基于主题分析的文本分割技术研究

刘 铭,王晓龙,刘远超
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 本文提出一种新颖的文本分割算法,算法首先将待分割文档划分为若干片段的集合,然后构造全文词汇链分析文中描述的多个子主题,并通过构造片段对子主题的覆盖图将描述相同子主题的相似片段归类.针对段落分割点可能落在片段内部的情况,算法对片段进行二次划分.实验表明:在对文档进行主题分析后,算法能够过滤掉与主题无关的特征对分割结果的干扰;构造的片段对子主题的覆盖图融合了相邻及相间片段的相似性,加大了划分的准确度;对片段进行二次划分使得分割的结果更加合理.

关键词: 主题分析; 词汇链; 知网; 二次划分
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2009) 02-0278-07

Research on Text Segmentation Based on Topic Analysis

LIU Ming, WANG Xiao-long, LIU Yuan-chao
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: A novel topic segmentation algorithm is proposed in this paper. This algorithm first partitions text into some blocks. After that it constructs whole-length lexical chains to analyze multiple subtopics of this text. By constructing graph which describes blocks covering subtopics, the similar blocks which describe same subtopic can be classified. In order to solve the situations that segmentation points drop inside blocks, it segments blocks again. Experiment results demonstrate that by analyzing topic of text, this algorithm can remove interferences, which are aroused by irrelative features, from segmentation results. By constructing graph which describes blocks covering subtopics, it can mix similarities of adjacent and disconnected blocks together, and increases segmentation precision. The second segmentation makes segmentation results more reasonable.

Key words: topic analysis; lexical chain; HowNet; second segmentation

1 引言

文本分割是按照文档主题叙述的线性变换将文档划分成为若干个语义片段或分割单元,以形成语义片段序列的过程,其能够使不同的分割单元描述不同的子主题信息[1].文本分割的应用范围相当广泛,如自动文摘、问答检索等系统中文本分割均起到很大的作用[2,3].

目前无论是基于局部片段相似性的文本分割算法[2,3]还是基于全局片段相似性的文本分割算法[4,5]均以文中所有词作为片段相似度计算的依据.此方法非常容易引入与文档描述的主题无关的噪声词,而增大或缩小描述不同或相同子主题的片段间的相似度,使得片段划分的结果不准确.同时由于上述算法没有对文档主题进行分析,很可能将描述相同子主题的片段分裂到不同的分割单元中.针对上述问题,本文提出一种基于主题分析的文本分割算法-LexicalSeg.算法通过分析文档主题去除与主题信息无关的词语对片段相似度计算的干扰.同时将图的思想引入到文本分割中,使分割转化为一种在全局范围内寻找连通分量的过程,结合全局和局部特征寻找片段划分的最优值,增加了划分的可信度.

2 义类获取及词汇链构造

词汇链是 1991 年由 Hirst 首先提出的,以相关或相似的词语构成的一条链.词汇链与文本结构有一种对应关系,提供了关于文档主题的重要线索[6].由于词汇链是具有相关含义的词语的集合体,因此在构建词汇链时需要知道词在文中某个上下文中的确切含义.本文以“知网”(HowNet)[7]作为词义获取和词汇链构建的语义词典.“知网”是由董振东博士完成的中文语义辞典,其定义了 1500 多个义原,并通过义原反映中文词义.“知网”中的每个词条均含有 NO.、W_C、G_C、E_C、DEF,其中 NO.为概念编号,剩下的符号分别对应于词条的词语、词性和例子.DEF 是概念定义项,表达了词条的语义信息,由两部分组成,如“气”(NO. = 084195)的 DEF 为

收稿日期:2007-12-10;修回日期:2008-09-16
基金项目:国家自然科学基金重点项目(No. 60435020);国家 863 高技术研究发展计划(No. 2006AA01Z197, No. 2007AA01Z172)

{Strength| 力量:host = {human| 人}}.以冒号为界,第一部分为基本义原:{Strength| 力量},能够很大程度上反映 DEF 的含义;第二部分为关系义原:{host = {human| 人}},代表了 DEF 的关系结构特性,如 host = {human| 人}表明该词的宿主为人."知网"以树形组织义原,越相似的义原在义原树内的位置越接近.

