自动文摘系统中的意义段划分问题研究*

万敏 罗振声 郭玉箐
清华大学计算语言学研究室，北京 100084
E-mail:luozhs@mail.tsinghua.edu.cn

摘要：意义段划分是自动文摘系统中文本结构分析阶段所要解决的一个重要问题。对有子标题的文章，本文总结了中文文章的各种子标题模式，根据标题进行意义段划分。对无子标题的文章，则建立以段落为基础的向量空间模型，根据段落相似度进行聚类实现意义段的划分。

关键词：自动文摘 意义段划分 向量空间模型 聚类

Study on Semantic Paragraph Partition in Automatic Abstracting System

Wan Min Luo Zhensheng Guo Yuqing
Laboratory of Computational Linguistics, Tsinghua University, Beijing 100084
E-mail:luozhs@mail.tsinghua.edu.cn

Abstract: Semantic paragraph partition is a significant problem during text structuring in automat abstracting system. For article those has headings, the paper presents heading models in Chinese text to divide article to semantic paragraphs according headings. For article those has not headings, the paper establish vector space model for the whole article based on paragraph, then cluster semantic relative paragraphs to semantic paragraph.

Keyword: Automatic abstracting, Semantic Paragraph Partition, Vector Space Model, Clustering

一、引言

在自动文摘系统的研究中，不仅需要对文章字词句进行精细考察，同时也要求系统能对文章文本结构进行分析，保证文摘对原文内容的覆盖度。文本结构包括：(1)文章主题数，即由几个相对独立的部分组成；(2)各段落所属的主题；(3)各个主题或段落之间的相关度。

对各个段落之间相关程度的考察，是为了确定哪些段落与文章的主题关系最为紧密，以作为文摘抽取的重点；确定文章主题数的目的，则是为了可以从不同的主题段中分别抽取句子，从而确保文摘信息的全面性。在对文章文本结构进行分析的过程中，研究意义段划分及意义段之间的联系是一个很重要的内容。所谓意义段，是指介于篇章与自然段之间的一个语言单位，它由若干个相邻自然段构成，在意义上表达或阐述一个相对独立的主题。正确地对文章进行意义段划分，可以使文摘系统对文章的各主题及其联系有所把握，确保

* 国家自然科学基金项目，批准号[69972025]

摘取的文摘能全面地、详略得当地反映文章的各个主题，使文摘能涵盖文章的最大信息量。

多数科技文献具有比较规范的组织结构和清晰的各级标题，对这类文献通过标题的识别判断，就能很好地划分意义段。对不能使用标题法划分意义段的文章，由于表达同一主题使用的特征词相对集中，故可通过分析文本各段落含有的特征词，实现意义段自动划分。

二、 基于标题层次的意义段划分

2.1 文章子标题的形式

子标题的形式在不同的文章中也往往各不相同。从形式上看，一个子标题可以被划为如下三部分：（1）前缀部分（可为空）；（2）序数字部分（不能为空），标识该标题的序号，是整个标题的核心；（3）后缀部分（可为空）。下面是一些常见的标题及其组成部分实例：

标题形式	前缀部分	序数字部分	后缀部分
第壹章	第	壹	章
§2	§	2	
iii)		iii	)
（五）	（	五	）
§2－4－7	§2－4－	7	
⑧		⑧	

表 2.1 常见标题实例

对标题三个部分的划分，可判明标题的序号与层次。两标题若前缀部分与后缀部分都相同，就是同一层次标题，否则不是。如 1.2.1）与 1.2.2），前缀都是 1.2.，后缀都是)，故是同一层次标题。而 1.2)与 1.2.1)，前缀是 1.和 1.2.，并不不相同，故不是同一层次标题。

2.2 子标题的用字集合

从以上实例来看，标题后缀部分的构成相对较为简单，至多是一个符号；序数字部分则是数字的各种形式，如阿拉伯数字、汉字、罗马数字、带括号数字、带圈数字等等；前缀部分的构成比较复杂，可以为空，可以是单个符号，也可以是一个带序数字的串。在对大量的文章标题进行考察后，整理得到以下几种子标题的用字集合：

