一种优化的 AP-CAPSA 中文文本结构分析算法 *

刘亚亮, 卢选民, 冯　莎, 王兴亮
(西北工业大学 电子信息学院, 西安 710129)

摘　要: 从段落向量空间模型、特征词权重等多个侧面分析比较了已有的文本结构分析算法,提出了优化的 AP-CAPSA 算法;给出了算法性能的评价准则,并在不同的语料集合上进行了测试。实验表明该方法能够更有效地反映文本的结构,为中文自动摘要系统奠定了重要的理论基础。

关键词: 文本结构; 中文自动摘要; 段落向量空间模型; 特征词权重; 评价准则
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1001-3695(2012)01-0137-03
doi:10.3969/j.issn.1001-3695.2012.01.038

Improved AP-CAPSA Chinese text structure analysis algorithm

LIU Ya-liang, LU Xuan-min, FENG Sha, WANG Xing-liang
(*School of Electronics & Information*, *Northwestern Polytechnical University*, *Xi' an* 710129, *China*)

Abstract: This paper studied the existing text structure analysis methods from several aspects including the paragraphic vector space model and the term weighting. It proposed and illustrated an improved structure analysis algorithm in detail. Also, gave the algorithm performance evaluation criteria. Finally, according to tests of a collection of different corpus, this method can reflect the structure of text more effectively. This lays a good foundation for the research of the Chinese automatic summarization system.

Key words: text structure; Chinese automatic summarization; paragraphic vector space model; term weighing; evaluation criteria

0　引言

面对因特网上快速增长的海量文本数据,自动摘要能够快速、高效、便捷地获取和使用这些数据,使其使用价值得到充分体现。文本结构分析是中文自动摘要中的关键性问题[1]。目前而言,文本结构分析有三种典型的算法,即基于相邻段落相似度的算法[2,3](adjacent paragraphic similarity algorithm, APSA)、基于连续段落相似度的算法[4,5](consecutive paragraphic similarity algorithm, CPSA)、相邻段落相似度和连续段落平均相似度相结合的算法[6](adjacent paragraphic similarity and consecutive average paragraphic similarity algorithm, AP-CAPSA)。

APSA 算法基本思想是相似度低的相邻段落之间存在话题的转移作为子主题划分的边界,它计算简便容易实现,但准确率不高且稳定性低。CPSA 算法将相似度的考察范围扩展到连续多个段落,若某一段与上一子主题的所有段落相似度都较小且与该段后面连续若干段的相似度都较大,则认为该段为子主题的首段;其动态确定子主题数并考虑了子标题段的特殊情况,但是对文本规范度要求较高。AP-CAPSA 算法同时适用于有子标题和无标题组织的文章,考虑了文本有主旨段的情况,使文本结构划分更加合理;但其在建立段落向量空间模型的过程中,对于特征词的段落权重计算考虑了特征词的频率和特征词在文本中的分布情况,但是没有考虑特征词在文本中分布的差异,从而影响了文本结构划分的最终结果。

本文通过改进 AP-CAPSA 算法,在建立段落向量空间模型时,全面考虑了特征词出现的位置对特征词权重的影响和特征词在段落中的相对长度对特征词权重的影响,提出了优化的 AP-CAPSA 算法。从中国知网获取了 170 篇科技论文,人为去掉论文中所有子标题,以最高层次子标题作为子主题的首段而得到多主题实验语料集合,从中文自然处理开放平台获取了 200 篇篇幅较短且单主题文本作为单主题实验语料集合。经过在多主题及单主题语料集合的测试,比较了 AP-CAPSA 及优化的 AP-CAPSA 算法对文本结构分析的效果。

1　优化的 AP-CAPSA 文本结构分析算法

1.1　段落向量空间模型(VSM)

段落向量空间模型是将文章中的一个特征词 T 视为空间中一个维度,并以某种方式为特征词赋权。

假设段落中含有 n 个特征词,记为 $T_1, T_2, \cdots, T_n$(其中 T_k $(k=1,2,\cdots,n)$ 是特征词)。由于特征词的重要程度不同,可以附加权重 W_k 进行量化,那么段落 P 就可以视为 n 维空间中的一个向量 $P(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$,简记为 $\phi(W_1, W_2, \cdots, W_n)$。如果把 $T_1, T_2, \cdots, T_n$ 看成一个 n 维坐标系,则 $W_1, W_2, \cdots, W_n$ 是相应的坐标值,所以可以利用向量间的夹角来衡量两个段落之间的相似性。

设有两个段落 $P_i = (W_{i1}, W_{i2}, \cdots, W_{in})$ 和 $P_j = (W_{j1}, W_{j2},$

收稿日期: 2011-06-14; **修回日期**: 2011-07-22　　**基金项目**: 2011 年西北工业大学研究生创新性实验项目

作者简介: 刘亚亮(1985-),女,陕西兴平人,硕士研究生,主要研究方向为中文文本结构分析(yaliangliu@126.com);卢选民(1972-),男,副教授,博士(后),主要研究方向为智能信息处理、多媒体通信与计算机网络等;冯莎(1989-),女,硕士研究生,主要研究方向为中文文本结构分析;王兴亮(1986-),男,硕士研究生,主要研究方向为中文文本结构分析.