2.1　义类获取

由上文得知"知网"将词的词义表示为 DEF,但是观察"知网"可知,"知网"中对于词义(DEF)的区分过于严格,并且每个词的 DEF 中的基本义原在很大程度上决定了该词的含义,至少对于本文的应用来说这个结论是成立的.因此本文将"知网"中每个词条的 DEF 集合划分为多个子集,使每个子集仅包含基本义原相同的 DEF,并将此子集视为该词的一个义类.将每个义类表示为两部分:一部分是该义类对应的基本义原,另一部分是该义类的每个 DEF 所包含的关系义原的并集.

2.2　词汇链构造

对文档进行分词及停用词过滤后即可获得文档的词空间,本文即通过计算词空间中的词语之间反映信息的相似程度将相似或相关的词构成一条词汇链,下面即详细介绍了词汇链的构造过程:

(1) 过滤掉"知网"中含义过大、过于宽泛的抽象义原,如"属性"、"事件"、"实体"等;

(2) 设待分割文档 Doc 的词空间为 $WordSet$,词汇链集合为 L;

(3) 顺序扫描 $WordSet$,设当前正在扫描的词为 W_i,该词具有 t 个义类;

(4) 顺序扫描 W_i 的 t 个义类,设当前正在扫描第 j 个义类,记其为 W_{ij};

(5) 按式(1)计算 W_{ij} 与 L 中每条链的关联关系,并找到与 W_{ij} 具有最大关联关系的链,设此词汇链为 L_m;

(6) 按式(2)计算 W_{ij} 与 L_m 的 $Select$ 值,如值为 1,则将词 W_i 插入到词汇链 L_m 中,并标记 W_i 在 L_m 中的义类为 W_{ij},否则新建一条词汇链包含 W_i,同时标记 W_i 在新建链中的义类为 W_{ij};

(7) 如果 $j \geq t$,即 W_{ij} 为词 W_i 的最后一个义类则转(8),否则循环步骤(4)～(7);

(8) 如果词 W_i 为 $WordSet$ 的末尾则转(9),否则循环步骤(3)～(8);

(9) 计算 L 中每条词汇链的权值,权值为词汇链包含的词数;

(10) 取大于平均链权重的词汇链作为文档 Doc 描述的子主题的代表,并记这些大于平均链权重的词汇链为强链;

$$R(W_{ij}, L_k) = \max\left(Sim(W_{ij}, L_k), Cor(W_{ij}, L_k)\right) \quad (1)$$

式(1)中 $Sim(W_{ij}, L_k)$ 反映的是义类 W_{ij} 和词汇链 L_k 所描述的信息之间的相似性. $Cor(W_{ij}, L_k)$ 反映的是 W_{ij} 和 L_k 所描述的信息之间的相关性[8].

$$Select(W_{ij}, L_m) = \begin{cases} 1; \text{ if } R(W_{ij}, L_m) \geq TH \\ 0; \text{ else} \end{cases} \quad (2)$$

式(2)中 TH 为义类 W_{ij} 与词汇链 L_m 反映的信息之间是否相互关联的阈值.由式(3)和式(9)可见:我们以词与词汇链包含的所有词的最大相似度和最大相关度作为词与链的关联度.因此词与词汇链之间的关联度阈值也就是词与词之间的关联度阈值.实验中发现,如果词与词之间的关联度超过 0.7,则两个词较为相似,因此本文设 TH 为 0.7.

$$Sim(W_{ij}, L_k) = \max_{p=1}^{|L_k|}\left(SimWord(W_{ij}, LW_{kp})\right) \quad (3)$$

式(3)计算了义类 W_{ij} 与词汇链 L_k 反映的信息之间的相似度.由于在词汇链构造过程中已经标记了 L_k 中的每个词在链中对应的义类,因此我们以 LW_{kp} 代表链 L_k 中的第 p 个词在链中对应的义类,并以 W_{ij} 与 L_k 中所有词的最大相似度作为 W_{ij} 与 L_k 的相似度[9].其中 $|L_k|$ 为词汇链 L_k 包含的词数.

$$SimWord(W_{ij}, LW_{kp}) = \lambda \times BSim(W_{ij}, LW_{kp}) + (1 - \lambda) \times RSim(W_{ij}, LW_{kp}) \quad (4)$$

