汉语文摘系统中文本结构的自动分析*

薛翠芳　李晓黎　郭炳炎
（山西大学计算机科学系 030006）

摘　要：本文试图运用向量空间模型来确定文本段落之间内容的相关性，从而实现文本主题的自动分析。这不仅为自动文摘技术探索一条新途径，而且也为全文检索等信息处理技术提供了一些有用的信息。

关键词：向量空间模型　　自动文摘　　文本结构

Automatic Analyses of Chinese Text Structure in Abstracting System

Xue Cuifang LI Xiaoli Guo Bingyan
(Department of Computer Science, Shanxi University)

Abstract: This paper aims at determining the relevance between text paragraphs by making use of Vector Space Model, and thereof realizing automatic analysis of text structure. this paper will propose a new way to the study of automatic abstraction techniques, It also provides some useful information for the other information processing techniques, includeing the technique of document retrieval.

Keyword: Vector Space Model　　automatic abstraction　　text structure

一、引言

当今，人们对信息自动处理的需求越来越迫切，在我国中文信息的自动处理也正吸引着越来越多的研究者，自动文摘技术便是其中之一。人工摘要是在理解原文的基础上做出来的，但是让计算机象人一样去理解一篇文章至少在目前还是不可能的。虽然从八十年代以来有不少学者对利用自然语言理解技术来实现自动文摘作了一些探索，但这些研究一般都要求选择一个较窄的领域，而且要求有可利用的领域知识，这些限制对一般的自动文摘任务是无能为力的。在自动文摘领域中，目前较成功的研究大都以机械文摘技术为主[1][2]，即从原文中抽出句子来生成摘要。在这种背景下，文本结构的自动分析对自动文摘来说就显得尤为重要。

通过对文本结构的自动分析，让计算机自动确定给定文本是单一主题还是多主题结构。对

* 本项目得到国家自然科学基金资助(69673011)

单一主题的文本，能自动确定对表达主题有贡献的若干个段，我们称之为主题段。实现自动文摘时，可从这些主题段中而不用从全文中抽取句子。对多个主题的文本，可以采用特定的处理策略。例如对有两个主题的文本在实现自动文摘时，可以从两个小主题各自的主题段中抽取句子，以保证所生成摘要内容的完整性。

二、文本结构分析的基本思想

我们用向量空间模型来实现文本结构或主题的分析[3][4][6]。在向量空间模型中，文本或文本段落被表示为等长词向量。通过比较每一向量对来确定对应的文本或文本段落的内容是否相似。

文本的内容和使用的词是对应的，对给定的每个文本，我们选定了特征词向量来表示对应文本，用向量单位化内积的平方来衡量它们的相似程度。

传统的相似性度量方法是上下文无关的，即任意两个对象之间的相似性只取决于所考虑对象的特性，而不会受到上下文的影响（这里的上下文指围绕对象的“环境”）。这种方法有一定的局限性，我们认为在度量两个对象之间的相似性时，不仅应该考虑所涉及的单个对象的特征，还应考虑适用于刻画对象整体特性的外部概念，需要捕捉能从整体上刻画聚类特征的特性。为了检测这种特性，系统必须具备识别或表达对应于一定“概念”的对象机制的能力，在我们的系统中，“概念”即全文的主题。

从概念聚类的角度看，向量 A、B 的相似度不仅取决于 A 向量和 B 向量各自对应的段和邻近的段集 E，而且还 取决于包含 A、B 段的整体概念集 T。

相似性(A,B)=F(A,B,E,T)

在文本结构的自动分析中，首先运用向量空间模型计算出文本每个段落与全文的相似程度，以此为依据找出若干个候选主题段，再计算出这些候选主题段落间内容上的相关程度，以此为依据判断这些候选主题段是否表达同一个主题，若是，则可判定对应的文本是结构简单的单一主题文本；若不是，则试图找出能共同表达某一主题的某几个候选主题段落，即一个文本块，我们将这样的文本块定义为一个小主题(subth)。对多主题的文本，应该能找出多个这样的小主题，而将全文的主题定义为一个大主题(TH)。

