意义段划分问题研究

刘挺　吴岩　王开铸

（哈尔滨工业大学计算机系　150001）

邵艳秋

（黑龙江交通高等专科学校　150050）

Research on Semantic Paragraph Partition

Liu Ting　WuYan　WangKaizhu

(Dept. of Computer, Harbin Institute of Technology, 150001)

Shao Yanqiu

(Heilongjiang College of Communications, 150050)

自然段是具有换行标志的，表示一个相对完整意思的语言单位。意义段是界于篇章和自然段之间的语言单位，它是若干个相邻自然段构成的集合。意义段划分问题是如何将自然段序列转换为意义段序列的问题。

篇章是一个结构体，对于没有章节标记的文章，必须首先进行意义段的划分，这样才能够从宏观上把握原文的结构，分清主次详略。句子甚至自然段都不是篇章的直接成份，只有意义段才是篇章的直接成份，要研究篇章的宏观结构必须研究意义段以及意义段之间相互关系。

属于同一意义段的若干自然段应谈论相同的主题，从形式上看，这些自然段应含有某些共同的关键词，这是笔者所提出的划分意义段的基本思想。这里所谓的关键词是指从原文中抽取出来的，能够反映原文内容的词。

几个相邻自然段之间意义上的相关程度称为关联度。为了用量化值衡量关联度，我们设定若m(m>1)个相邻自然段含有n(n>0)个共同的关键词，那么这m个自然段构成了一个潜在意义段，其关联度为n。此外，特别设定任意一个自然段本身均构成一个潜在意义段，且关联度为1。设N为文章中所有潜在意义段的关联度之和，其中某个潜在意义段的关联度为n，那么该潜在意义段的关联率为n/N，关联费用为-log(n/N)。

现在我们来描述意义段划分问题：假定文章中的n个自然段依次排列，段1之前存在初始结点0，段i之后存在结点i，结点n为目标结点。若段j，j+1，…，k构成一个潜在意义段，则从结点j引一条弧指向结点k，弧的费用等于潜在意义段的关联费用。至此构成了意义段划分问题的状态空间图，而意义段划分问题本身则可以归结为从初始结点到目标结点的最小费用路径搜索问题。

最小费用路径上各意义段的关联费用之和最小，由于关联费用=-log(关联率)，因此最小费用路径上各意义段的关联率的乘积最大，而关联率反映了一个潜在意义段成为真正意义段的可能性，故根据最小费用路径进行的意义段的划分是各种划分方法中可能性最大的一种。

本文提出的算法适用于社会实用文体，包括科技文献、政论文、公文等。对于文艺类作品不适用。