1）强序数字集合，出现在标题的序数字部分。强序数字一般可独立作为一个标题，其后不再跟后缀。有以下 4 类：

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩　㈠ ㈡ ㈢ ㈣ ㈤ ㈥ ㈦ ㈧ ㈨ ㈩

1. 2. 3. 4. 5. 6. 7. 8. 9. 10.　(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

2）弱序数字集合，主要出现在标题的序数字部分，也可能出现在标题的前缀部分。弱序数字一般不独立作为一个标题，其后要跟一定后缀才形成完整的标题。有以下 12 类：

1 2 3 4 5 6 7 8 9 10　　1 2 3 4 5 6 7 8 9 1 0

i ii iii iv v vi vii viii ix x xi xii　　I II III IV V VI VII VIII IX X XI XII

i ii iii iv v vi vii viii ix x　　Ⅰ Ⅱ Ⅲ Ⅳ Ⅴ Ⅵ Ⅶ Ⅷ Ⅸ Ⅹ Ⅺ Ⅻ

一 二 三 四 五 六 七 八 九 十　壹 贰 叁 肆 伍 陆 柒 捌 玖 拾
abcdefghijklmnopqrstuvwxyz　a b c d e f g h i j k l m n...
A B C D E F G H I J K L M N...　A B C D E F G H I J K L M N...

3）前缀符集合，出现在标题的前缀部分的开头。有以下 14 个：

· § ☆ ★ ○ ● ◎ ◇ ◆ □ ■ △ ▲ ※

4）后缀符集合，出现在标题的后缀部分。有以下 9 个：　. . 〉 〉 〉 〉)> 〉

5）配对符集合，成对出现在标题的前缀与后缀部分。每一对由前符与后符两个符号组成，有以下 14 对：〈〉 <> 〔〕 （） 〈 〉〈 〉() [] 〖〗【】[] {} 〔 〕{}

6）分隔符集合,出现在标题前缀部分，用于隔开各序数字，有以下 6 个:. . · - — _

2.3 子标题模式及识别

根据各类文章中出现的子标题情况，总结出如下几种子标题模式：

1）直接用强序数字作为标题。模式：强序数字　例子：⑴ ② 3. ㈣

2）带后缀的标题。　模式：序数字 后缀符 例子：一）壹. 1) 1 ）i) ii）a) A) A. a ）

3）带配对符的标题。 模式：配对符前符 序数字 配对符后符 例子：（一）【壹】(a)

4）带前缀的标题。模式:[前缀符][弱序数字 分隔符]*序数字[后缀符] 例子: § 1.2>

5）章节式标题。　模式：第 汉字序数字 [讲篇章节]　例子：第二章　第三节

6）汉字顿号式标题。 模式：汉字序数字 、例子：一、 二、 壹、

根据这些标题模式，可对文本各段开头进行模式匹配，确定文章标题所在位置，再通过判断标题前缀与后缀部分是否相同，确定各标题层次，从而划定文章题目及全部子标题。

2.4 利用标题识别进行意义段划分

在得到全文的标题情况后，还不能就将每一个标题下所统领的部分划分一个意义段，因为标题的使用情况十分复杂，不同的作者也有不同的习惯。有些小标题下统领的内容过少，不适宜将其划为一个意义段。因此需要预先设定一个段数或字数的阈值，只有当标题所统领的部分超过规定的阈值，才将其划为一个意义段。另外，通常情况下，文章的标题是带有层次关系的，因此这样划分出的意义段也会带有层次关系。

通过大量文本的测试，发现以上总结的标题用字和标题模式涵盖了实际文本的各类标题情况，能由此有效地识别出文章的标题。对有子标题的文章，这种由标题引导的意义段划分具有很高的准确性与实用性，在自动文摘系统中取得了较好的实际效果。

