文章编号:1001 - 9081(2005)09 - 2022 - 03

基于连续段落相似度的主题划分算法

傅间莲,陈群秀
(清华大学 计算机科学与技术系 智能技术与系统国家重点实验室,北京 100084)
(gzcfu@163.com)

摘 要:主题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。文中提出了一个通过建立段落向量空间模型,根据连续段落相似度进行文本主题划分的算法,解决了文章的篇章结构分析问题,使得多主题文章的文摘更具内容全面性与结构平衡性。实验结果表明,该算法对多主题文章的主题划分准确率为 92.4%,对单主题文章的主题划分准确率为 99.1%。

关键词:自动文摘; 向量空间模型; 段落相似度; 主题划分
中图分类号: TP391.1 **文献标识码**:A

Study on topic partition based on sequential paragraphic similarity

FU Jian-lian, CHEN Qun-xiu
(*State Key Lab of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Topic partition is a significant problem during text structuring in automatic abstracting system. VSM was established for the whole article based on paragraph, and then algorithms for multi-topic text partitioning based on sequential paragraphic similarity were proposed. It solved the problem of chapter structural analysis in multi-topic article and made the abstract of the multi-topic to have more general content and more balanced structure. Experiments on close test show that the precision of topic partition for multi-topic text and single-topic text reaches 92.4% and 99.1% respectively.

Key words: automatic abstraction; VSM; paragraphic similarity; topic partition

0 引言

自动文摘,是利用计算机自动地从自然语言电子文本中提取重要的文字内容,生成一篇语义连贯的能涵盖或索引原文核心内容的文摘。根据 1995 年自动文摘测试大纲的要求,自动文摘应具有概况性、客观性、可理解性和可读性[1]。

一篇优秀的文摘应该将原始文献的主要信息全面地反映给读者,使读者不需查阅原文就可以获得有用的信息。而很多文章(例如政府工作报告等)是多主题文章,往往从几个不同的方面和角度进行论述,若抽取文摘时只从句子重要度从高到低抽取,则容易造成对次重要主题的遗漏或忽略,完整性差,因此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,即文章由几个相对独立的部分组成;(2) 各段落所属主题;(3)各主题或段落之间的相关程度。[3]

在对文章文本结构进行分析的过程中,研究主题划分及主题之间的联系是一个很重要的内容。所谓主题,是介于篇章与段落之间的一个语言单位,一个主题表达或阐述一个相对独立的意义或话题,从形式上由文章的若干个相邻的自然段组成,各个主题相连构成整个篇章。正确地对文章进行主题划分,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。

1 基于连续段落相似度的主题划分算法

该算法选择向量空间模型 VSM 实现对篇章结构的自动分析和主题划分。

所谓 VSM,是将文章中的一个词视为空间中的一个维度[4,5],这样段落 P 就可视为 n 维空间中的一个向量 $P(W_1, W_2, \cdots, W_n)$,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:$P_i = (W_{i1}, W_{i2}, \cdots, W_{in})$,$P_j = (W_{j1}, W_{j2}, \cdots, W_{jn})$,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:

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

1.1 基本思想

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及内容的一个较小范围,具有一定的重复性[6]。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,表现为这两个段落有较小的夹角,即相似度较大,可初步认为两段谈的是同一主题,即应划在同一个意义段中[7,8]。而不同主题的段落所含词语尤其是高频词一般并不很相同,通常表现为主题划分段与上一个主题的所有段落的夹角都较大,也即相似度都较小,同时与同一主题内的后面连续的若干段的夹角则较小,也即相似度较大。

收稿日期:2005 - 03 - 02

作者简介:傅间莲(1972 -),女,广东南海人,讲师,硕士研究生,主要研究方向:信息提取; 陈群秀(1947 -),女,江西南昌人,教授,主要研究方向:自然语言理解、机器翻译、信息检索、信息提取、机器词典.

