

基于归一化割的主题划分算法研究

王洋洋¹, 刘柏嵩^{2*}, 刘 薇¹

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 宁波大学 网络中心, 浙江 宁波 315211)

摘要: 为了保证抽取信息的全面性, 主题划分成了不可或缺的工作。借助同义词词林, 从词语的语义角度计算文本中各个段落间的相似度, 建立段落文本关系图。基于文本关系图对归一化割分割准则中权值矩阵的构建做出调整, 使之更能体现出段落间的相似程度, 并使用该准则对文本进行主题划分。结果表明, 该方法无论是对连续段落还是跨段落表达同一主题的主题划分均较为有效。

关键词: 同义词词林; 文本关系图; 段落相似度; 主题划分; 归一化割

中图分类号: TP391.1

文献标志码: A

文章编号: 1001-5132 (2013) 04-0040-05

随着信息技术的快速发展, 文本信息也随之快速增加, 人们迫切要求借助计算机迅速从中找到自己感兴趣的信息, 并全面地掌握获取信息的核心思想。如何通过计算机将文本中的主要思想抽取出来, 并全面、准确地展示给读者, 显得非常有意义。

很多文本反映的不是单纯一个主题, 而是多主题; 在一篇文章中, 作者常常从多角度论述思想, 若是采用基于句子重要度抽取信息的传统方法, 则容易造成抽取的信息集中在主要主题上, 忽略次要主题的信息, 这样导致抽取信息完整性较差^[1]; 同时也会造成抽取信息的冗余度较大。主题划分的预处理工作, 可以确保抽取信息能够全面、准确地反映文章的主题, 这样更容易获取新知识, 也为知识的链接^[2]和发现工作做准备; 对于读者也能够更便捷地找到感兴趣的内容。

对于自动文摘而言, 通常一个主题由文本中若干段相邻自然段组成^[3], 针对文本中相邻段落表达同一主题的情况进行主题划分工作, 可以提升自动文摘工作抽取信息全面性及准确性的质量。然而在一些文章中, 作者为了凸显某些主题的重要性, 常常在文章中多次强调这些主题, 即跨段落表达相同主题, 若是采用基于连续段落相似度主

题划分算法, 抽取出的主题同样避免不了重复、冗余的问题。此时, 自动文摘则依赖于对跨段落表达同一主题的情况进行主题划分工作。另外, 跨段落主题划分工作在新闻主题追踪领域也显得非常有意义。以新闻微博为例, 与一个主题相关的新闻大都孤立地分布在不同时间段, 跨段落主题划分工作有助于主题追踪工作。Hearst 等^[4]主题划分算法采用基于词频统计计算特征间的相似度, 文献[1]和文献[3]是将特征表示为向量空间模型计算特征间相似度, 这些方法忽略了词语间相关性。另外, 这些方法主题由连续的特征集组成, 对非连续特征集表达同一主题的情况效果不佳。

笔者首先针对向量空间模型忽略词语间相关性的缺陷, 从词语的语义角度出发基于同义词词林计算词语相似度, 进而得出句子、段落间的相似度, 然后建立段落文本关系图, 这样的描述更能体现段落间的相似程度。在保证对连续段落进行主题划分质量的同时, 针对文献[1]及文献[3]没有考虑到对文本中跨段落表达同一主题进行主题划分的情况, 引入归一化割分割准则。归一化割分割准则是一个全局优化的算法准则, 其权值矩阵构建的质量影响着分割的结果。在图像分割领域中使用归一化割分割准则时, 权值矩阵的值反映了像

收稿日期: 2013-06-09.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家社科基金 (08CTQ014); 浙江省教育厅科研项目 (Y200908634)。

第一作者: 王洋洋 (1988—), 男, 河南禹州人, 在读硕士研究生, 主要研究方向: 知识工程与数据挖掘。E-mail: wangyang@mail.nbu.edu.cn

*通信作者: 刘柏嵩 (1971—), 男, 安徽安庆人, 博士/研究员, 主要研究方向: 人工智能及计算机网络。E-mail: lbs@nbu.edu.cn

素点间的相似程度。而要将归一化割分割准则应用到主题划分领域中, 则需对权值矩阵的构建工作做出相应调整, 使其能够反映文本段落间的相似程度。笔者基于段落的文本关系图构建了权值矩阵, 然后采用归一化割分割准则进行主题划分, 将表达同一主题的段落归为一类, 完成对文本主题划分的工作。