三、 基于向量空间模型的意义段划分

对无子标题的文章，意义段划分通过考察用词情况而实现。作者在表达阐述一个主题时，其所用重点词汇通常局限在能代表该主题所涉内容的一个较小范围，具一定的重复性。若两个段落所含词语，特别是高频词，在一定程度上发生重复，即可初步认为这两段谈的是同一主题，若位置合适的话，即应划在同一个意义段中。基于这一假设，本文选择向量

空间模型（Vector Space Model，简称 VSM）实现对篇章结构的自动分析和意义段划分。

3.1 向量空间模型的建立

所谓 VSM，是将文章中的一个词视为空间中的一个维度，设考察的特征词共有 n 个（一般要去掉低频词与禁用词），记它们分别是 T_1，T_2，…，T_n，则构成一个 n 维的向量空间。

文章中的段落可形式化为 P（T_1，W_1；T_2，W_2；…；T_n，W_n），其中 $W_k(1\leqslant k\leqslant n)$为特征词 T_k 在段落 P 中的权值，这样段落 P 就可视为 n 维空间中的一个向量 P（W_1，W_2，…，W_n）。权值 W_k 的计算一般采用信息论中的熵值公式：$W_k = f_k * \log F_k$　　（公式 3.1）

其中 F_k 为特征词 T_k 在文章中出现的频度，f_k 为特征词 T_k 在该段落中出现的频度。$\log F_k$ 代表一个特征词 T_k 所含信息量，故 W_k 就是该段落中出现的全部特征词 T_k 所含的信息量。由于文摘主要任务是提取原文信息，故用信息量作为段落所含特征词的权值是十分恰当的。

为了划分意义段，还必须刻画两个段落间的相关程度。由于段落已被映射成为 n 维向量，所以可以利用向量间的夹角余弦来衡量两个段落间的相似性。设有两个段落 P_i 和 P_j：P_i =（W_{i1}，W_{i2}，…，W_{in}），P_j =（W_{j1}，W_{j2}，…，W_{jn}），用 Sim（P_i，P_j）来记它们之间的相似度，又记向量空间的原点为 O，则利用向量间的夹角余弦公式可有：

$$Sim(P_i,P_j)=\cos\angle P_iOP_j=\frac{\sum_{k=1}^{n}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{n}W_{ik}^{\ 2})(\sum_{k=1}^{n}W_{jk}^{\ 2})}}$$　　（公式 3.2）

3.2 段落相关图

为文章建立以段落为基础的 VSM 的具体过程：设一篇文章共 m 段 P_1，P_2，…，P_m。统计词频，去掉低频词与禁用词后，需考察的特征词共 n 个：T_1，T_2，…，T_n，在全文中的频度分别为 F_1，F_2，…，F_n。对每个段落 $P_i(1\leqslant i\leqslant m)$，统计出词频后利用公式（1）计算出各项权值，从而建立起 m 个段落向量。然后对每两个段落 P_i，P_j $(1\leqslant i, j\leqslant m)$，利用公式 3.2 计算出段落相关系数 Sim($P_i$，$P_j$)。这样就建好了整篇文章的 VSM。