式(4)计算了义类 W_{ij} 和义类 LW_{kp} 的相似度.以加号为界,此式分为两部分:第一部分计算了两个义类中的基本义原的相似程度,如式(5)所示;第二部分计算了两个义类中的关系义原的相似程度,如式(8)所示.参数 λ 对应于两部分的重要性,由于基本义原较能反映词的主要信息,因此 λ 的设置偏重于第一部分,本文设 λ 为 0.6.

$$BSim(W_{ij}, LW_{kp}) = DL(W_{ij}, LW_{kp}) \times PS(W_{ij}, LW_{kp}) \quad (5)$$

式(5)计算了两个义类中的基本义原的相似程度.如文献[7]所述,"知网"将义原组织为树形结构,相似的义原在义原树内有较短的距离.因此可以通过计算两个义原在义原树内的位置确定两个义原的相似度.以乘号为界,此公式的第一部分反映了两个基本义原是否在同一义原树内,第二部分反映了两个基本义原在义原树内的位置关系.

$$DL(W_{ij}, LW_{kp}) = \begin{cases} 0; \text{ if } W_{ij} \text{和} LW_{kp} \text{的基本义原不在同一义原树内} \\ \max\left(Layer(W_{ij}), Layer(LW_{kp})\right); \text{else} \end{cases} \quad (6)$$

其中 $Layer(W_{ij})$ 为 W_{ij} 的基本义原在义原树内的层号.

式(7)计算了同一义原树内的两个基本义原的位置关系.本文以"知网"中的义原层次与义原相似度的相互关系为基础对位于不同层次的义原间的相似度进

行赋值[7],其中 $interval(W_{ij}, LW_{kp})$ 指义类 W_{ij} 和 LW_{kp} 的基本义原在义原树内的层次间隔.

$$PS(W_{ij}, LW_{kp}) = \begin{cases} 0; \text{if } interval(W_{ij}, LW_{kp}) > 2 \\ 0.6; \text{if } interval(W_{ij}, LW_{kp}) = 2 \\ 0.8; \text{if } interval(W_{ij}, LW_{kp}) = 1 \\ 1.0; \text{if } interval(W_{ij}, LW_{kp}) = 0 \end{cases} \quad (7)$$

式(8)计算了两个义类中的关系义原的相似程度.由于关系义原在很大程度上反映了"知网"中 DEF 的关系结构特性,因此该公式为两个义类的结构相似性.

$$RSim(W_{ij}, LW_{kp}) = \frac{IS(W_{ij}, LW_{kp})}{RC(W_{ij}) + RC(LW_{kp})} \quad (8)$$

$IS(W_{ij}, LW_{kp})$ 为义类 W_{ij} 和 LW_{kp} 的关系义原集合的交集大小. $RC(W_{ij})$ 为义类 W_{ij} 具有的关系义原总数.

式(9)计算了义类 W_{ij} 与词汇链 L_k 反映的信息之间的相关度.类似于相似度,我们以 W_{ij} 与 L_k 中所有词的最大相关度作为 W_{ij} 与 L_k 的相关度.

$$Cor(W_{ij}, L_k) = \max_{p=1}^{|L_k|}(CorWord(W_{ij}, LW_{kp})) \quad (9)$$

其中各符号的含义可以参见式(3).

$$CorWord(W_{ij}, LW_{kp}) = \frac{I(W_{ij}, LW_{kp}) + I(LW_{kp}, W_{ij})}{RC(W_{ij}) + RC(LW_{kp})} \quad (10)$$

式(10)中 $I(W_{ij}, LW_{kp})$ 指义类 W_{ij} 的关系义原集合中是否包含义类 LW_{kp} 的基本义原,如包含值为 1,否则为 0.由于基本义原代表了义类的主要信息,关系义原代表了义类的关系特性,因此 $I(W_{ij}, LW_{kp})$ 能够说明 LW_{kp} 反映的信息是否与 W_{ij} 具有一定的相关性.

