动态流形方法在多文档文摘模型上的应用

刘美玲[1],郑德权[2],王慧强[3],于 洋[1]
(1. 东北林业大学 信息与计算机工程学院,黑龙江 哈尔滨 150040;
2. 哈尔滨工业大学 教育部-微软语言语音重点实验室,黑龙江 哈尔滨 150001;
3. 哈尔滨工程大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

摘 要:网络动态演化内容的识别和分析是人们快速获取有效信息的主要手段之一,已经成为人们迫切需要解决的关键问题。动态多文档文摘建立在时间信息基础上,从网络动态演化性出发,对同一话题不同时段的文档集合进行分析,在识别信息内容差异性的基础上,对信息的动态演化性进行建模。文中在经典流行排序思想的基础上,进一步提出了动态流行排序模型。该模型中不仅融入了信息的重要性特征,而且融入了信息与历史信息的关联特征以及信息的时间特征,使文摘信息动了起来,即文摘系统具有了动态性。该模型在国际标准评测 TAXT ANYNASIS CONFERENCE 2008 的 Update task 任务语料上进行了测试,获得了较好的实验结果。

关键词:动态多文档文摘;动态演化性;差异性分析;相似度;质心整体选优
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2018)03-0026-06
doi:10.3969/j.issn.1673-629X.2018.03.006

Application of Dynamic Manifold Method in Multi Document Summarization Model

LIU Mei-ling[1], ZHENG De-quan[2], WANG Hui-qiang[3], YU Yang[1]
(1. School of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China;
2. Ministry of Education-Microsoft Key Laboratory of Speech Language,
Harbin Institute of Technology, Harbin 150001, China;
3. School of Computer Science and Technology, Harbin University of Engineering, Harbin 150001, China)

Abstract: The identification and analysis of evolutionary information on the internet is an efficient means to get useful information, which has become a critic issue urgent to work out. Based on time information, starting from network dynamic evolution, the dynamic multi_document summarization analyzes the document sets of different period about a same topic. On the basis of identifying the difference of information content, a summarization model can be built. Based on the classic manifold ranking model, we propose a dynamic manifold ranking model which not only adds some significant features, but also introduces some historical redundancy features and some time information feature, which make the information contained by abstract dynamic. An evaluation based on this model is conducted on the update task corpus of TAXT ANYNASIS CONFERENCE 2008 and a good testing result is obtained.

Key words: dynamic multi_document summarization; dynamic evolution; difference analysis; similarity; overall centroid optimized

0 引 言

在 Web2.0 时代,网络上的各种新闻、论坛、博客、在线聊天等信息跟静态网页信息相比体现出非常明显的动态演化性。网络信息随着时间的变化而出现、发展直至消亡,一个话题在不同的时刻具有不同的侧重点,而不同时刻的话题内容之间具有关联性,如何针对这类持续发展变化的话题或者事件提供动态摘要已经成为一个新的研究方向。

传统的多文档文摘[1]技术是一种静态文摘,即针对某个封闭的静态文档集生成摘要,不考虑文档集的对外联系。动态文摘是传统静态文摘的延伸和扩展,除了需要保证文摘信息的主题相关性和内容的低冗余性之外,还需要针对内容的动态演化性分析已出现信息和新出现信息的关系,使文摘随话题的演化而动态

收稿日期:2017-01-26 **修回日期**:2017-05-31 **网络出版时间**:2017-12-04
基金项目:中央高校基本科研业务费专项资金(2572014CB26);黑龙江省自然科学基金(F2015037)
作者简介:刘美玲(1981-),女,博士,讲师,CCF 会员(20378M),研究方向为信息检索、数据挖掘、智能交通。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20171204.1647.006.html
万方数据