1 基于归一化割的主题划分

归一化割分割准则中, 权值矩阵的构建是该分割准则的关键, 需要合理地度量出两个节点的相似程度。为了保证归一化割进行文本主题划分的效果, 笔者从语义的角度描述段落间的相似度, 以文本中各个段落为节点, 段落间相似度为权重构建权值矩阵。

从语义相似度的角度计算文本中各个段落间的相似度, 这样描述段落间的相似度更为准确。作者在阐述一个主题时, 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, 通常两个段落间的相似度可以通过衡量段落间包含词汇的重复程度来表示, 然而这种方法仅粗略地计算了段落相似度。

有些段落间包含词汇的重复程度不高, 但是它们包含词语间的词义相同或是词义有很强的相关性, 那么它们表达的内容很有可能属于同一主题。比如“计算机”、“微处理器”和“电脑”, “大豆”、“黄豆”和“毛豆”词义相同; “飞机”和“直升机”、“大炮”和“坦克”、“战斗机”和“武器”、“地铁”和“交通”其词义间有很强的相关度。若是纯粹地以段落中含有高频词的吻合程度衡量段落间的相似度, 显然不够准确。我们从语义的角度计算段落间相似度能较好地解决此问题。

1.1 文本关系图的建立

文本关系图是描述文本之间关系的一种形式化模型, 它是以文本中的某些特征项为顶点, 特征项之间的关系为边建立的一种关系图结构^[5]。以段落为基本单元、段落相似度为权重建立段落文本关系图, 如图1所示。

图1中顶点 p 和 q 分别表示文章中的段落, 连接顶点 p 和 q 的权值 C_{pq} 表示两段落间的相似度。段落间的相似度计算, 依赖于段落中句子相似度

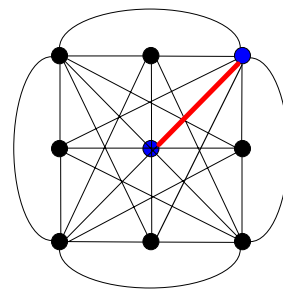


图1 文本关系图

的计算, 词语相似度又是句子相似度计算的基础, 它们的相似度算法描述如下。

1.1.1 词语相似度的计算

基于同义词词林的词语相似度, 取值范围在 $[0,1]$ 之间, 一个词语与其自身的语义相似度为1, 若是两个词语在任何上下文都不可替换, 其相似度为0。词语相似度和词语相关度是影响词语相似度的两个重要指标。词语相似度可以用词语间的语义距离来衡量, 两个词语的相似度与它们的词语距离成反比; 词语相关度一般用同一个语境中词语的共现程度衡量。词语相似性反映的是词语之间的聚合特征, 而词语相关性反映的是词语之间的组合特征^[6]。

一个中文词语往往有多种意思, 即有多个义项, 计算词语相似度时要考虑到所有的义项。我们采用文献[6]中的词语相似度计算算法, 计算词语相似度时把两个词语的所有义项分别两两计算, 然后取最大值作为两个词语的相似度值。

1.1.2 句子相似度的计算

两个句子在语义上的匹配程度表示句子相似度, 相似度取值为 $[0,1]$ 之间的实数, 定义为:

$$\text{sim}(s_1, s_2) = \sum_{i=1}^4 \alpha_i \text{sim}W_i(s_1, s_2),$$

其中: $\text{sim}W_i(s_1, s_2)$, $i=1,2,3,4$ 分别表示两个句子中名词、动词、代词和形容词的相似度。参数 α_i 是各部分相似度的权系数, 其满足 $\alpha_1+\alpha_2+\alpha_3+\alpha_4=1$, 这里参照文献[7], 取 $\alpha_1=0.5$, $\alpha_2=0.2$, $\alpha_3=0.17$, $\alpha_4=0.13$ 。