基于这一假设,我们通过构建段落 VSM 模型,对全文所有段落与其他若干个连续的段落进行段落相似度比较,若某个段落与前面连续的若干个段落相似度都较小而与后面连续的若干个段落相似度都较大,则初步认为该段是主题划分段。另外,还有一种特殊情况,就是若某一段同时与前面和后面连续若干个段落的相似度都较小,则应该再看该段的下一段与其后面的连续若干个段落的相似度是否都较大,若是,则把该段也当作主题划分段。因为此段极有可能是标题段,而标题段含有的词汇少,通常与其他段落的相似度都较小。

假设全文共 m 段,记为 $P_0,P_1,\cdots,P_{m-1}$。统计词频,去掉低频词和禁用词后考察的特征词共有 n 个,记为 $T_0,T_1,\cdots,T_{n-1}$,则构成一个 n 维的向量空间。计算各特征词 T_k 的权重 W_k,则文章中的段落可形式化为 $P(T_1,W_1,T_2,W_2,\cdots,T_n,W_n)$。然后对每两个段落 P_i,P_j:$(0 \leqslant i,j \leqslant m-1)$ 利用公式 1 计算出段落相似度 $Sim(P_i,P_j)$。

为了直观地考察文章各段落的联系情况,以矩阵形式列出文章各段落相似度是一个较好的办法,由于每个段落与自己的夹角为 0,即相似度为 1,故根据上面相似度计算得到的是一个主对角线为 1 的对称矩阵,其中第 i 行第 j 列上的值代表 $Sim(P_i,P_j)$ 且 $Sim(P_i,P_j)$ 等于 $Sim(P_j,P_i)$。

表 1 是由两篇标题分别为"第七届全国美展中国画获奖作者新作展在京开幕"和"常世琪微雕艺术作品赴台展出"的文章合在一起所生成长文章计算所得的段落相似度矩阵。全文共 8 段,描述两个相近但不同的主题,其中第 $P_0 \sim P_2$ 描述第一个主题,$P_3 \sim P_7$ 描述第二个主题。由于段落相似度矩阵是一个对称矩阵,所以可只列出矩阵的上三角阵。

表 1　段落相似度矩阵实例

	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7
P_0	1.0	0.98	0.49	0.0	0.08	0.0	0.10	0.06
P_1		1.0	0.49	0.12	0.15	0.10	0.37	0.21
P_2			1.0	0.08	0.11	0.18	0.09	0.10
P_3				1.0	0.33	0.48	0.17	0.52
P_4					1.0	0.69	0.39	0.77
P_5						1.0	0.49	0.67
P_6							1.0	0.55
P_7								1.0

由前述,从表 1 中可以看到:

(1) P_3 与前面的 $P_0 \sim P_2$ 的相似度分别为 0.0、0.12、0.08,其值都较小,而 P_3 与后面的 $P_4 \sim P_7$ 的相似度则较大,这说明 P_3 与前面连续 3 个段落的关联度都较小,而与后面连续 4 个段落的关联则都较大;

(2) 矩阵中二个带阴影的区域数值都较大,而不带阴影的区域数值则较小。前者表明从 P_0 段到 P_2 段之间较相似,从 P_3 段到 P_7 段亦如此,后者则表明这二个部分之间的联系较小。

故 P_3 是主题划分段,这与原文相吻合。

1.2　算法步骤

根据前述设计出一个基于 VSM 的连续段落相似度主题划分算法如下:

(1) 设待处理文本共有 m 个段落,$P_0,P_1,\cdots,P_{m-1}$,构建文章 VSM;

(2) 设定阈值 ε_1、ε_2 和 ε_3(取 $\varepsilon_1 = 0.20,\varepsilon_2 = 0.30,\varepsilon_3 = 0.40$);

(3) 计算每两个段落之间的相似度 $Sim(P_i,P_j)(0 \leqslant i, j \leqslant m-1)$,得出文章的段落相似度矩阵;

(4) 逐个找出主题转换的候选点 $P_{k_1},P_{k_2},\cdots,P_{k_r},\cdots,P_{k_R}$,其中 P_{k_r} 满足如下三个条件:

① $\sum_{k_{r-1}\leqslant i\leqslant k_r} Sim(P_i,P_{k_r})/(k_r - k_{r-1}) < \varepsilon_1$

② $Sim(P_{k_r},P_{k_{r-1}}) < \varepsilon_2$

③ $k_r - k_{r-1} > 1$

上述各式中 k_r 是第 r 个主题划分候选点的段落下标,其中:条件 ① 表示当前主题划分候选点与从上一个主题划分候选点至上一个相邻段的段落相似度的平均值要小于阈值 ε_1;条件 ② 表示当前主题划分候选点与上一个主题划分候选点的段落相似度要小于阈值 ε_2;条件 ③ 表示当前主题划分候选点与上一个主题划分候选点至少要间隔一个段落。