更新。动态文摘与静态文摘的最大区别在于分析已出现信息和新出现信息的关系,从而对内容的动态演化性进行建模。

TAC2008 的评测任务中 Update Summarization 作为文摘研究的标准备受关注,文中对动态多文档文摘动态演化的内容选择问题展开进一步的研究。流形排序(manifold-ranking)是经典的排序方法,之前在话题相关文摘中的应用效果不错,但该方法并不能捕捉时间片进化的信息。文中以动态信息的模拟演化为目标,通过建立动态流形排序模型来为动态多文档文摘话题相关的内容选择提供重要性排序。

提出了一种动态流形排序模型(dynamic manifold ranking model,DMRM),将其用于动态多文档文摘的研究中,使文摘同时融入了文档的流形结构和动态演化性。在动态多文档文摘领域,对相关文档集进行特征抽取是文摘技术的核心。主流思想是以信息显著性和信息新颖性为主要特征,根据句子信息显著度和信息新颖度对句子加权排序,抽取排序值最高的句子作为文摘句;对已经提取的文档集特征,根据信息显著度对句子加权排序,进而根据信息新颖度过滤句子,过滤掉信息新颖度低的句子,最后从剩余的句子集合中抽取排序值高的句子作为文摘句。在上述两种思想中,都把文档集中的句子看成是孤立的,认为句子之间没有关联,这是一种错误的假设。文档集中的句子,有相当一部分相互之间具有关联性。

文中提出的动态流形排序思想弥补了上述两种模型的不足,基于动态分析,考虑了句子之间的相关性。动态流行排序是一种迭代算法,考虑了句子集合中数据点的流行结构,经迭代后,相似的句子趋向于具有相同的排序值,同类的句子趋向于具有相同的排序值,克服了常规文摘方法的缺点。

1　相关工作

1.1　动态多文档文摘和流行排序的相关研究

美国 NIST[2] 承办的 Document Understanding Conference (DUC) 2007[3] 首次提出了动态文摘任务,在 IARPA[4] 的支持下于 2007 年举行了第一届评测会议,并且在 Text Analysis Conference (TAC)2008[5] 中仍然被作为重要的评测任务之一。在时序信息高速演化的背景下,快速的动态信息获取技术成为数据挖掘和自然语言处理的研究重点。

国内很多学者在文摘方面的研究效果显著,例如,静态文摘和动态文摘相结合就是一种基于改进文摘模型的动态文摘解决方法。张瑾等[6]提出了一种基于模糊隶属度的文档过滤模型。该方法从对动态内容的建模入手,通过模式识别和传统文摘生成方法,对动态内容进行提取和分析。在动态网络演化信息中,句子选择和排序也需要动态变化,因此需要解决如何在排列策略中体现动态内容的演化性问题。文中主要对信息显著度(information significance,IS)[7] 和信息新颖度(information novelty,IN)两种指标进行评估和分析,在此基础上改进设计一种基于动态时序内容的句子排列流形策略。

流形这个概念最早产生于人类对感知的研究[8],最初阶段关系到与物理世界(地球的表面)和几何公理研究有关的多维参数思想的分析[9]。从拓扑学角度出发,流形表示一个局部为欧几里德的拓扑空间。局部欧几里德特性意味着对于空间上任一点都有一个邻域,在这个邻域中的拓扑与 R_m 空间中的开放单位圆相同,R_m 表示 m 维欧氏空间,从拓扑空间的一个开集(邻域)到欧氏空间的开子集的同胚映射,使得每个局部可坐标化。它的本质是分段线性处理[10]。流形学习的主要目标是从非线性高维数据中发现嵌入其中的低维光滑流形,以进行维数约简和数据分析。

流形排序[11-12] 在话题相关的静态多文档文摘中得到了很好的应用,在传统文摘技术中应用流形排序学习算法中得到了启发。文中面向动态多文档文摘领域,提出了一种面向查询的动态流形排序模型,该模型更好地体现了文档的流行结构和动态演化性。