$TH=SUBTH_1 \cup SUBTH_2 \cup \cdots\cdots \cup SUBTH_n$，n 为小主题的个数。

三、文本结构分析算法

在给出汉语文本结构自动分析算法之前，先对一些概念作如下定义：

定义 1：$F(P_i)=(W_{i1},\cdots\cdots W_{ik})$称为局部特征向量，其中：$W_{ij}$表示文本 T 的特征词列表中第 j 个元素在段落 i 中的权值。k 为特征向量中元素的个数。

定义 2：T 为一文本段落集合，则 $F(T)=(W_1,W_2,\cdots\cdots W_k)$称为对应文本的全局特征向量，$W_i$ 表示文本 T 的特征词列表中第 i 个元素在全文中的权值。

定义 3：某个段落 P_i 与全文 T 的相关系数 $M(P_i,T)$称为全局相关系数，定义为：

$$M(P_i,T)=[M(F(P_i),F(T))]^2=\frac{[F(P_i)*F(T)]^2}{[\|F(P_i)\|*\|F(T)\|]^2}$$

其中：$F(P_i)*F(T)=\sum_{j=1}^{k}W_{ij}*W_j$，　$W_{ij}\in F(P_i)$，$W_j\in F(T)$。

$\|F(P_i)\|=(W_{i1}^2+W_{i2}^2+\cdots+W_{ik}^2)^{\frac{1}{2}}$　　$\|F(T)\|=(W_1^2+W_2^2+\cdots+W_k^2)^{\frac{1}{2}}$

定义 4：即段落 P_i 与段落 P_j 之间的相关系数 $M(P_i,P_j)$ 称为局部相关系数，定义为：

$$M(P_i,P_j)=[M(F(P_i),F(P_j))]^2=\frac{[F(P_i)*F(P_j)]^2}{[\|F(P_i)\|*\|F(P_j)\|]^2}$$

其中：$F(P_i)*F(P_j)=\sum_{t=1}^{k}W_{it}*W_{jt}$，　$W_{it}\in F(P_i)$，$W_{jt}\in F(P_j)$.

汉语文本结构分析算法为：

INPUT：一篇文本 T

1. 对文本 T，构建全局向量 F(T)
2. 假设文本 T 共有 n 段

   FOR I=1 TO n

   构造局部特征向量 $F(P_I)$，$P_I\in T$，计算全局相关系数 $M(P_I,T)$
3. 选取一阈值 Q_1，若 $M(P_I,T)>Q_1$，则认为对应的段落 P_I 为一个候选主题段，假设共选了 m 个候选主题段。
4. 计算所选的 m 个候选主题段相互之间的局部相关系数，得出局部相关系数矩阵 P_{m*m}

段落号：P_1	P_2	P_3 ……	P_{m-1}	P_m
P_1		ˇ	ˇ	ˇ
P_2		ˇ	ˇ	ˇ
P_3		• •	ˇ	ˇ
•		• • •		ˇ
•		• • •		ˇ
•		• • •		ˇ
P_{m-1}				ˇ
P_m				

矩阵 P_{m*m} 是一对称矩阵，故只需计算出相应的上（或下）三角矩阵即可. 任取 $X_{ij}\in P_{m*m}$，有 $X_{ij}=M(P_i,P_j)$。

5. 确定一阈值 Q_2，参考矩阵 P_{m*m}，构建矩阵 S_{m*m}，方法为：任取 $X_{ij}\in S_{m*m}$，有

$$\begin{cases}X_{ij}=1,\text{若对应的 }Y_{ij}\in P_{m*m},\text{满足 }Y_{ij}\geqslant Q_2;\\ X_{ij}=0,\text{若对应的 }Y_{ij}\in P_{m*m},\text{满足 }Y_{ij}<Q_2;\end{cases}$$

6. 分析矩阵 S_{m*m}，确定输入文本是单一主题还是多主题。同时对每一小主题，确定表达它的是哪几个段落。

OUTPUT：小主题的个数及与每个小主题相对应的段落。

为便于分析，利用算法输出的相关系数矩阵，构建输入文本的文本结构图。其方法是：对确定的 m 个主题段，用 m 个结点来代表. 再依据确定的阈值 Q_2，任取 $X_{ij}\in P_{m*m}$，若 $X_{ij}\geqslant Q_2$，则在对应段落 P_i，P_j 的结点间连线，否则对应结点间无连线。

显然，阈值 Q_1，Q_2 取不同的值便会有不同的结果。选取一个测试集，Q_1，Q_2 的取值根据对测试集的实验来确定，在实验过程中动态调整，人工判断是否合理，在进行一段测试后，便可有一缺省值。在我们的实验中，取缺省值 $Q_1=0.3$，$Q_2=0.3$