1.1.3 段落相似度的计算

假定文本 P 由 n 个自然段 P_i 构成, 即 $P=\{P_1, P_2, \dots, P_n\}$, 段落 P_i 与 P_j 之间的相似度定义为:

$$\text{Similarity}(P_i, P_j) = (\sum_{s=1}^m a_s / m + \sum_{t=1}^n b_t / n) / 2,$$

其中:

$$a_s = \max \{ \text{sim}(P_{is}, P_{j1}), \\ \text{sim}(P_{is}, P_{j2}), \dots, \text{sim}(P_{is}, P_{jn}) \}, \\ b_t = \max \{ \text{sim}(P_{i1}, P_{jt}), \\ \text{sim}(P_{i2}, P_{jt}), \dots, \text{sim}(P_{im}, P_{jt}) \},$$

$\text{Similarity}(P_i, P_j) \in [0, 1]$, 其表示该相似距离越大两者描述的主题内容越相似, 相似距离越小两者描述的主题内容越不相似。

1.2 归一化割准则

1.2.1 最小化割

一幅加权无向图 $G=(V, E)$ 可以通过删去某些边, 将其分为两个非连接性点集 A, B , 使得: $A \cup B = V, A \cap B = \emptyset$. 原来连接两部分而现在被删去的所有边的权的总和, 在图论中称之为割。

$$\text{cut}(A, B) = \sum_{i \in A, j \in B} w(i, j),$$

其中: $w(i, j)$ 是连接顶点 i 和顶点 j 边的权值, 表示两点之间的相似程度。这样只要使 $\text{cut}(A, B)$ 最小, 就可得到图 $G=(V, E)$ 的最优分割。然而最小化割准则常把一些孤立点分割为一类, 这样最小化割分割准则容易受到噪声的影响^[8]。

1.2.2 归一化割

为了克服这一缺点, Shi 等^[9]提出了归一化割准则, 定义为:

$$N\text{cut}(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, V)},$$

表示类间的相似度, 值越小分割效果越好。其中:

$$\text{assoc}(A, V) = \sum_{u \in A, v \in V} w(u, v),$$

表示 A 部分内的结点与整幅图结点 V 的权重之和。同时, 定义:

$$N\text{assoc}(A, B) = \frac{\text{assco}(A, A)}{\text{assoc}(A, V)} + \frac{\text{assoc}(A, B)}{\text{assoc}(B, V)},$$

表示类内的相似度, 其值越大分割效果越好。可以推出:

$$N\text{cut}(A, B) = 2 - N\text{assoc}(A, B),$$

这样可以得出: 归一化割准则在保证类间相似度最小的同时, 也保证了类内相似度最大^[10]。

1.2.3 权值矩阵的构建

权值矩阵构建的质量影响着分割的效果, 其权重要能够反映段落间的相似程度。我们以段落为节点, 段落相似度为权重构建权值矩阵。

1.2.4 归一化割算法步骤

最小化归一化割是个 NP-hard 问题, 其随着 $G=(V, E)$ 节点的增加, 计算将会变得复杂。在实际应用中我们采用近似的方法, 求解归一化割算法步骤为: (1) 基于前面的工作, 以段落为节点, 段落相似度为权值的文本关系图计算权值矩阵 W 和 D 。 W 是 $N \times N$ 的对称矩阵, 其元素值 $W(i, j) = w_{ij}$, 即段落间的相似度; D 为 $N \times N$ 对角矩阵, 对角线上的元素为 d_i , $d_i = \sum_j w(i, j)$ 是节点 i 与所有节点的关系之和。(2) 求出特征系统 $(D - W)y = \lambda Dy$ 的特征值和与其相对应的特征向量。(3) 使用上一步骤