1.2　主流的评测方法

目前在时序多文档文摘的评价方面完全沿用传统静态多文档文摘的评价方法,包括自动评价 ROUGE[10]、BE[13] 方法和人工评价金字塔(PYRAMID)[14] 方法。文摘评价主要面向文摘的内容选择和语言质量。自动评价针对文摘的内容选择进行评测,而人工评价则针对文摘的内容选择、语言质量和整体的反映度(综合考虑面向话题的覆盖度和流利度)进行评测。

TAC 是多文档文摘领域最有影响的国际评测会议,由美国国家技术标准局(national institute of standards and technology,NIST)主办的 DUC 和 TREC 中的问答评测演化而来。TAC 评测由美国 IARPA(intelligence advanced research projects activity)资助,每年由 NIST 的信息技术研究室中的信息检索组主办,由来自政府、企业和学术界的顾问委员会监督。Update summarization 评测面向英语,测试语料主要来自 TREC 中 QA 评测的 AQUAINT-2 数据集。

2　DMRM 多文档文摘模型

2.1　动态流形排序思想

经典流形排序主要用于数据点查询问题中,即数据挖掘领域。其主要排序特征是查询数据点,查询数

万方数据

据点一般来说是静态的,这是经典流形排序为静态模型的原因。在动态多文档文摘领域,其主要的排序特征是信息显著性和信息新颖性。具体而言,信息显著性包括的特征有:句子与所有其他句子相似度累加值特征;句子在文档中的位置特征;句子的长度特征。信息新颖性包括的特征有:与历史文摘的相似度值,相似度愈小,新颖性愈强;句子的时间特征。文中提出的动态流行排序模型主要使用这五个特征对句子加权,进行文摘内容的选择和排序。

2.2 DMRM 的算法流程

DMRM 的算法流程如图 1 所示。

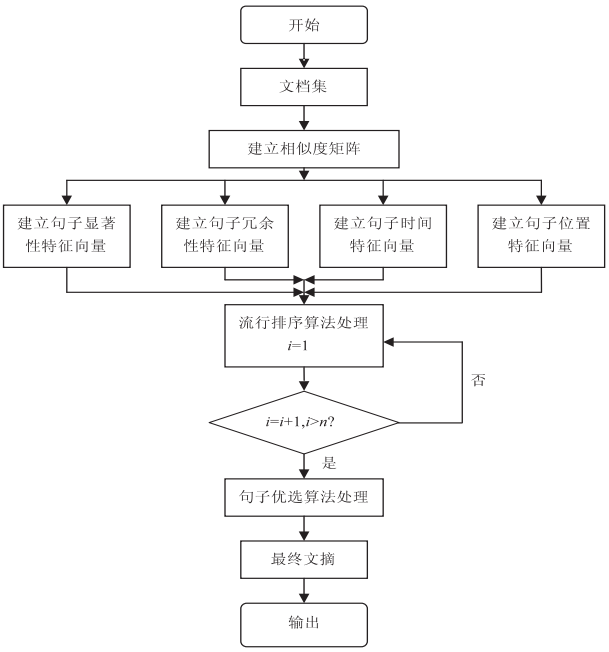

图 1 DMRM 的算法流程

2.3 DMRM 的建立

2.3.1 句子相似度矩阵 $\boldsymbol{W}$

该模型的第一步为相似度矩阵的建立,用来度量句子集合中句子之间的相关性,是动态流行排序思想的基础。相似度矩阵的建立过程也是为文档集中的句子集建立带权无向图的过程。该矩阵的建立要依赖于两句子之间的相似度算法,所以相似度算法的选择至关重要。虽然该领域中已存在不少相似度算法,但是其在该模型中的应用效果均不佳。基于此,文中提出了基于 TII 的句子相似度计算算法,其算法公式如下:

$$\mathrm{Sim}(s_i,s_j)=\frac{\sum_{w\in s_i,s_j}\mathrm{Weight}(w)}{\mathrm{length}(s_i)+\mathrm{length}(s_j)} \tag{1}$$