四、举例

例 1　发表在 1990 年 6 月 7 日人民日报第 5 版上的一篇标题为“重点工程不可铺张浪费”的文本，该文共四段。全文的 4 个段落都覆盖了文中讨论的主要问题。这意味着该文内容紧凑，易于阅读，全文只有一个主题。文本语义上的一致可以由文中各段内容都与某个段相关来证实。该例中第 3 段便是一个中心段。各段间相关系数矩阵见图 1－1 所示，相应的文本结构图见图 1－2：

	1	2	3	4
1		0.571	0.563	0.388
2			0.741	0.18
3				0.422

图 1－1　例 1 的局部相关系数矩阵

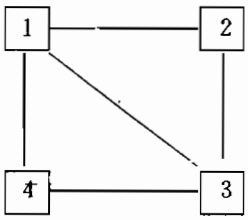

图 1－2　例 1 的文本结构图

表 1 给出了各段的主要内容，可以看出，文本每一段都是围绕着“重点工程的建设”而展开，主要是论述应厉行节约，不可铺张浪费。

表 1　例 1 各段的主要内容

段名	基本内容
P_1	论述秦皇岛码头工程的生活福利设施十分简朴
P_2	与秦皇岛码头工程相对的另一工程
P_3	详细说明搞工程应以工程为主，不应在生活福利上大量投资的理由
P_4	再次提倡秦皇岛码头工程的精神

表 2　例 2 各段主要内容

段号	基本内容
P_1	鸦片战争的经验教训表明：只有共产党才能救中国，只有社会主义才能发展中国。
P_2	鸦片战争惨败的原因及留给我们的另一启示

这种文本的每一段都与全文主题密切相关，因此作摘要时应从所有段落中抽取句子。

例 2 发表在 1990 年 6 月 1 日人民日报第三版上的一篇文章，题为“北京数十名理论工作者探讨鸦片战争与中国发展道路”，副标题为“结论：只有社会主义 才能发展中国”。该文一共只有两段。取 $Q_1=0.3$ 则可以认为两段都与全文的主题有不同程度的相关，但两段之间的局部相关系数 $M(P_1,P_2)=0.14$，取 $Q_2=0.3$，则可认为该段讨论了两个不太相关的主题。各段主要内容见表 2。可见该文的结构与例 1 完全不同，要对这种文本作摘要，应从这两个不太相关的小主题对应的段落中都抽取信息才能保证摘要主题内容的完整性。

例 3　发表在人民日报 1990 年 6 月 9 日第二版的一篇“中华人民共和国 1990 年特种国债条例”该条例一共 11 条，对每一条构造一局部向量，运用向量空间模型对其进行处理。取 $Q_1=0.3$，则可选出 10 个候选主题段：$P_1,P_2,P_3,P_4,P_6,P_7,P_8,P_9,P_{10},P_{11}$。（其对应的相关系数矩阵如图 3－1）

由于本条例不同条款从不同的方面（如：发行对象、发行数额、偿还期、认购任务等）对特种国债加以说明，因此这种文本主题的确定比较困难。对这种类型的文本，进行自动文摘结果不易被人理解，我们认为对这种文本不宜进行自动文摘。

	1	2	3	4	6	7	8	9	10	11
1		0.666	0.666	0.028	0.5	0.615	0.666	0.166	0.166	0.25
2			0.444	0.038	0.5	0.615	0.666	0.166	0.111	0.166
3				0.038	0.333	0.410	0.444	0.111	0.111	0.166
4					0.02	0.035	0.038	0.238	0.038	0.057
6						0.037	0.333	0.166	0.083	0.125
7							0.410	0.102	0.102	0.153
8								0.166	0.111	0.166
9									0.111	0.166
10										0.75

图 3—1 例 3 局部相关系数矩阵

其对应的文本结构图见图 3—2($Q_2=0.2$)和图 3—3($Q_2=0.5$)