(5) 对主题划分候选点 P_{k_r},若 P_{k_r} 满足条件:

$$\sum_{k_r<i<k_{r+1}} Sim(P_{k_r},P_i)/(k_{r+1} - k_r - 1) > \varepsilon_3$$

则 P_{k_r} 为主题划分段,继续处理下一个候选点,若全部主题划分候选点处理完毕则结束,若 P_{k_r} 不满足上述条件则转(6)。

上面的条件表示当前主题划分段与从下一个相邻段至下一个主题划分候选段的上一个相邻段的段落相似度的平均值要大于阈值 ε_3。

(6) 若该主题划分候选点的下一段落 P_{k_r+1} 满足条件:

$$\sum_{k_r+1<i<k_{r+1}} Sim(P_{k_r+1},P_i)/(k_{r+1} - k_r - 2) > \varepsilon_3$$

则 P_{k_r} 为主题划分段,否则为非主题划分段。返回(6)继续处理下一个主题划分候选点,若全部主题划分候选点处理完毕则结束。

上面的条件表示当前主题划分段的下一个段落要与从其下一个相邻段至下一个主题划分候选段的上一个相邻段的段落相似度的平均值要大于阈值 ε_3。

2　实验与结果分析

2.1　实验资源

为了测试本文提出的主题划分算法的准确性,最好是能有大量的经专家进行了主题划分的语料用于测试。但是目前这方面的语料还较为缺乏,而且主题的划分标准也较难制定。文章的主题是一个较为主观的概念,并没有一个客观的判断标准。同一篇文章如果没有各子标题的指引,不同人也可能会有不同的主题划分方法,得到的结果中主题数目与主题的分界都有可能不完全相同。

为了克服上面所说的困难,我们采用了一个折中的办法来进行主题划分算法的测试。从语料库中随机地抽取一些篇幅较短小的文章组成一篇长的文章。由于每篇文章的篇幅较短,所以可以假定在一定的语义层次上,每一篇文章都属于同一个主题,而不同的文章则表述不同的内容,分属不同的主题。于是对这些生造的长文章,其主题划分就有了一个可操作的标准。另外为了检测主题相近的文章的主题划分准确性,还应该构造一些主题较相近的长文章。

从中文自然语言处理开放平台(http://www.nlp.org.cn)取得了文本分类语料库测试语料,按上述原则随机取了 2 000 篇含有教育、艺术、通信、电子、航天航空、历史、文学、经济、体

育、交通等不同类型的文章,组成了 600 篇长文章,每篇文章含 2 到 5 篇或相同或不相同类型的短文,每篇短文章的篇幅不超过 10 个段落。另外各取了 120 篇教育、120 篇政治、120 篇艺术、120 篇农业类的文章各自组成了 40 篇共 160 篇长文章,每篇文章都由 3 篇主题相近的短文组成。最后为了测试该算法对只有单个主题的文章是否有影响,还随机对 1 000 篇原语料文章进行主题划分测试。测试文章的组成结构如表 2 所示。

表 2 测试文章的组成结构

	多主题文章数/篇	单主题文章数/篇	文章总数/篇
原语料文章数	2480	1000	3480
生成文章数	760	1000	1760

2.2 实验评价方法

如果文章划出的主题分界处正好位于各短文章的交界处,则认为是一个正确的划分,否则认为是一个错误的划分。我们定义主题划分的准确率为:

$$p = \frac{n}{N} \times 100\% \qquad (2)$$

上式中 p 为准确率,n 为划分结果准确的文章数,N 为进行测试的总文章数。

2.3 实验结果

表 3 实验结果

实验结果	多主题文章		单个主题文章	
	文章数/篇	所占比例	文章数/篇	所占比例
主题划分正确	702	92.4%	991	99.1%
主题个数正确但有一划分相差一段	21	2.7%		
多划分一个主题	8	1.1%	9	0.9%
少划分一个主题	22	2.9%		
把全文当作一个主题	7	0.9%		
总和	760	100%	1000	100%

最后得到的测试结果:(1) 对 760 篇生造的多主题文章的主题划分中有 702 篇文章划分正确,有 21 篇主题个数划分正确但其中一个划分相差一段,有 8 篇和 22 篇分别多划分和少划分一个主题,还有 7 篇则认为全文是一个主题,且这 7 篇都是由主题相近的短文章生成的长文章;(2)对 1 000 篇原语料文章的主题划分结果中有9篇将文章划分为两个主题,而其余 991 篇划分正确即认为文章只有一个主题。结果如表 3 所示。