其中, $\boldsymbol{W}$ 为句子 s_i 和 s_j 中的同现词; $\mathrm{Weight}(w)=\mathrm{TF}(w)*\mathrm{IDF}(w)*\mathrm{ISF}(w)$ 为词语 $\boldsymbol{W}$ 的权重,其中 $\mathrm{TF}(w)$ 表示词语 $\boldsymbol{W}$ 的频率, $\mathrm{IDF}(w)$ 表示词语 $\boldsymbol{W}$ 的反文档频率, $\mathrm{ISF}(w)$ 表示词语 $\boldsymbol{W}$ 的反句子频率。此三值的统计范围为当前文档集句子集合,其中 $\mathrm{IDF}(w)=1/\mathrm{DF}(w)$, $\mathrm{DF}(w)$ 为整个文档集合中包含词 $\boldsymbol{W}$ 的文档数, $\mathrm{ISF}(w)=1/\mathrm{SF}(w)$, $\mathrm{SF}(w)$ 为整个文档集中包含词 $\boldsymbol{W}$ 的句子数; $\mathrm{length}(s_i)$ 和 $\mathrm{length}(s_j)$ 分别表示 s_i 和 s_j 的长度。

运用该相似度算法对文档集句子集合中所有句子其相互之间的相似度值进行计算,即可建立相似度矩阵 $\boldsymbol{W}$。

2.3.2 句子显著度向量 $\boldsymbol{A}$

动态流行排序模型的第二步为句子特征值的提取。定义向量 $\boldsymbol{A}$,其元素表示当前文档集句子集合中相应句子与所有其他句子的相似度累加值,这个值是衡量句子重要性的一个特征。这种思想基于投票原理:句子集合中的句子之间具有关联性,这种关联性的强弱可通过其与其他句子间的相似度大小来体现,同时与其具有关联性的句子数量同样能体现出这种关联性强弱。综合考虑以上两项因素,文中提出用句子间的相似度累加值作为衡量句子关联性的参数,若某句子拥有相当大的关联性度量值,即表示该句子所含信息的显著度值很大,那么该句子将成为一重要的候选文摘句,因此该特征将成为候选文摘句选择的一重要指标。计算某句子 sent 相似度累加值的公式如下:

$$\mathrm{Current\_Weight}(\mathrm{sent})=\sum_{i=1}^{n}\mathrm{Sim}(\mathrm{sent},s_i) \tag{2}$$

其中, n 表示当前文档集中句子的总数; $\mathrm{Sim}(\mathrm{sent},s_i)$ 可由式(1)的计算方法得到,表示句子 sent 和句子 s_i 之间的相似度。

运用该算法计算句子集合中所有句子的相应值,即可建立向量 $\boldsymbol{A}$。

2.3.3 句子冗余度向量 $\boldsymbol{B}$

向量 $\boldsymbol{B}$ 中的元素表示句子与历史文摘中所有句子的相似度累加值,这个值是衡量句子信息新颖度的一个参数值。基于上述投票原理,句子与历史文摘句子集合的相似度累加值愈大,该句子与历史文摘中的句子具有的关联性愈大,表明该句子包含更多冗余信息。在动态流形排序模型中使用此特征可过滤掉信息冗余度高的句子,这是动态流形排序模型动态性的表现之一。文中提出的计算公式如下:

$$\mathrm{History\_Weight}(\mathrm{sent})=\sum_{i=1}^{n}\mathrm{Sim}(\mathrm{sent},s_i) \tag{3}$$

