过程中的特征冗余量的参数,即表达熵(Representation Entropy)被借鉴、改造为我们评价文摘冗余的新的因子。

针对此目的,定义助记符如下:

N:原文中的术语总数(即特征总数);

Nz:文摘中的句子总数;

Lz:由文摘中的所有句子向量所构成的 $Nz\times N$ 阶的句子-术语向量矩阵;

Σz:文摘句子向量之间的 $Nz\times Nz$ 阶的协方差矩阵;

λ_i:Σz 的特征值(eigenvalues) i:$1\sim Nz$;

r_i:Σz 的特征值 λi 的归一标准化,具体定义为:$r_i=\lambda_i/\sum_{i=1}^{Nz}\lambda_i$;

文摘句之间的表达熵的计算公式[6]定义如下:

$$RE=-\sum_{i=1}^{Nz} r_i * \log r_i$$

相应地,基于表达熵的文摘冗余的评判原则如表2所示:

表2 基于表达熵的文摘冗余的评判原则

评判文摘冗余量的原则
在不同文摘方法取相同数目的文摘句的前提下:
若由文摘句子向量间的协方差矩阵计算出的表达熵值越高→说明文摘的冗余越低
若由文摘句子向量间的协方差矩阵计算出的表达熵值越低→说明文摘的冗余越高

③ 主题—冗余信噪比 F

$$F=TC/e^{-RE}$$

我们提出又一新的评价因子,该因子能有效地将上述两个因子综合在一起,客观地反映文摘的质量。F 值越大说明产生的摘要质量越好。

2) 实验结果及评价

我们从国家语委现代汉语语料库中随机抽取了100篇不同题材的文章做为实验语料库,鉴于对篇幅太短的文章进行摘要没有太大实际意义[4],因而我们从中选取了30篇长度在400字以上的语料作为实验样本,采用基于主题区域发现的自动文摘方法(以下简称方法1)及传统的无主题区域检测的文摘方法(以下简称方法2)对上述所有的样本进行了摘要,表3、表4将显示具体的实验数据及比较结果。

表3 实验数据

语料题材	语料ID	语料字数	段落数	发现的主题区域数	主题覆盖度 TC		表达熵 RE	
					方法 1	方法 2	方法 1	方法 2
经济类	d10000801	1461	11	5	0.6	0.56	1.44	1.25
	d10000901	1192	7	5	0.64	0.6	1.36	1.35
	d10100101	1936	14	9	0.66	0.64	2.14	2.06
	d10100201	1778	12	6	0.8	0.5	1.62	1.54
	d10100301	2472	4	3	0.64	0.4	0.81	1.05
	d10100601	1553	11	7	0.9	0.64	1.79	1.83
	d29600501	2400	6	4	0.7	0.56	1.33	1.01
	d29800101	670	4	3	0.64	0.6	1.06	1.01
	d40000301	2026	8	5	0.56	0.52	1.45	1.54
	d40100101	1529	7	4	0.6	0.58	1.19	1.31
文艺类	e10000101	907	4	2	0.72	0.56	0.64	0.24
	e10000201	845	5	3	0.9	0.6	1.06	0.89
	e29600201	2035	5	4	0.72	0.5	1.36	1.21
	e29800201	1831	7	2	0.56	0.52	0.67	0.57
散文类	f20000101	2354	12	7	0.58	0.5	1.92	1.79
	f20000201	1769	9	6	0.64	0.52	1.72	1.50
军体类	g00000201	1163	5	4	0.84	0.56	1.34	1.21
	g00000501	790	6	4	0.64	0.54	1.31	1.26
	g00001201	425	5	5	0.92	0.62	1.45	1.49
	g00100101	1629	10	3	0.84	0.6	0.93	0.82
	g00100301	817	6	4	0.76	0.7	1.32	1.26
	g00100501	1355	4	4	0.84	0.5	1.31	1.12
	g09600901	2179	7	6	0.72	0.62	1.75	1.73
	g09601601	1271	5	3	0.7	0.52	1.03	0.98
生活类	h00000401	1224	6	6	0.72	0.54	1.75	1.60
	h00000601	1331	15	7	0.6	0.5	1.88	1.80
	h00000901	1507	7	3	0.64	0.68	1.05	0.83
	h00001801	1604	8	6	0.68	0.64	1.73	1.66
	h00100301	960	6	3	0.9	0.4	1.04	1.05
	h00100601	1228	6	3	0.8	0.6	1.06	0.89

表4 两种方法的因子评价结果

语料题材	包含语料总数	主题覆盖度均值$\overline{TC}$		表达熵均值$\overline{RE}$		主题—冗余信噪比 F	
		方法1	方法2	方法1	方法2	方法1	方法2
经济类	10	0.68	0.56	1.42	1.40	2.81	2.27
文艺类	4	0.72	0.54	0.93	0.73	1.82	1.12
散文类	2	0.62	0.52	1.82	1.65	3.83	2.71
军体类	8	0.78	0.58	1.31	1.23	2.89	1.98
生活类	6	0.72	0.56	1.42	1.31	2.98	2.08

对30篇实验样本的综合评价表明，我们的方法在上述评价因子的得分上优于传统的无主题区域检测的文摘方法，且对于部分文风自由、主题灵活的文本，我们的方法要优于仅基于相邻段落语义相似性计算的方法。

结论及今后的工作 本文提出了一种基于主题区域发现的中文自动文摘的研究方法，该方法产生的文摘能在尽可能全面地覆盖原文多个不同主题的同时，显著地缩减自身的冗余，从而能有效地平衡两者之间的矛盾。在我们的实验评价过程中，此方法取得了比较好的评价结果。