$\cdots, W_{jn})$，用 $\mathrm{sim}(P_i, P_j)$ 表示段落 P_i、P_j 之间的相似度，记向量空间原点为 o，如图 1 所示。

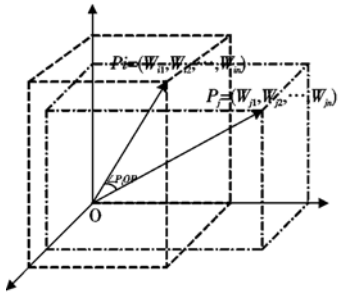

图1　段落向量空间模型

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

$\mathrm{sim}(P_i, P_j)$ 的值越大，则段落 P_i、P_j 之间的相似性越高，相关性也越高。

1.2　基本思想

为了使主旨段单独成段，通过构建段落 VSM，文本首段与全文范围内的段落进行相似度比较。若首段与多个不完全相邻段落的相似度都较大，说明首段与全文所述的话题都有较密切的关系，该段是主旨段，起到总领全文的作用。

为了避免像子标题段或解释性、举例性等段落单独成段的情况，通过构建段落 VSM，按照文本顺序对全文所有相邻段落进行相似度比较，若相似度较高则可认为两段属于同一子主题；在初步得到了所有子主题划分候选点之后，将仅含有一个段落的子主题段分别与前后子主题段进行相似度比较，并将其归入与其平均相似度较大的意义段中。

在计算特征词的段落权重时，由于特征词在文本中的不同位置会代表不同的权重，特征词出现在标题中的重要程度会更高，其权重也应该设置得较高。为了解决特征词在文本中分布的差异，本文提出了一种改进的特征词权重的计算方法，特征词 W_j 的段落权重值的计算式为

$$WP_{ij} = \lambda \times Wf_j \times \log(Wf_{ij} + 1) \times \log(M/M_j + 0.5) + \mathrm{len}(W_j)/\mathrm{len}(P_i)$$

其中：WP_{ij} 表示 W_j 在第 i 个段落中的权重值；λ 为位置加权系数，表示特征词在不同位置段落的加权处理系数；Wf_j 表示 W_j 在文本中出现的频率；Wf_{ij} 表示 W_j 在第 i 个段落中出现的频率；M 表示文本所含段落的数目；M_j 表示包含 W_j 段落的数目；$\mathrm{len}(W_j)$ 表示特征词 W_j 的词语长度；$\mathrm{len}(P_i)$ 表示第 i 个段落的段落长度。

对于位置加权系数 λ 的计算，按照文本段落的不同，一般分为标题、主旨段、结论段这三个位置，分别赋予不同的加权系数。采用各部分的频率与不同位置加权系数的乘积和来表示：

$$\lambda = \lambda_1 \times Wf_t + \lambda_2 \times Wf_m + \lambda_3 \times Wf_c$$

其中：Wf_t、Wf_m、Wf_c 分别表示特征词 W_j 在标题、主旨段和结论段的词频，λ_1、λ_2、λ_3 分别为其加权系数。由此可见，位置加权系数 λ 的设置考虑了特征词出现的位置对特征词权重的影响，$\mathrm{len}(W_j)/\mathrm{len}(P_i)$ 的设置考虑了特征词在段落中的相对长度对特征词权重的影响。

1.3　算法步骤

a）设待处理文本共有 m 个段落，记为 $P_1, P_2, \cdots, P_m$。

b）构建段落 VSM。

c）计算每两个段落之间的相似度 $\mathrm{sim}(P_i, P_j)$（$0 \leqslant i \leqslant m-1$；$0 \leqslant j \leqslant m-1$），得出文本的段落相似度矩阵。

d）逐个找出各个子主题转换的候选点 $P_{k_1}, P_{k_2}, \cdots, P_{k_r}, \cdots, P_{k_R}$，并设定三个阈值 ε_1、ε_2、ε_3，其中 P_{k_r} 满足如下各个条件：

（a）当前子主题首段候选段与相邻的上一段落的相似度小于阈值 ε_1，即

$$\mathrm{sim}(P_{k_r}, P_{(k_r-1)}) < \varepsilon_1$$

（b）当前子主题首段候选段与上一候选子主题所有段落的相似度平均值小于 ε_2，即

$$\sum_{k(r-1) \leqslant i < k_r} \mathrm{sim}(P_i, P_{k_r}) / (k_r - k_{(r-1)}) < \varepsilon_2$$