其中, n 表示历史文摘中的句子总数; $\mathrm{Sim}(\mathrm{sent},s_i)$ 同式(2)。

运用该公式计算当前文档句子集合中所有句子的相应值,即可得到向量 $\boldsymbol{B}$。

2.3.4 动态特征选择

(1)句子时间特征向量 $\boldsymbol{C}$ 的建立。

由于句子时间特征是文摘动态性的一个重要体

万方数据

现,因此系统融入了对其的考虑。直接考虑每个句子的时间特征涉及到时间短语的提取和归一化,这是时序多文档文摘的研究内容,考虑起来过于复杂,而且该系统的研究内容为动态多文档文摘,与时序多文档文摘有一定的区别,没有必要考虑所有的时间短语。所以该系统将避开直接考虑句子级的时间特征,而从文档集整体角度去考虑时间特征,这为问题的解决提供了方便。考虑到文档集中各个文档的出版时间有先有后,以及动态多文档文摘具有动态演化特性,所以出版时间靠前的文档具有小的新颖性,出版时间靠后的文摘具有大的新颖性。基于此原理,文中以文档在文档集合中出现的时间顺序来衡量该文档的新颖性,进而衡量该文档中句子的新颖性。句子信息新颖性度量值计算公式如下:

$$\text{Time\_Weight}(\text{sent}) = i \tag{4}$$

其中, Time_Weight(sent) 为句子 sent 的时间特征权值; i 为句子 sent 所属文档在文档集中根据时间排序的排序值。

运用该公式即可计算当前文档句子集合中所有句子的相应值,形成时间特征权重向量 $\boldsymbol{C}$ 。

(2)句子位置特征向量 $\boldsymbol{D}$ 的建立。

句子的位置特征对于多文档文摘系统是不可或缺的。句子在文档中的位置决定了其重要性,根据文章的规律,位置靠前和靠后的句子比在中间的句子具有更高的重要性,加入句子位置特征能使文摘系统具有更好的性能。所以文中在动态流形排序模型算法中加入句子的位置特征,其计算公式如下:

$$\text{Position\_Weight}(\text{sent}) = \frac{1}{\text{ps}} \tag{5}$$

其中,Position_Weight(sent)表示句子 sent 的位置特征值;ps 表示句子 sent 在所属文档中的位置值。

利用该公式即可计算当前文档中所有句子的相应值,从而建立句子位置特征向量 $\boldsymbol{D}$ 。

(3)句子长度特征。

无论对于静态多文档文摘系统,还是动态多文档文摘系统,句子长度特征都是必不可少的。若句子太短,则该句子不具有很高的重要性;若句子太长,即使重要,由于占用文摘的空间太大,也不利于文摘的效果的提高,因为在动态多文档文摘中,文摘是有字数限制的。例如,TAC 是国际上著名的文摘评测会议,其 update summary 任务是专门针对动态多文档文摘评测的,其要求文摘字数在一百字以内,因而对句子长度的考虑是必须的。文中按如下方法使用该特征:当句子长度大于 n_1 与小于 n_2 时,考虑该句子;否则舍去。该模型算法中,设置 n_1 为 10, n_2 为 25。在算法设计阶段没有用到句子长度特征,而在文摘句优选阶段考虑句子长度特征,有助于动态流形排序算法的实现。

万方数据

2.3.5　动态流形排序思想的核心

经典流形排序思想主要用于早期的数据点查询问题,描述如下:令 f 表示一个排序函数,其赋予每一个节点 x_i 一个排序值 f_i ,如此, f 可表示为一个向量 $\boldsymbol{f} = [f_1, f_2, \cdots, f_n]^{\mathrm{T}}$ 。同时,定义向量 $\boldsymbol{y} = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$,若 x_i 是一个查询,则令 $y_i = 1$;否则,令 $y_i = 0$。首先定义相邻矩阵 $\boldsymbol{W} = \{W_{ij} \mid i, j = 1, 2, \cdots, n\}$,其中 W_{ij} 表示从 x_i 到 x_j 的相似度。再定义另外一个矩阵 $\boldsymbol{S}$,其计算公式为 $\boldsymbol{S} = \boldsymbol{D}^{-1}\boldsymbol{W}$,其中 $\boldsymbol{D}$ 为对角阵,其第 (i,i) 个元素等于 $\boldsymbol{W}$ 第 i 行之和,其他值均为 0,矩阵 $\boldsymbol{S}$ 称拉普拉斯矩阵,其值 S_{ij} 即为从 x_i 到 x_j 的转移概率。在矩阵拉普拉斯矩阵 $\boldsymbol{S}$ 的基础上,句子 $x_1, x_2, \cdots, x_n$ 的重要性权重 f 可由与之相邻的其他句子推导出来。f 的计算公式可以递归地表示为:

$$f(t+1) = \alpha * S * f(t) + (1-\beta) * y \tag{6}$$

其中, α 和 $1-\alpha$ 分别表示相邻节点和初始的查询数据点的排序值对当前排序值的相对贡献。

分析经典流形排序模型算法可知,整个算法只使用了一个特征,即查询数据点。因为对数据查询问题就只依赖于这一个特征,所有元素的排序值都由此特征决定。动态多文档文摘的目的是抽取最重要的指定数量的句子作为文档集的文摘,其排序对象是当前文档集的句子集合。由前面的分析可知,句子的重要性程度依赖于五个特征:与当前文档集中句子集合的相似度累加值;与历史文摘中句子集合的相似度累加值;句子的位置特征;句子的时间特征;句子的长度特征。由于动态流形排序算法暂不考虑句子的长度特征,故还有四个需考虑的特征,根据这四个特征建立了四个向量。其中句子与当前文档集中句子集合相似度累加值向量 $\boldsymbol{A}$ 和句子位置特征向量 $\boldsymbol{D}$ 的加入意味着在句子排序值中加入了句子的信息显著度;句子与历史文摘中句子集合相似度累加值向量 $\boldsymbol{B}$ 和句子时间特征向量 C 的加入意味着在句子排序值中加入了信息新颖度,体现了该动态流形排序模型的动态性。

动态流形排序思想的核心-迭代计算句子的排序值向量 $f(t)$ 受经典流形排序思想的启发,文中提出的动态流行排序迭代算法公式为:

$$f(t+1) = \alpha * S * f(t) + \beta * A - \gamma * B + \eta * C + \lambda * D \tag{7}$$

其中, $f(t)$ 和 $f(t+1)$ 分别表示一次迭代前后的排序值; α 为相邻点对该句子排序值的贡献; β 为当前文档集中与之相似的句子集合对该句子排序值的贡献; γ 为历史文摘中与之相似的句子对该句子的惩罚; η 为该句子时间特征对之排序值的贡献; λ 为句子位置特征对该句子排序值的贡献。

该公式计算完成之后的 $f(t+1)$ 的最终值记为向量 $\boldsymbol{f}$,其第 i 个元素为句子 $sent_i$ 的权重,也就是 Weigth($sent_i$)。

由式(7)可知,该算法基于迭代算法,算法的迭代次数理所当然地会影响实验结果。迭代次数过多,句子集合中的所有句子排序值差异将非常小,那么对后面的其他算法,很小的参数波动都会使得实验结果有很大的差异性;评测语料的不同也会使实验结果产生很大的差异性,使算法的稳定性变差。迭代次数过少,句子之间的关联性所起的作用就相当小,达不到动态流形排序原本的目的。因此,迭代次数的确定也是算法的一个重要环节。

2.3.6 文摘句优选算法

动态流形排序的优点是考虑了句子之间的关联性,使重要的句子之间互相推荐,使得抽取的文摘句都具有很高的重要性;缺点恰巧也在此,因为该算法导致抽取的句子都是相互之间有很高相似度的句子,用此句子集合形成的文摘具有很高的冗余性,使得文摘的概括面非常窄,导致结果不理想。若想通过此模型得到好的文摘,必须解决文摘句的优选问题。传统的 MMR 文摘句优选算法,句子之间的相似度计算基于词频统计方法,由于该算法不能很好地计算句子之间的相似度,传统的 MMR 文摘句优选算法的效果不佳。基于此,文中提出了一种新的文摘句优选算法,其计算公式如下:

$$New\_Weight(sent) = \alpha * Old\_Weight(sent) - (1-\alpha) * \frac{\sum_{i=1}^{n} Sim(sent, sent_i)}{\sum_{i=1}^{n} Weigth(sent_i)} \tag{8}$$