当然，本方法也存在不足之处，如术语抽取的质量还有待进一步提高、聚类算法和聚类分析算法还需进一步完善等。对于这些不足之处，我们将在今后的工作中逐一改进。

参 考 文 献

1 王继成，武港山，等. 一种篇章结构指导的中文 Web 文档自动摘要方法. 计算机研究与发展，2003,40(3):398～405

2 刘建舟，何婷婷，姬东鸿. 基于开放式语料的汉语术语的自动抽取. 见：第二十届东方语言计算机处理国际学术会议论文集，2003. 43～49

3 Nomoto T, Matsumoto Yuji. A New Approach to Unsupervised Text Summarization. In: Proc. of ACM SIGIR'01, 2001. 26～34

4 Gong Yihong, Liu Xin. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In: Proc. of ACM SIGIR'01, 2001. 19～25

5 Pantel P, Lin Dekang. Document Clustering with Committees. In: Proc. of ACM SIGIR'02, 2002. 199～206

6 Mitra P, Murthy C A, Pal S K. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions of Pattern Analysis and Machine Intelligence, 2002. 1～13

7 MANI I. Summarization Evaluation: An Overview. In: Proc. of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. Tokyo: National Institute of Informatics, 2001

8 MANI I. Recent Developments in Text Summarization. In: Proc. of CIKM'01, 2001: 529～531

9 杨晓兰，钟义信. 基于文本理解的自动文摘系统研究与实现. 电子学报，1998,26(7):155～158

10 Kaufmann L, Rousseeuw P J. Clustering by means of medoids. In Statistical Data Analysis based on the L1 Norm. In: Dodge Y, ed. Amsterdam, 1987. 405～416

11 Rissanen J. Modeling by the shortest description. Automatica, 1978(14): 465～471

（上接第176页）

通常使用的用于估计丢失帧的插值方法。将特征矢量序列 $\{X_0, X_1, \cdots, X_N\}$ 输入到插值器组，特征矢量的每一维单独使用一个插值器，如第 m 维使用 $I_0(m)$。这样根据特征矢量轨迹信息可以估计出丢失的矢量 X_n，其第 m 维的估计 $\hat{x}_n(m)$ 可估计如下：

$$\hat{x}_n(m)=I_m(x_{n-B}(m)\cdots x_{n+F}(m)) \quad (1)$$

上式在估计丢失数据时，使用了其前 B 个特征和后 F 个特征信息。需要注意的是，对实时性的操作，F 要尽可能地小。

多项式插值的方法有很多，一般使用拉格朗日插值，对 $N+1$ 个特征矢量中的第 m 维，其插值形式为：

$$P_N(t)=L_0(t)x_0(m)+L_1(t)x_1(m)+\cdots+L_N(t)x_N(m) \quad (2)$$

其中拉格朗日系数 $L_n(t)$ 是 N 阶多项式。

一般为简化计算取一阶拉格朗日多项式，这样有：

$$\hat{x}_n(m)=\frac{t_n-t_q}{t_p-t_q}x_p(m)+\frac{t_n-t_p}{t_q-t_p}x_q(m) \quad (3)$$

其中，$\hat{x}_n(m)$ 是丢失的第 n 个特征矢量的第 m 维参数的估计，$p<n<q$，$x_p(m)$ 和 $x_q(m)$ 分别是 n 前后两个特征矢量的第 m 维参数。图5给出了这种插值的情况。

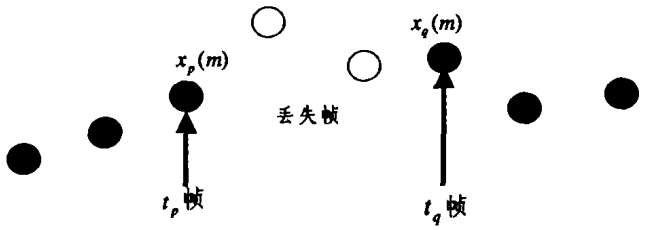

图5 丢失特征帧的多项式插值示意图

以上我们讨论了网络环境下由于语音编码和丢包所引起的语音识别性能下降的情况，以及相应的解决方法。从总体上看，本领域的研究工作才刚刚起步，还有许多问题亟待解决。

结束语 网络环境下的语音识别是近年来随着网络技术的发展而出现的一个新的研究课题。国际上该方面的研究已经开展起来，国内这方面的工作还比较少。相信不久的将来一定会有越来越多的人从事该方面的研究，也必将会有所突破。

参 考 文 献

1 Jim Van S, Jeff Z M. Investigation of Speech Recognition over IP Channels, CASSP 2002, IEEE Press. 3812～3815

2 Pelaez-Moreno C, Gallardo-Antolin A, Diza-de-Maria F. Recognizing Voice Over IP: A Robust Front-End for Speech Recognition on the World Wide Web. IEEE Transactions on Multimedia, 2001, 3(2): 209～218

3 Miner B. Robust Voice Recognition over IP and Mobile Networks. In: Proc. of the Alliance Engineering Symposium, 2000. 1197～1200

4 Miner B, Semnani S. Robust Speech Recognition over IP Networks. In Akansu A N. ICASSP 2000, Istanbul, Turkey: IEEE Press. 1791～1794

5 Miner B, et al. Bobust Distributed Speech Recognition Across IP Networks. In: Proc. IEE Colloquium ISDS, 1999. 6/1-6/6

6 Quercia D, Docio-Fernandez L, Garcia-Mateo C, Farinetti L, De Martin J C. Performance Analysis of Distributed Speech Recognition Over IP Networks on the AURORA Database, ICASSP 2002, IEEE Press. 3820～3823

7 韩纪庆，张磊，吕成国，王承发. 可穿戴计算机中的语音处理技术. 计算机科学，2002,29(5):107～109