（c）当前子主题首段候选段的下一段落与上一候选子主题所有段落的相似度平均值小于 ε_3，即

$$\sum_{k(r-1) \leqslant i < k_r} \mathrm{sim}(P_i, P_{(k_r+1)}) / (k_r - k_{(r-1)}) < \varepsilon_3$$

e）判断是否存在子主题候选段仅含有一个段落的情况。若 $P_{k(r+1)} - P_{k_r} = 1$，则存在这种情况，进行如下判断。若这一单独的段落与其上一子主题候选段的平均相似度大于与其下一子主题候选段的平均相似度，则划分到上一子主题中；否则划分到下一子主题中。即若：

$$\sum_{k(r-1) \leqslant i < k_r} \mathrm{sim}(P_i, P_{k_r}) / (k_r - k_{(r-1)}) > \sum_{k(r+1) \leqslant i < k(r+2)} \mathrm{sim}(P_i, P_{k_r}) / (k_{(r+2)} - k_{(r+1)})$$

将段落 $P_{k(r+1)}$ 划分到上一子主题中。

若 $$\sum_{k(r-1) \leqslant i < k_r} \mathrm{sim}(P_i, P_{k_r}) / (k_r - k_{(r-1)}) < \sum_{k(r+1) \leqslant i < k(r+2)} \mathrm{sim}(P_i, P_{k_r}) / (k_{(r+2)} - k_{(r+1)})$$

将段落 $P_{k(r+1)}$ 划分到下一子主题中。

f）判断文本首段是否为主旨段，若文本第一个子主题含有三个及以上的段落 $P_{k_2} - P_1 > 2$ 且文本第一个段落 P_1 在全文范围内与其他段落的相似度平均值相对较大，即

$$\sum_{1 < i \leqslant m} \mathrm{sim}(P_1, P_i)/m > \min(\sum_{1 < i < k_2} \mathrm{sim}(P_1, P_i)/k_2)$$

则将 P_1 独立划分为子主题为主旨段，下一个相邻的段落 P_2 成为新的子主题候选点。

2　评价方法与实验结果

2.1　实验资源

在这个实验中主要通过在两种不同的语料上进行实验，比较 AP-CAPSA 与本文提出的优化的 AP-CAPSA 算法对文本结构分析的结果。实验选取的语料：a）从中国知网上获取了 170 篇科技论文作为实验语料，将论文中所有的子标题人为地去掉，并以最高层次子标题作为子主题的首段；b）从中文自然处理开放平台[7]获取了文本分类语料库，对其中 200 篇篇幅较短且为单主题的语料文本进行测试。

2.2　实验评价标准

在实验中，对于 170 篇科技论文，先识别出各个文章的最高层次子标题，再人为地去掉这些标题，如果文本结构分析的结果中的各子主题首段正好位于原来文本中各最高层次子标题位置相邻下一段，则认为是一个正确的结构分析结果。对于其中 200 篇篇幅较短的单主题文本，若文本结构分析结果有且

只有一个子主题首段,则认为是一个正确的结构分析结果。

通过结构分析准确率 P 和子主题划分准确率 precision、子主题划分召回率 recall 来对两种方法以进行评价,它们的定义如下:

$$\text{结构分析准确率}(P)=\frac{\text{结构分析结果正确的文本数}}{\text{所测试的所有文本数}}$$

$$\text{子主题划分准确率}(\text{precision})=\frac{\text{正确划分出的子主题数}}{\text{划分出的所有子主题数}}$$

$$\text{子主题划分召回率}(\text{recall})=\frac{\text{正确划分出的子主题数}}{\text{答案中的所有子主题数}}$$

2.3　实验数据

将优化前后的 AP-CAPSA 文本结构分析算法在多主题文本和单主题文本两种不同的语料库上进行测试,其实验数据如表 1 所示。

表 1　优化前后的 AP-CAPSA 文本结构分析算法的实验数据

(A 为 AP-CAPSA 算法,B 为优化的 AP-CAPSA 算法)

比较项	语料类型			
	多主题文本		单主题文本	
	A 算法	B 算法	A 算法	B 算法
所测试的所有文本数	170	170	200	200
结构分析结果正确的文本数	154	162	191	197
答案中的所有子主题数	567	567	200	200
划分出的所有子主题数	539	556		
正确划分出的子主题数	521	544	188	195
结构分析准确率 P	0.906	0.952	0.955	0.985
子主题划分准确率 precision	0.967	0.978		
子主题划分召回率 recall	0.919	0.959	0.940	0.975

在多主题语料集合测试中,图 2 为两种结构分析算法的结构分析准确率、子主题划分准确率和子主题划分召回率;在单主题语料集合测试中,图 3 为两种结构分析算法的结构分析准确率、子主题划分召回率。