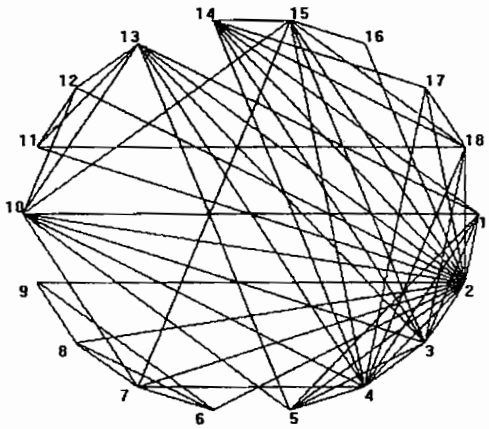

图 3.1《数码照相机的常识与选购》的段落关系图

利用 VSM 对篇章结构进行分析时，为了直观地考察文章各段落的联系情况，画出段落关系图是一个较好的办法。将文章的 m 个段落以平面上 m 个点表示，设定一个阈值 $Q(0\leqslant Q<1)$，对任意的两个段落 P_i，P_j $(1\leqslant i, j\leqslant m)$，如果 Sim($P_i$，$P_j$)$\geqslant$Q，则在对应的两点间连线；如果 Sim($P_i$，$P_j$)<Q，则不连。有连线的两点表明对应的两个段落有较大的相关度，在内容上比较相近，在意义上有所联系；而无连线的两点则表明对应的两个段落相似程度较小，在意义上的联系也较少。Q 值的选取决定了段落关系图的连线多寡，若取值过低，则图中连线增多，不容易分清真正有着强联系的段落；若取值过高，又容易造成图中连线过少，不能很好地反映文章整体结构。根据我们的经验，Q 值取在 0.2~0.35 之间比较合适。

图 3.1 是对语料库中的一篇文章《数码照相机的常识与选购》建立向量空间模型后，取阈值 Q=0.24 后得到的段落相关图。图中数字 1~18 表示全文的 18 个段落。从该图上可以直观地看出，此篇文章应划为 4 个意义段：1~5 段，6~9 段，10~13 段，14~18 段。这与实际情况是相符的，原文从概念、工作原理、最新发展情况、购买常识等 4 个方面向读者介绍了数码照相机，段落起始位置与段落相关图上反映出的情况相一致。

3.3 利用段落相似度进行聚类实现意义段划分

观察段落相关图可以发现，各意义段内的段落间联系十分紧密，而位于不同意义段内的段落，特别是位于相邻意义段中的段落，则联系不够紧密。因此，同属一个意义段的段落之间具有高度的相似性，可视为具有一定关键词特性的一个聚合类。通过建立文章的 VSM，文章中的 m 个段落可看成高维空间中的 m 个点，对文章意义段的划分可看成是对这 m 个点的聚类问题，两点间的相似性由对应段落的相似度来衡量，以此为聚类标准，并要求聚在同一类中的各段在原文中连续出现，即可设计出一个聚类式的意义段划分算法。

为了利用聚类的方法实现意义段划分，一个意义段被看成是聚为一类的高维空间中的一些点的集合，我们需要给出一种衡量这些点集的相聚度的办法。设一个意义段含有文章中连续的 k 个段落，对这 k 个段落的段落向量求其算术平均，作为该意义段的中心向量 c，然后分别计算这 k 个段落与中心向量 c 的相似度，将其总和作为该意义段的相聚度。

划分问题就是要经过调整，使分出的各意义段的相聚度的总和最大。具体算法如下：

1）设文章共有 m 段 $P_1,P_2,...P_m$，建立文章的 VSM

2）按照某种初始划分算法将文章分为 p 个初始意义段 $L_1,L_2,...,L_p$

3）计算各 Li 的相聚度 $J(L_i)$及相聚度总和 J

4）对任意两个相邻意义段 L_i,L_{i+1}，若改动其划分线可使总相聚度 J 增加则作出调整

5）返回到第 4)步，直至已作不出任何调整

由于文摘系统中 VSM 是广泛用到的一种全文分析方法，该算法建立在 VSM 的基础上，不用采取新的计算方式，模拟了对段落相关图的观察，在理论上有较高可信度，在实际系统中能方便地实现并达到一定效果。在算法第 2)步中要用到一个初始划分算法，这一步在整个划分过程中也十分重要。一个好的初始划分算法可减少第 4)步的循环调整次数，便于快速达到最优划分结果。下面就来介绍一个基于相邻段落相似性的意义段初始划分算法。

3.4 基于相邻段落相似性的意义段初始划分算法

再次观察段落相关图，可看到，在两个意义段相接处的两个相邻段落，即上一个意义段的末段与下一个意义段的首段之间，一般都无连线或连接强度较低，说明这两个段落的相似性较低。由于在两个意义段之间存在着话题的转移，所以它们的用词也会存在一定的差异，反映到向量空间上就是两个段落向量之间有较大夹角，从而出现相似度较低的情况。