3 片段分割

本文介绍的算法 LexicalSeg 首先将待分割文档划分成为固定大小的片段(BLOCK).如果片段的结尾不为复句标点("。"、"!"、";"等),则扩充此片段到最近的复句标点.此方法可以使每个片段的结尾均为有意义的分割点.首先按照文中第 2 章的词汇链构造方法为每个片段建立词汇链集合以反映该片段所描述的信息,然后计算每个片段的词汇链集合与文中强链集合的相交模式,并将结果表示为矩阵 $A = A_{ij}$.矩阵 A 的第 i 个行向量为片段 i 与文中强链集合的相交模式.A 的行为强链集合,列为片段集合,A_{ij} 为待分割文档中的第 i 个片段覆盖第 j 条强链所述信息的程度.由于强链能够在一定程度上反映文中的多个子主题,因此矩阵 A 能够在一定程度上反映片段对子主题的侧重程度.

设待分割文档中的第 i 个片段的词汇链集合为 $BL(i)$,大小为 $|BL(i)|$,其中第 m 条词汇链为 $BL(i)_m$.设文中第 j 条强链为 SL_j.式(11)即为 A_{ij} 的计算方法:

$$A_{ij} = \sum_{m=1}^{|BL(i)|} \frac{LS(BL(i)_m, SL_j)}{|BL(i)|} \quad (11)$$

设词汇链 $BL(i)_m$ 和文中强链 SL_j 中词的并集为 $CWSet$,其大小记为 $|CWSet|$.其中的第 l 个词为 CW_l,$Blockfre(i, CW_l)$ 和 $Articlefre(CW_l)$ 分别为词 CW_l 在文中第 i 个片段和在全文中的词频.如果词 CW_l 在 $BL(i)_m$ 和 SL_j 中均出现,$Same(CW_l)$ 为 1,否则为 0.式(12)为词汇链 $BL(i)_m$ 和强链 SL_j 的相似度:

$$LS(BL(i)_m, SL_j) = \frac{\sum_{l=1}^{|CWSet|} Same(CW_l) \times \frac{Blockfre(i, CW_l)}{Articlefre(CW_l)}}{|CWSet|} \quad (12)$$

以乘号为界,式(12)中分母的第一部分反映了词汇链 $BL(i)_m$ 和 SL_j 中相同词的个数,代表了两条词汇链反映的信息的相似程度.而乘上相同词 CW_l 在片段 i 中的频率与该词在待分割文档中的频率之商,则反映了此相似程度就全文来说的比例.因此该公式能够反映词汇链 $BL(i)_m$ 和 SL_j 所述信息的相似度.

计算任意两个片段 i、j 在矩阵 A 中对应的行向量 A_i 和 A_j 的余弦相似度,以无向图反映此片段相似度.以片段代表图中顶点,以边(弧)代表两个顶点(片段)之间的相似性,边上的权值则反映了此相似性的大小.由于矩阵 A 能够反映片段对子主题的侧重程度,则此图即为片段对子主题的覆盖图,具体图示可参见实验部分图 1、2.

上述片段对子主题的覆盖图中的某些顶点间的相似弧相对于其它顶点间的相似弧是一种弱相关,应该依据阈值予以去掉以减少分割干扰.文献[10]通过分析片段内部以及片段间的词分布情况来确定分割阈值,但是此方法需要计算每种可能的分割情况所对应的不同阈值,计算量非常大.同时待分割文档中的所有词并非都对分割阈值的计算有贡献,那些与文档描述的主题无关的词可能会在阈值计算中引入误差,降低分割结果的准确性.

本文以矩阵 A 作为阈值计算的依据.由本章的起始部分可知,矩阵 A 能够反映片段对文中不同子主题的侧重程度,因此可以通过矩阵 A 计算片段内部以及片段间子主题的分布情况,即片段内距离 BI 和片段间距离 BA,并通过线性回归融合 BI 和 BA 以确定分割阈值[11].

定义片段内距离 BI 为:

$$BI = \sum_{i=1}^{b} \frac{\log_2(P_i + 1)}{b} \quad (13)$$

其中 P_i 为片段 i 在矩阵 A 中对应的行向量的非零列数,反映了片段 i 对文中多个子主题的侧重程度,b 为片段总数.则 BI 反映了各片段对文中多个子主题的侧

重程度的平均值.

定义片段间距离为:

$$BA = \sum_{i=1}^{b} \frac{\| A_i - M \|^2}{b} \quad (14)$$

其中 M 为矩阵 A 中各行向量的平均向量. 则 *BA* 反映了各片段对文档中心的离散程度.