表 1 权值矩阵

段号	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁	P ₁₂
P ₁	1.0	0.32	0.42	0.32	0.34	0.58	0.37	0.30	0.43	0.27	0.29	0.40
P ₂	0.32	1.0	0.31	0.32	0.58	0.32	0.31	0.27	0.39	0.23	0.40	0.31
P ₃	0.42	0.31	1.0	0.38	0.34	0.44	0.31	0.34	0.36	0.36	0.34	0.33
P ₄	0.32	0.32	0.38	1.0	0.35	0.31	0.23	0.34	0.38	0.30	0.41	0.25
P ₅	0.34	0.58	0.34	0.35	1.0	0.37	0.38	0.28	0.56	0.25	0.49	0.31
P ₆	0.58	0.32	0.44	0.31	0.37	1.0	0.36	0.29	0.39	0.25	0.26	0.29
P ₇	0.37	0.31	0.31	0.23	0.38	0.36	1.0	0.19	0.33	0.21	0.24	0.46
P ₈	0.30	0.27	0.34	0.34	0.28	0.29	0.19	1.0	0.31	0.40	0.26	0.24
P ₉	0.43	0.39	0.36	0.38	0.56	0.39	0.33	0.31	1.0	0.25	0.29	0.28
P ₁₀	0.27	0.23	0.36	0.30	0.25	0.25	0.21	0.40	0.25	1.0	0.24	0.24
P ₁₁	0.29	0.40	0.34	0.41	0.49	0.26	0.24	0.26	0.29	0.24	1.0	0.34
P ₁₂	0.40	0.31	0.33	0.25	0.31	0.29	0.46	0.24	0.28	0.24	0.34	1.0

注: 表中统计的是文本中各个段落间的相似度, 文本是从新闻热点微博中按照时间顺序随机抽取组成, 其中, 段落 P₁、P₆、P₇、P₁₂ 描述第 1 个主题, P₂、P₄、P₅、P₉、P₁₁ 描述第 2 个主题, 剩下的段落 P₃、P₈、P₁₀ 描述第 3 个主题。

的第二最小特征向量分割段落为两部分。(4)检查 $Ncut$ 值是否稳定或主题包含的最小段落数决定是否继续分割。(5)若需要继续分割,递归此算法分割上一次分割好的部分。

1.3 整体算法步骤

基于前面工作,设计出的基于归一化割主题划分的算法为:(1)输入文档进行预处理,设待处理文档的段落为 m 段。(2)基于同义词词林计算词语间的相似度,并在此基础上计算段落相似度。(3)以 m 段段落为节点, m 段落间的相似度为权值建立文本关系图,并构建表示段落间相似程度的权值矩阵。(4)在构建好的权值矩阵基础上,进行归一化割算法处理。(5)最终完成对文档的主题划分。

2 实验与结果分析

2.1 实验资源

选用大量由专家进行了主题划分的语料,测试本文提出的主题划分算法的准确性。但是,目前这方面的语料较少,另外主题的划分标准也较难制定,是一个较为主观的概念,对于同一篇文章进行主题划分,各个人可能会有不同主题划分的方法。鉴于此,我们随机从语料库中抽取一些篇幅较短的文章组成一篇长的文章,由于篇幅较短,可以假定每一篇文章都属于同一个主题,而不同的文章则分属不同的主题。这样对一篇由一些描述同一主题的短文章组成的长文章,其主题划分就有了可操作的标准。

为了测试该算法的准确性,分别从搜狗语料库及当前各大互联网中随机抽取 1000 篇含有科技、军事、财经、娱乐、汽车、体育、时尚等不同类型的文档,组成 300 篇长文章,每篇文章含有 3~5 篇不同类型的短文。基于这些实验资源,对本文算法进行了有效性测试。

为了测试本算法对跨段落表达同一主题的有效性,我们将之前建立的资源库中各个长文章内部的段落随机分布,然后用算法再对此实验资源进行处理,分析本算法在跨段落主题划分中的有效性。

2.2 实验评价方法

假如划分为同一主题下的段落恰好是各短文章包含的段落,那么认为这是一个正确的划分,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:

$$p = n / N \times 100\%,$$

其中: n 为划分结果正确的文章数, N 为进行测试的文章总数。

2.3 实验结果及分析

由表 1 可以看出, P_3 段与 P_1 、 P_6 段的相似度明显大于 P_3 与 P_8 、 P_{10} 段的相似度,但最终结果显示依旧将 P_3 同 P_8 、 P_{10} 段划分到了一个主题下,这体现了本算法的全局优化特性及对连续段落和跨段落主题划分的有效性。