基于这一现象，可通过考察相邻段落间的相似度来作出意义段的初始划分。将相邻段落相似度较低地方作为意义段的划分边界。同时，还应注意到，意义段一般来讲是由若干段落组成的，故在考虑意义段的划分边界时还应照顾位置因素，要求初始划分出的各意义段边界相隔不能太近，否则有可能造成大量单个段落被划成一个意义段。具体算法如下：

1）设全文共有 m 段，记为 P_1，P_2，...，P_m

2）分别计算各相邻段间的相似度 $Sim(P_i, P_{i+1})$，i=1,2,..,m-1

3）从所得的诸 $Sim(P_i, P_{i+1})$结果中选出最小的一个作为一个意义段划分边界

4）再从剩下的诸 $Sim(P_i, P_{i+1})$结果中选出最小的一个来，如果它与现有的各边界相距大于一定的阈值，则也作为边界选出，否则不选

5）重复第 4）步直至意义段数达到预定要求或剩下的 $Sim(P_i, P_{i+1})$都已大于设定的阈值

该方法着重考察相邻两段间的关系，计算复杂度小。由于它实际上只注意了最终划出的意义段的首末段，未全面考察内部的段落，故划分准确率不很高，只能作为意义段的初始划分算法，而不能独立进行意义段划分。但作为一个初始划分算法其效果还是很理想的。

此方法的难点在于第 5）步的停止条件。算法通过预先设定意义段数目或相似度阈值来停止划分，但意义段数目的选取与相似度阈值的选取都没有很好的可循方法，只能通过经验得到，而对不同的文章设定的值也往往不一样，这是本方法在通用性方面的一个缺陷

四、 结语

综合以上提的方法，在我们所实现的自动文摘系统中，是这样处理意义段划分问题的：

1）读入文章，利用模式匹配的办法进行题目、子标题及标题层次识别

2）对于有子标题文章，转 3），否则转 5）

3）扫描各标题，如果其统领部分超过规定的阈值，则将其划为一个意义段

4）将所得的意义段按标题层次组织起来，即完成了带层次的意义段划分，结束

5）统计全文词频，去除禁用词及低频词

6）建立以段落为基础的向量空间模型，计算各段落间的相关系数

7）考察相邻段落相似度，进行意义段初划分

8）利用段落相似度进行段落聚类，完成全文的意义段划分，结束

在自动文摘系统的文本结构分析阶段，对文章进行意义段划分是一个很重要的环节。本文对该问题进行了初步的探讨，并在实际系统中实现了所提的算法，取得了预期的效果。

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] Marti A. Hearst, Multi-Paragraph Segmentation of Expository Text, Processdings of the 32nd Annual Meeting of the Association for Computational Linguistics(ACL 94), June 1994

[3] G.Salton, A.Wong, C.S.Yang, A Vector Space Model for Automatic Indexing, Communicating of the ACM, 1995, Vol.18

[4] G.Salton, A.Singhal, M.Mitra & C.Buckley, Automatic Text Structuring and Summarization, Information Processing & Management 1997, Vol:33 No.2, 193－207

[5] 王永成，许慧敏，OA-1.4 版中文自动摘要系统，高技术通讯，1998，1

[6] 刘挺，吴岩，王开铸，绍艳秋，意义段划分问题研究，语言工程，清华大学出版社，1997

[7] 林鸿飞，战学刚，姚天顺，文本层次分析与文本浏览，中文信息学报，第 13 卷，第 4 期，7－15

[8] 刘挺,吴岩,王开铸,基于信息抽取和文本生成的自动文摘系统设计,情报学报,1997 年 12 月,第 16 卷(增刊)

[9] 薛慕芳,李晓黎,郭炳炎.汉语文摘系统中文本结构的自动分析.语言工程,清华大学出版社, 1997: 332~337