4　片段内部划分

文中第 3 章将片段划分为固定大小,然而片段的大小大多是根据经验值得到的[2],这样某些片段的结尾并不一定是真实的段落分割点,真实的分割点很可能落在算法确定的分割点上下两个片段的内部. 此时我们将片段内部可能的分割点记为疑似分割点. 本文以片段内的复句标点作为疑似分割点,并以这些疑似分割点重新划分片段. 下面即详细叙述了如何寻找某些落在片段内的分割点:

(1) 设分割点 *s* 的上下两个片段分别为 *BU*(*s*) 和 *BD*(*s*),设 *BU*(*s*) 和 *BD*(*s*) 的相似度为 *SimUD*(*s*),设 *BU*(*s*) 和 *BD*(*s*) 内疑似分割点的集合为 *SegUSet*(*s*) 和 *SegDSet*(*s*);

(2) 设片段 *BU*(*s*) 和 *BD*(*s*) 内所有疑似分割点对应的上下两个片段的相似度集合为 *SegSimSet*(*s*);

(3) 顺序扫描 *SegUSet*(*s*),设当前正在扫描的疑似分割点为 $SegU(s)_p$;

(4) 以 $SegU(s)_p$ 作为划分点,将片段 *BU*(*s*) 中位于 $SegU(s)_p$ 以上的内容作为一个单独的片段,将 $SegU(s)_p$ 以下的内容归入到片段 *BD*(*s*) 中,这样即可以形成两个新的片段,将其记为 $BU(s, SegU(s)_p)$ 和 $BD(s, SegU(s)_p)$;

(5) 按文中第 3 章所述方法分别计算 $BU(s, SegU(s)_p)$ 和 $BD(s, SegU(s)_p)$ 与强链集合的相交模式,然后计算这两个片段对应的相交模式的余弦相似度,记此相似度为 $SimUD(s, SegU(s)_p)$,并将其插入到 *SegSimSet*(*s*) 中;

(6) 如果 $SegU(s)_p$ 为 *SegUSet*(*s*) 的末尾则转步骤(7),否则循环步骤(3)～(6);

(7) 按步骤(3)～(6)处理 *SegDSet*(*s*) 中的疑似分割点;

(8) 设 *SegSimSet*(*s*) 中相似度的最小值为 $SegSim(s)_{min}$,检测其是否小于 *SimUD*(*s*),如小于,则以 $SegSim(s)_{min}$ 对应的疑似分割点作为新的片段分割点;

5　实验结果及分析

现实应用中几乎没有标准的中文文本分割测试数据集,这是因为文本分割一般不单独作为独立的系统出现,而多是将其作为系统的一部分用于实际应用中. 同时文本分割算法的评价是一种主观的评测方法,即使对于相同的分割结果,不同的评价标准也会得到不同的评价结果. 针对上述问题,本文以"任常霞先进事迹"、"2008 年奥运会"、"圆明园水污染治理"、"山野菜的种植"为主题,使用搜索引擎 Google 分别检索 100 篇共 400 篇文档作为测试语料,同时为各测试语料的文档人为标定分割结果. 测试算法 LexicalSeg 在上述四种语料集下的准确率(*P*)、召回率(*R*)和 *F* 值,并与 Hearst 提出的 TextTiling[2] 算法和 Reynar 提出的 Dotplot 的改进算法[4]进行比较.

文本分割算法的准确率和召回率类似于信息检索中的准确率和召回率,其中正确识别的分割点为与人为标定的分割点相同的算法返回的分割点.

$$\text{准确率}(P) = \frac{\text{正确识别的分割点数}}{\text{算法返回的分割点数}} \quad (15)$$

$$\text{召回率}(R) = \frac{\text{正确识别的分割点数}}{\text{文中正确的分割点数}} \quad (16)$$

$$F = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \quad (17)$$

传统的准确率和召回率并不能够全面和公正的评价文本分割算法的性能,其原因在于准确率和召回率主要考虑绝对匹配的结果. 实际上,离正确分割点较近的错误分割点比较远的错误分割点的性能更好,但是准确率和召回率却无法体现出这种差别. 为了克服上述缺点,本文同时采用 Beeferman 等人[12]提出的 P_u 评测方法来评价文本分割算法的性能.