其中,New_Weight(sent)表示选优之后候选句 sent 的权重;Old_Weight(sent)为候选文摘句 sent 优选之前的权重;n 表示已选入的文摘句数;$\sum_{i=1}^{n} Sim(sent, sent_i)$ 表示候选句 sent 与已选入的所有文摘句的累加,见式(1);$\sum_{i=1}^{n} Weigth(sent_i)$ 表示已选入的所有文摘句 $sent_i$ 的权重累加值;α 表示比例因子,当 $\alpha=0.2$ 时,性能达到最优。

3 实 验

3.1 实验语料及评测方法

在 TAC2008 中,Update Summarization 任务的测试语料由来自 AQUAINT-2 的 48 个话题组成,每个话题包含两个时间片,且均由 10 个文档组成。评价标准采用文摘评测领域著名的 ROUGE 工具,其中最主要的两个指标 ROUGE-2 和 ROUGE-SU4 * 是评价 Update 文摘的。

3.2 实验结果及分析

迭代算法中的所有参数都影响系统的性能。不同的参数设置,相应的实验结果差异性很大,因此针对文中提出的模型,测试了以下的参数设置。

做了三组实验,第一组比较信息新颖度和信息显著度对_DMRM 的贡献,第二组比较不同的迭代次数对_DMRM 的影响,第三组比较 DMRM 性能与 TAC2008 Update,分别如表 1 ~3 所示。

表 1 不同参数设置的比较

α	β	γ	η	λ	R-2	R-SU4 *
0.2	0.2	0.2	0.2	0.2	0.034 07	0.064 95
0.3	0.3	0.2	0.1	0.1	0.056 04	0.082 47
0.4	0.3	0.1	0.1	0.1	0.107 07	0.135 05
0.2	0.3	0.3	0.1	0.1	0.063 32	0.103 14
0.1	0.1	0.1	0.2	0.3	0.045 12	0.114 54

从表 1 可以看出,当 $\alpha=0.4$,$\beta=0.3$,$\gamma=0.1$,$\eta=0.1$,$\lambda=0.1$ 时,文摘性能最好。最好效果出现在 $\alpha=0.4$ 的参数设置上,表明了句子集合中句子之间的关联性对文摘句抽取影响很大,由得分可以看出动态流行排序在动态多文档文摘领域中具有一定适用性。

表 2 不同迭代次数的比较

iterative frequency	R-2	R-SU4 *
10	0.054 14	0.088 75
30	0.074 12	0.010 54
50	0.107 07	0.135 05
70	0.078 96	0.112 44
100	0.063 32	0.098 74

从表 2 可以看出,当迭代次数为 50 时,ROUGE-2(R-2)和 ROUGE-SU4(R-SU4)得分最高,即文摘性能最好,说明此模型的时间复杂度适中,系统运行起来速度较快。

表 3 与 TAC2008 发布结果的比较

SYSTEM	R-2	R-SU4 *
DMRM	0.107	0.135
Rank_1	0.101	0.137
Rank_2	0.097	0.134
Rank_3	0.092	0.132

TAC2008 总共发布了 73 个系统分数,表 3 中列出了前三名系统的 ROUGE-2 和 ROUGE-SU4 打分,与此分数比较,本模型在 TAC2008 的发布结果中排名第 1,说明动态流行排序模型很有潜力,是一种不错的动态多文档文摘建模方法。

万方数据

综上,在动态多文档文摘领域,动态流行排序思想值得研究,是一种有效的动态多文档文摘方法。

4　结束语

在认真研究国内外多文档文摘领域最新发展的基础上,创新性地对动态内容的演化关系进行了差异性分析。考虑到文摘句的信息新颖度和信息显著度对文摘的重要性,运用流行排序思想整合信息新颖度和信息显著度对句子集合中所有句子进行排序,根据排序值抽取句子形成文摘。同时融入对句子历史信息特征的惩罚和时间特征的奖励后,还能实现对文档集所含信息动态演化性的建模,使文摘具有动态性,对于推动动态多文档文摘领域的发展起到了一定的作用。下一步将是研究如何与其他模型更好地融合,使动态文摘具有更好的显著性和新颖性。

参考文献:

[1] NENKOVA A, MASKEY S, LIU Y. Automatic summarization [C]//Proceedings of the 49th annual meeting of the association for computational linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001.