表 2 连续段落表达同一主题

结果	文章数	所占比例/%
主题划分正确	268	89.3
相差一个段落	22	7.3
其他	10	3.3
总和	300	100.0

表 3 跨段落表达同一主题

结果	文章数	所占比例/%
主题划分正确	260	86.7
相差一个段落	22	7.3
其他	18	6
总和	300	100

由表 2 可知:对 300 篇连续段落表达同一主题的多主题文章的主题划分中有 268 篇文章划分正确,有 22 篇划分为同一主题中的段落与其对应的短文章中包含段落相差一段。

由表 3 可知,对 300 篇跨段落表达同一主题的多主题文章的主题划分中有 260 篇文章划分正确,有 22 篇划分为同一主题中的段落与其对应的短文章中包含段落相差一段。虽然这种结果还不够精确,但是由于主题划分用在自动文摘长度的具体分配上,所以对它们的后续工作影响不大。

从以上结果可以看出,本文的主题划分算法对于连续段落表达同一主题的情况进行主题划分的准确率为 89.3%,对于跨段落表达同一主题的情况进行主题划分的准确率为 86.7%。表明本文方法对跨段落进行主题划分的准确率略低于其对连续段落进行主题划分的准确率。对跨段落进行主题划分的准确率虽还不是非常理想,但已克服了文

献[1]和文献[3]中算法不能对跨段落进行主题划分的不足。另外其效果也能满足对跨段落情况进行主题划分的要求。所以本文算法对文本中表达同一主题的段落是否连续的情况都是可行的,而且其准确率也令人满意。因此,对于自动文摘的主题划分或是从网络新闻中找出谈论主题相似信息的工作都是较为有效的。

3 结语

借助同义词词林计算词语间相似度,从语义相似度的角度计算了文章中各个段落间的相似度,建立段落间的文本关系图,并转换为归一化割中的权值矩阵,然后使用归一化割分割准则对表达同一主题的段落进行划分。结果表明,文中基于归一化割进行主题划分的算法思想,无论是对表达同一主题的连续段落,还是对表达同一主题而分散于文章中不同区域段落都比较有效,这为自动文摘、新闻主题追踪等工作奠定了基础,使其抽取的内容更具全面性和准确性。

参考文献:

[1] 童毅见,唐慧丰.面向自动文摘的主题划分方法[J].

北京大学学报:自然科学版,2013,49(1):39-44.

[2] 刘薇,刘柏嵩,王洋洋.基于知识元的数字图书馆学科标引研究[J].宁波大学学报:理工版,2013,26(1):100-103.

[3] 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.

[4] Hearst M A, Plaunt C. Subtopic structing for full-length document access[C]//Proceedings of the 16th Annual International ACM/SIGIR. Pittsburgh, 1993:56-68.

[5] 段晓丽,王宇.基于主题分割与 PageRank 算法的文本主题抽取[J].现代图书情报技术,2010,12:34-39.

[6] 田久乐,赵蔚.基于同义词词林的词语相似度计算方法[J].吉林大学学报:信息科学版,2010,28(6):602-608.

[7] 金博,史彦军,滕弘飞.基于语义理解的文本相似度算法[J].大连理工大学学报,2005,45(2):291-297.

[8] Shi Jianbo, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[9] Shi Jianbo, Malik J. Motion segmentation and tracking using normalized cuts[C]//ICCV Proceedings of the Sixth International Conference on Computer Vision. Bombay, India: 1998:1154-1160.

[10] 席秋波.基于 Ncut 的图像分割算法研究[D].成都:电子科技大学,2010.

Probe: Normalized Cuts Based Topic Partition

WANG Yang-yang¹, LIU Bai-song^{2*}, LIU Wei¹

(1.Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China;

2.Network Center, Ningbo University, Ningbo 315211, China)

Abstract: To ensure the completeness of information extraction, the topic partition is one of the indispensable tasks. With the aid of Tongyici Cilin, we first seek the similarity between paragraphs from the point of semantic computing, based on which we then establish text relation map. Using and accordingly adjusting the weight matrix, the degree of similarity between paragraphs can be more accurately obtained, in which the Normalized Cut approach is adopted to complete the topic partition of text. The experimental results show that the method is effective either for consecutive paragraphs or for cross-paragraphs expressing a similar topic.

Key words: Tongyici Cilin; text relation map; paragraph similarity; topic partition; Normalized Cut

(责任编辑 史小丽)