P_u 用于计算从文档中随机选取的两个句子,能够被正确识别为属于同一个分割单元,或者分属于不同分割单元的概率. 其定义如下:

$$P_u(ref, hyp) = \sum_{1 \le i \le j \le n} D_u(i, j) \left(\delta_{ref}(i, j) \overline{\oplus} \delta_{hyp}(i, j)\right) \quad (18)$$

其中 *ref* 为标准的分割模式,*hyp* 为算法给出的分割模式,*n* 为待分割文档包含的句子数;$\delta_{ref}(i, j)$ 为指示函数,当句子 *i* 和 *j* 在 *ref* 中属于同一个分割单元时,其值为 1,否则为 0;同理,当句子 *i* 和 *j* 在 *hyp* 中属于同一个分割单元时,$\delta_{hyp}(i, j)$ 为 1,否则为 0. D_u 为随机选取的句子在待分割文档中的距离概率分布函数,其值依赖于参数 *u*. D_u 为参数 *u* 的指数分布函数,其中 r_u 为归一化因子. 本文设参数 1/*u* 为测试文档集中分割单元包含的平均句子数.

$$D_u(i, j) = r_u e^{-u|i-j|} \quad (19)$$

由表 1 可见:当使用词汇链进行主题分析后,算法 LexicalSeg 能够去除与主题无关的词语对分割结果的干扰. 同时根据片段对子主题的覆盖图结合了相邻及相间片段间的信息,对相似片段的分布进行了很好的预

测,其效果比 TextTiling 和 Dotplot 改进算法有了一定的提高.

表 1 分割结果对比

	TextTiling	改进的 Dotplot	LexicalSeg	LexicalSeg (片段内划分)
Data Set 1				
P	0.41	0.37	0.45	0.48
R	0.36	0.38	0.51	0.51
F	0.38	0.37	0.48	0.49
P_u	0.69	0.67	0.75	0.79
Data Set 2				
P	0.45	0.41	0.44	0.46
R	0.32	0.35	0.47	0.50
F	0.37	0.38	0.45	0.48
P_u	0.70	0.65	0.72	0.76
Data Set 3				
P	0.41	0.38	0.47	0.50
R	0.39	0.35	0.46	0.53
F	0.40	0.36	0.46	0.51
P_u	0.72	0.62	0.77	0.83
Data Set 4				
P	0.38	0.39	0.44	0.47
R	0.44	0.42	0.49	0.51
F	0.41	0.40	0.46	0.49
P_u	0.67	0.63	0.74	0.78

以下采用实际文本对算法 LexicalSeg 的运行步骤及结果进行描述,使读者能够更加清晰的了解算法 LexicalSeg 的执行过程及性能. 测试文档的标题为"驻沪海军某部党委机关倾心为基层解难"(http://military.people.com.cn/GB/1076/52965/5269015.html),主要叙述了海军党委对海军官兵学习,工作,生活上的关心.

按固定大小划分得到的片段(BLOCK)结果为:

片段 1 驻沪海军某部党委机关倾心……基层提出的建议项项有答复,件件有着落.

片段 2 新年伊始,记者在……收集整理了 21 条对部队党委机关的意见和建议.

片段 3 面对这些全部来自基层官兵……委婉地向机关提出能否建几个晒衣场.

片段 4 该部闻讯后,不仅相继建设……解决官兵生活上的不便.

片段 5 此问题一经提出,该部后勤……官兵们纷纷拍手叫好.

片段 6 (刘贵朝)、(崔晓龙)1. 遵守……转载或引用.

按文中第 3 章所述方法计算片段对子主题的侧重程度或相交模式,并计算上述片段对应的相交模式的余弦相似度,将各片段间的相似度值列在表 2 的矩阵中.

表 2 片段相似度矩阵

片段	1	2	3	4	5	6
1	1	0.161	0.048	0.020	0.039	0.016
2	0.161	1	0.079	0.135	0.075	0
3	0.048	0.079	1	0.118	0.069	0
4	0.020	0.135	0.118	1	0.104	0
5	0.039	0.075	0.069	0.104	1	0
6	0.016	0	0	0	0	1

以片段(BLOCK)为顶点,以片段间的相似度为边构造片段对子主题的覆盖图,结果如图 1.