[2] ALLAN J, JIN H, RAJMAN M, et al. Topic-based novelty detection[R]. Baltimore: Center for Language and Speech Processing, Johns Hopkins University, 1999.

[3] TRIPATHY A, AGRAWAL A, RATH S K. Classification of sentimental reviews using machine learning techniques[C]//Proceedings of 3rd international conference on recent trends in computing. [s. l.]: [s. n.], 2015: 821-829.

[4] ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking[C]//Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM, 1998: 37-45.

[5] GILLICK D, FAVRE B. A scalable global model for summarization[C]//Proceedings of the workshop on integer linear programming for natural language processing. [s. l.]: [s. n.], 2009: 10-18.

[6] 张　瑾,许洪波. 基于动态内容的文摘方法研究[C]//全国信息检索与内容安全学术会议. 出版地不详:出版者不详,2007.

[7] XIE X, LIU Y, LE W, et al. S-looper: automatic summarization for multipath string loops[C]//International symposium on software testing and analysis. New York, NY, USA: ACM, 2015: 188-198.

[8] SEUNG H, LEE D D. The manifold ways of perception[J]. Science, 2000, 290(5500): 2268-2269.

[9] 陈惠勇. 流形概念的起源与发展[J]. 太原理工大学学报:社会科学版, 2007, 25(3): 53-57.

[10] 徐　蓉,姜　峰,姚鸿勋. 流形学习概述[J]. 智能系统学报, 2006, 1(1): 44-51.

[11] NASTASE V. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation[C]//Conference on empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008: 763-772.

[12] SILVEIRA S B, BRANCO A. Extracting multi-document summaries with a double clustering approach[M]//Natural language processing and information systems. Berlin: Springer, 2012: 70-81.

[13] LIN C Y, HOVY E. Automatic evaluation of summaries using n-gram cooccurrence statistics[C]//Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 71-78.

[14] FERREIRA R, CABRAL L D S, FREITAS F, et al. A multi-document summarization system based on statistics and linguistic treatment[J]. Expert Systems with Applications, 2014, 41(13): 5780-5787.

(上接第 25 页)

trends in software methodologies, tools and techniques. [s. l.]: IOS Press, 2010: 3-36.

[9] LEVESON N G, HEIMDAHL M P E, HILDRETH H, et al. Requirements specification for process-control systems[J]. IEEE Transactions on Software Engineering, 1994, 20(9): 684-707.

[10] MOIR I, SEABRIDGE A, JUKES M. Civil avionics systems [M]. [s. l.]: John Wiley & Sons, 2013.

[11] PARNAS D L. Tabular representation of relations[D]. Canada: Telecommunications Research Institute of Ontario McMaster University, 1997.

[12] HEITMEYER C L, JEFFORDS R D, LABAW B G. Automated consistency checking of requirements specifications[J]. ACM Transactions on Software Engineering & Methodology, 1996, 5(3): 231-261.

[13] 张　鹏,刘　磊,刘华虓,等. Tabular 表达式的指称语义研究[J]. 软件学报, 2014, 25(6): 1212-1224.

[14] HATTON L. What is a formal method (and what is an informal method)? [C]//Proceedings of the 12th annual conference on computer assurance 1997. [s. l.]: IEEE, 1997: 125-126.

[15] WELLS A T, RODRIGUES C C. Commercial aviation safety [M]. [s. l.]: McGraw-Hill Professional, 2011.

万方数据