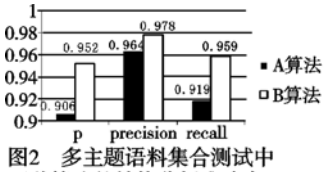

图2　多主题语料集合测试中两种算法的结构分析准确率、子主题划分准确率和召回率

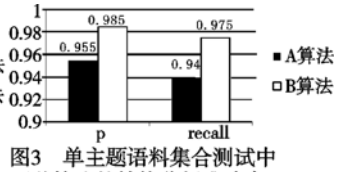

图3　单主题语料集合测试中两种算法的结构分析准确率、子主题划分召回率

2.4　实验分析

通过实验数据可以看出,优化的 AP-CAPSA 文本结构分析算法得到的结果在结构分析准确率、子主题划分准确率、子主题划分召回率均优于 AP-CAPSA 文本结构分析算法,并且该方法同时适用于单主题和多主题的文本,因此本文提出的优化的 AP-CAPSA 文本结构分析算法能够更有效地反映文本结构。

3　结束语

本文通过分析三种已有的典型的文本结构分析算法,提出了优化的 AP-CAPSA 分析算法。此算法在 AP-CAPSA 算法基础上,充分考虑了特征词在文本中分布位置对其权重的影响及特征词在段落中相对长度对其权重的影响。实验结果表明,该算法对于文本结构分析的效果令人满意,并适用于单个主题和多个主题的文章,这为中文自动摘要系统的后续工作奠定了良好的理论基础。

参考文献:

[1] CAI J S, CHI C Y, ZHAN X G. Query-biased summarization based on slide windows [J]. Computer Engineering, 2007, 33(6): 215-238.

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

[3] LUHN H P. The automatic creation of literature abstract [J]. IBM Journal of Research and Development, 1958, 2(2):159-165.

[4] 傅间莲,陈群秀.自动文摘系统中的主题划分问题研究[J].中文信息学报,2005,19(6):28-35.

[5] 傅间莲,陈群秀.基于连续段落相似度的主题划分算法[J].计算机应用,2005,25(9):2022-2024.

[6] 朱荷香,曲维光,卢俊之,等.面向自动文摘的文本结构划分[J].南京大学学报:自然科学,2008,44(2):204-211.

[7] 中文自然语言处理开发平台[DB/OL]. http://www.nlp.org.cn.

[8] WAN Min, LUO Zheng-sheng, GUO Yu-qing. Study on semantic paragraph partition in automatic abstracting system[C]//Proc of 2001 IEEE International Conference on Systems, Man, and Cybernetics. 2001:892-897.

[9] WAN Min, LUO Zheng-sheng. Study on topic segmenting method in automatic abstracting system[C]//Proc of 2003 IEEE International Conference on Nataral Language Processing and Knawledge Engineering. 2003:734-739.

(上接第 106 页)强度,为复杂战场环境的路线图构建提供了可能。由于提高了采样效率,进一步满足了无人机规划的实时性要求。

参考文献:

[1] KAVRAKI L, LATOMBE J. Randomized preprocessing of configuration space for fast path planning[C]//Proc of IEEE International Conference on Robotics and Automation. 1994:3020-3026.

[2] 陈岩,苏菲,沈林成.概率地图 UAV 航线规划的改进型蚁群算法[J].系统仿真学报,2009,21(6):1658-1662.

[3] BOOR V, OVERMARS M, Van der STAPPEN F. The Gaussian sampling strategy for probabilistic roadmap planners[C]//Proc of IEEE International Conference on Robotics and Automation. New York: IEEE Prass, 1999:1018-1023.

[4] HSU D, JIANG T, REIF J, *et al*. The bridge test for sampling narrow passages with probabilistic roadmap planners[C]//Proc of IEEE International Conference on Robotics and Automation. New York: IEEE Press, 2003:4420-4426.

[5] HSU D, SANCHEZ-ANTE G, CHENG Ho-lun. Multi-level free-space dilation for sampling narrow passages in PRM planning[C]//Proc of IEEE International Conference on Robotics and Automation. [S.l.]: IEEE Press, 2006:1255-1260.

[6] FLAVIGNE D, TAIX M. Improving motion planning in weakly connected configuration spaces[C]//Proc of IEEE International Conference on Intelligent Robot and Systems. 2010:5900-5905.

[7] PARK B, CHUNG W K. Adaptive node sampling method for probabilistic roadmap planners[C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, NJ: IEEE Press, 2009:4399-4405.

[8] 殷路,尹怡欣.基于动态人工势场法的路径规划仿真研究[J].系统仿真学报,2009,21(11):3325-3328.

[9] 孙汉昌,朱华勇.基于概率地图方法的无人机路径规划研究[J].系统仿真学报,2006,18(11):3050-3054.