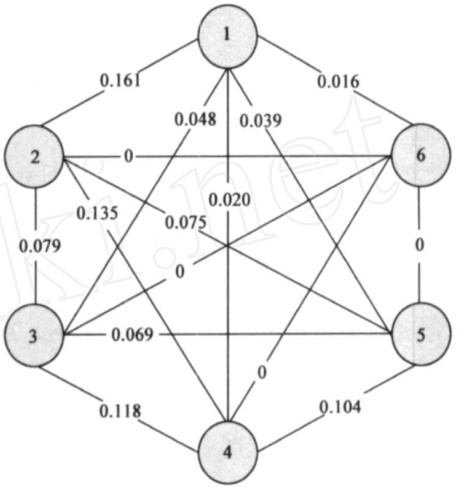


图1 片段对子主题的覆盖图

按文中第 3 章线性融合片段内距离 BI 和片段间距离 BA 后可得相似度阈值为 0.089,去掉相似度小于阈值的边后可得的片段对子主题的覆盖图,结果如图 2.

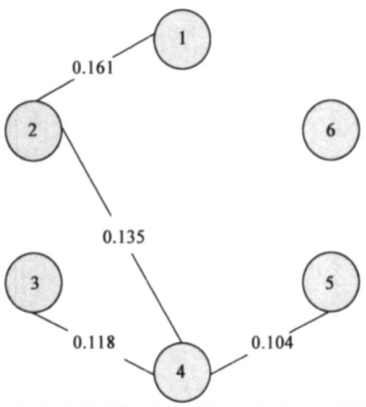


图2 根据阈值去掉弱相关弧后的片段对子主题的覆盖图

按顺序扫描图 2 中的顶点,首先取顶点 1、2,由图 2 可知,顶点 1、2 间有边连接,即 1、2 两个片段描述的信息具有很大的相关性,应该在同一个分割单元内. 继续扫描顶点,取顶点 3,得知顶点 3 与顶点 1、2 间均无边连接,可知片段 3 描述的信息与片段 1、2 描述的信息不相关,虽然顶点 2 与顶点 4 间有边连接,但由于片段 3 叙述的子主题与 1、2 不相同,因此片段 3 为段落分割点,而片段 1、2 为一个分割单元. 沿顶点 3 继续扫描,顶点

3、4 间有边连接,顶点 4、5 间有边连接,顶点 6 与顶点 3、4、5 间无边连接,可知片段 3、4、5 描述的信息具有一定的相关性,应该在同一个分割单元内,片段 6 单独为一个分割单元.测试文档的最终划分结果为 1、2| 3、4、5 | 6.观察每个分割单元的内容可知:片段 1、2 主要介绍了海军党委的政策,片段 3、4、5 为政策的具体实施,片段 6 为人民网公告,其结果有一定的合理性.但是观察划分结果可知,片段 3 的前半句"面对这些全部来自基层官兵 ……按时限完成. "仍然属于政策的介绍,因此该部分应该归属于第一个分割单元.我们将在片段内部划分中解决这个问题.

以每个片段内的复句标点作为片段内的疑似分割点,则片段 2、3 内各有 1 个疑似分割点.按文中第 4 章介绍的片段内部划分方法分别计算以这两个疑似分割点重新划分后得到的上下两个片段的相似度,相似度值分别为 0.097 和 0.037.同时由表 2 可知,片段 2、3 的相似度值为 0.079.比较这三个相似度,0.037 为最小值,则以相似度 0.037 对应的疑似分割点作为新的段落分割点.经上述划分后可以得到如下的分割结果:

表 3　片段分割结果

分割单元 1(段落 1)	
片段 1	驻沪海军某部党委机关倾心 ……基层提出的建议项项有答复,件件有着落.
片段 2	新年伊始,记者在 ……责令机关各部抓紧制定切实可行的对策措施,及时答复,切实进行整改,按时限完成.
分割单元 2(段落 2)	
片段 3	(东海)海域常年潮湿多雨,舰艇上没有专门的晒衣区域 ……委婉地向机关提出能否建几个晒衣场.
片段 4	该部闻讯后,不仅相继建设 ……解决官兵生活上的不便.
片段 5	此问题一经提出,该部后勤 ……官兵们纷纷拍手叫好.
分割单元 3(段落 3)	
片段 6	(刘贵朝)、(崔晓龙)1.遵守 ……转载或引用.