2.4 结果分析

从表 3 可以看出,基于连续段落相似度的主题划分算法对于多主题文章的准确率为 92.4%,对于单个主题文章的划分准确率为 99.1%,另外还有相当一部分的错误与正确结果之间只相差 1 个段落,这种结果虽不精确,但主题划分只用在对文摘长度的具体分配上,对最终文摘句的结果影响不大。而且该算法也适用于无标题组织结构的文章主题划分。另外本算法的计算复杂度和空间复杂度都较小,所以采用基于连续段落相似度的主题划分算法的效果对于用于自动文摘的主题划分步骤是令人满意的。

3 结语

本文提出了一种通过构建段落 VSM,根据连续的段落相似度对文本进行主题划分的算法。实验结果表明,该方法对于文章主题的划分效果令人满意,而且适用于有标题组织和无标题组织的文章,也适用于单个主题和多个主题的文章,为自动文摘系统的后续工作铺垫了良好的基础,使得多主题文章的文摘更具内容全面性与结构平衡性。

参考文献:

[1] 俞士汶,段慧明. 自动文摘评测报告[N]. 计算机世界报,1996-03-25.183.

[2] LUHN HP. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159 - 165.

[3] EDMUNSON HP. Problems of automatic abstracting[J]. Communications of ACM, 1964, 7(4): 259 - 263.

[4] SALTON G. A Blueprint for Automatic Indexing[J]. SIGIR Forum, 1981, 16(2).

[5] SALTON G, MCGILL M. Introduction to Modern Information Retrieval[M]. New York: McGraw-Hill, 1983.

[6] SALTON G, SINGHAL A, MITRA M. Automatic text structing and summarization[J]. Information Processing& Management, 1997, 33(2).

[7] WAN M, LUO ZS. Study On Topic Segments Method in Automatic Abstracting System[A]. Natural Language Processing and Knowledge Engineering(NLPKE) Mini Symposium of the 2001 IEEE Internation Conference on Systems, Man, and Cybernetics(SMC2003)[C], 2003.

[8] 万敏. 基于统计和语义分析的中英文自动文摘研究[D]. 北京:清华大学, 2003.

投稿注意事项

1. 本刊只受理 E-mail 投稿(先杀毒再传送),投稿邮箱为:bjb@computerapplications.com.cn。请在主题栏注明"新投稿"字样,如是修改稿请在主题栏写上稿件编号。稿件请用 Word 编排,并加上作者信息表(下表为 Word 表格,第 3 ~5 项不填写),以附件形式发送。一封邮件控制在 2M 内,同一篇稿件请不要反复发送。

2. 本刊收到作者稿件后,将在 3 ~4 个工作日内反馈给作者带有文章编号的回执(通过 E-mail 发送)。作者收到后,请在 30 天内办理相关手续,逾期视稿件已被作者撤销,不再安排审读。

3. 本刊实行三审制,审稿周期为 3 个月。无论稿件是否被录用,作者都将及时得到通知(E-mail)。

4. 编辑部将给刊用稿件发刊用通知(书面和 E-mail 两种形式)。收到刊用通知后,请作者速按通知要求,在限定时间内办理回执等相关手续。稿件将在投稿后的半年左右发表。

5. 本刊已整体进入《中国科学引文数据库》、《中国核心期刊(遴选)数据库》、《中国科技期刊数据库》、《中国学术期刊综合评价数据》、《中国科技期刊精品数据库》、《中国期刊全文数据库(CJFD)》,并在《万方数据数字化期刊群》全文上网,被《中国期刊网》、《中国学术期刊(光盘版)》、《CEPS 中文电子期刊服务》全文收录。

我刊专业技术领域划分(投稿方向)如下:

①数据库、数据仓库、数据挖掘;②专家系统、自动推理、模式识别;③多媒体、流媒体、超媒体技术;④网络结构、网络协议、网络管理;⑤移动计算、Agent 技术;⑥信息安全(加密、认证、检测);⑦CAD、CAE、CAM、CAPP 技术;⑧软件开发方法(软件模式、组件、中间件);⑨图形图像处理;⑩虚拟现实、计算机仿真;⑪电子商务、电子政务、电子金融;⑫操作系统、嵌入式技术;⑬现场总线、工业检测与控制;⑭先进制造技术、工作流技术;⑮无线与移动通信。

第一作者	其他作者			联系地址	邮编	省市	电话	E-mail	稿件名称	投稿方向
张三	王二,赵五			××大学××学院××系××信箱	610041	四川成都	028-85224283	bjb@computerapplicaitons.com.cn	××××	①