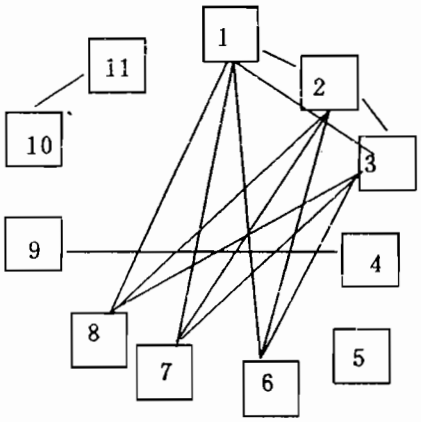

图 3—2 例 3 文本结构图($Q_2=0.2$)

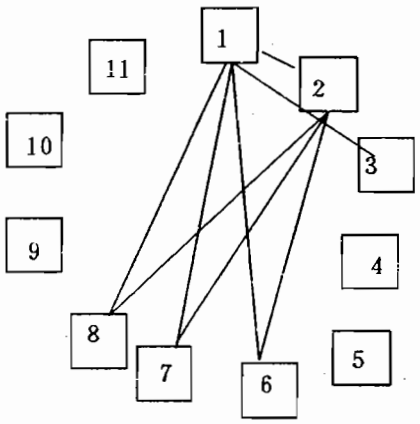

图 3—3 例 3 文本结构图(Q2=0.5)

五、测试与评价

我们的测试集是选自 1990 年 6 月 1 日至 6 月 10 日的人民日报的 37 篇文章。对测试集中的每一篇文章,采用本文提出的方法进行文章结构的分析,输出结果,并与人工判断相比较,得出小主题的划分以及主题段判断的正确率。

本方法的测试结果如表 5—1 和表 5—2 所示:

表 5—1 小主题划分的结果

	单主题	多主题
人工判断(篇)	4	33
真实结果(篇)	3	32
正确率(%)	75	96.96
平均正确率	94.59	

表 5—2 各小主题主题段的划分结果

	单主题	多主题
人工判断(篇)	3	32
测试结果	3	25
正确率(%)	100	78.12
平均正确率	80	

表 5－2 中的"人工判断"是指表 5－1 中正确找出的单(多)主题文章篇数,即:若小主题划分不正确,就无需再判断相应主题段的划分是否正确。

测试过程中出现错误是由于自动分词,词性标注以及特征词抽取错误所致。人工判断也会因人而异,产生主观随意性。

从测试结果中可以看出:小主题划分结果要比相应主题段的划分结果理想。这主要是由于:在本方法中文本全文及文本片断的内容是用所选取的特征词向量来表示的,而特征词的选取结果在表达某些段落的内容时并不准确,例如在某些概括性、总结性的段落中,行文简练,用词较少,相应地包含在特征词向量中的词也不会太多,因此会导致最终的主题段划分不准确。但初步的测试结果:80%的正确率还是比较鼓舞人的,相信随着该方法的进一步完善,会有更满意的结果。

六、结论

有关文本结构的研究对很多领域都是有用的,特别地,通过确定内容一致的小主题来区别单一线索、简单的叙述性文本和结构比较复杂的文本是可能的。我们的研究表明,通过文本结构的自动分析,可以有效地确定自动摘要时抽取句子的区域,这为自动文摘奠定了良好的基础,同时也为自动文摘技术探索了一条新的途径。为了进行自动摘要,对结构较复杂的文本,一个文本主题的确定看起来是必须的。进行自动摘要时,对文中的每一个小主题,应能确定表达它的那几个段落,从所选段落中抽取句子来构成摘要中该小主题的表达,这样生成的摘要条理性会更好。我们目前正在继续进行这方面的研究。

参考文献:

[1] Ranald Brandow, Karl Mitze & Lisa F. Rau(1995). Automatic Condensation of Electronic Publications by Sentence Selection, Information Processing and Management, Vol. 31 ,No. 5,PP. 675－－－685. 1995.

[2]李俊杰(1995)。非受限域中文自动文摘系统的研究与实现,哈尔滨工业大学工学博士学位论文。

[3] Gerard Salton, James Allan & Amit Singhal(1996). Automatic Text Decomposition and Structuring, Information Processing and Management, Vol. 32, No. 2, PP. 127－－－138. 1996.

[4] Salton G cedl(1971). The Smart retrieval system exeriments in Automatic Document processing. Englewood Cliffs NJ: Prentice—Hall.

[5] Richard F. E. Sutcliffe(1991). Distrbuted Representations in a Text Based Information Retrieval system: a New Way of Usingthe Vector Space Model. Proceeding of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.