由表 3 可见:通过对片段进行二次划分后能够寻找到位于片段内部的分割点,使得段落划分的结果更加合理.

6　结论

文本分割是自然语言处理的重要组成部分,可以应用于许多领域中.但是传统的分割方法大多以待分割文档中的所有词作为分割计算的依据,从而引入了分割噪声.针对上述问题,本文提出一种基于主题分析的文本分割算法,算法首先通过构建全文词汇链对文档主题进行分析,然后依此方法去除与文档描述的主题无关的词语对分割结果的干扰.算法还通过计算片段对词汇链的相交模式来计算片段间的相似度,并依此构造片段对子主题的覆盖图,以结合相邻及相间片段间的信息寻找片段划分的最优值,提高了划分的准确度.针对某些分割点可能落在片段内部的情况,算法对片段进行二次划分从而获得了更加合理的分割结果.实验表明,该算法能够使不同的分割单元描述不同的子主题信息,分割结果较好.

参考文献:

[1] Kauchak D, Chen F. Feature-based segmentation of narrative documents[A]. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics [C]. USA: ACL Press, 2005. 32 - 39.

[2] Hearst M A. TextTiling: segmenting text into multi-paragraph subtopic passages[J]. Computational Linguistics, 1997, 23(1): 33 - 64.

[3] Stokes N, Carthy J, Smeatona F. SeLeCT: a lexical cohesion based news story segmentation system[J]. Journal of AI Communications, 2004, 17(1): 3 - 12.

[4] Chen Qingcai, Wang Xiaolong, Liu Bingquan. Subtopic segmentation of Chinese document: an adapted dotplot approach [A]. Proceedings of 2002 International Conference on Machine Learning and Cybernetics[C]. Beijing: IEEE Press, 2002. 1571 - 1576.

[5] 石晶,戴国忠. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展, 2007, 44(2): 242 - 248.
Shi Jing, Dai Guozhong. Text segmentation based on PLSA mode[J]. Journal of Computer Research and Development, 2007, 44(2): 242 - 248. (in Chinese)

[6] Morris J, Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of the text[J]. Computational Linguistics, 1991, 17(1): 21 - 48.

[7] Kok Wee Gan, Ping Wai Wong. Annotating information structures in Chinese texts using HowNet[A]. Proceedings of the Second Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics[C]. HK: ACL Press, 2000. 85 - 92.

[8] Chan S W. Extraction of salient textual patterns: synergy between lexical cohesion and contextual coherence[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2004, 34(2): 205 - 218.

[9] Gonenc E, Ilyas C. Using lexical chains for keyword extraction [J]. Information Processing and Management, 2007, 43(6): 1705 - 1714.

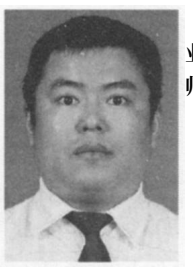
[10] 朱靖波,叶娜,罗海涛. 基于多元判别分析的文本分割模型[J]. 软件学报, 2007, 18(3): 85 - 94.
Zhu Jingbo, Ye Na, Luo Haitao. Text segmentation model based on multiple discriminant analysis[J]. Journal of Software, 2007, 18(3): 85 - 94. (in Chinese)

[11] 佘二永,王润生. 基于线性融合模型的多传感器图像融合[J]. 电子学报,2005,33(6):1008 - 1010.
She Eryong, Wang Runsheng. Multisensor image fusion based on linear fusion model[J]. Acta Electronica Sinica, 2005, 33(6):1008 - 1010. (in Chinese)
[12] Beeferman D, Berger A, Lafferty J. Text segmentation using exponential models[A]. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing[C]. USA: ACL Press, 1997. 35 - 46.

作者简介:

刘 铭 男,1981 年生于黑龙江. 哈尔滨工业大学计算机科学与技术学院博士研究生. 研究方向为聚类分析、文本挖掘、自然语言处理.
E-mail: mliu@insun.hit.edu.cn

王晓龙 男,1955 年生于黑龙江. 哈尔滨工业大学计算机科学与技术学院教授,博士生导师. 研究方向为信息检索、文本挖掘.



刘远超 男,1971 年生于黑龙江. 哈尔滨工业大学计算机科学与技术学院副教授,硕士生导师. 研究方向为自然语言处理、人工智能.