软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(2):289−298 [doi: 10.3724/SP.J.1001.2012.03999]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

动态多文档文摘模型*

刘美玲[1,2+], 郑德权[1], 赵铁军[1], 于 洋[2]

[1](教育部-微软语言语音重点实验室(哈尔滨工业大学),黑龙江 哈尔滨 150001)

[2](东北林业大学 信息与计算机工程学院,黑龙江 哈尔滨 150040)

Dynamic Multi-Document Summarization Model

LIU Mei-Ling[1,2+], ZHENG De-Quan[1], ZHAO Tie-Jun[1], YU Yang[2]

[1](Ministry of Education-Microsoft Key Laboratory of Speech Language (Harbin Institute of Technology), Harbin 150001, China)

[2](College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China)

+ Corresponding author: E-mail: mlliu@mtlab.hit.edu.cn, http://www.hit.edu.cn

Liu ML, Zheng DQ, Zhao TJ, Yu Y. Dynamic multi-document summarization model. *Journal of Software*, 2012,23(2):289−298. http://www.jos.org.cn/1000-9825/3999.htm

Abstract: This paper introduces two models to describe dynamic evolution of network information: identify and analysis the document collection on the same topic in different stages. In order to construct dynamic of evolution content differences, two dynamic multi-document summarization models are presented, which are matrix subspace analysis model, text similarity cumulative model. Based on these models, some efficient dynamic sentence weighting algorithms are implemented. Experiments on the test data of Update Summarization in TAC 2008 and comparative results between new models and TAC 2008 evaluation, shows the effectiveness of the models.

Key words: multi-document summarization; otherness analysis; matrix model; similarity cumulative; dynamic evolvement

摘 要: 从网络信息的动态演化性出发,对同一话题不同时序阶段的文档集合进行识别和分析,在度量演化内容差异性的基础上实现动态性,给出了两种实现动态多文档文摘的模型,即基于矩阵子空间分析和基于文本相似度累加的动态多文档文摘模型.在此基础上,提出了高效的动态句子加权方法.TAC 2008 的 Update Summarization 测试数据上的实验证明了所提出的动态多文档文摘模型的有效性.

关键词: 多文档文摘;差异性分析;矩阵模型;相似度累加;动态演化

中图法分类号: TP391 **文献标识码**: A

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,消除旧信息,提取新信息,使文摘随话题的演化而动态更新.

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.在 Web2.0 时代出现的 bbs、blog、twitter、在线评论等新媒体中的网络信息(如网络话题、热点事件等,

* 基金项目: 国家自然科学基金(60736014, 60773069, 61073130); 国家高技术研究发展计划(863)(2006AA010108)
收稿时间: 2010-10-14; 修改时间: 2010-12-09; 定稿时间: 2011-01-31

表现为一系列相关文章的集合)是动态演化的,它们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,不同时刻的话题内容之间具有关联性.动态文摘与静态文摘方法的最大区别在于,动态文摘需要在主题相关性的基础上考虑多个文档集之间的时序关系,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.

从分析历史文档与当前文档关系的角度,本文研究了动态演化环境下动态文摘求解模型,给出了多文档文摘动态模型的两种实现方法,即基于矩阵子空间分析(matrix subspace analysis method,简称 MSAM)和基于文本相似度累加(text similarity cumulative method,简称 TSCM)的动态多文档文摘生成方法.并在此基础上,利用静态文摘系统 ZStaticSummary[2]验证了本文所提出的方法的有效性.

本文第 1 节对相关工作进行介绍.第 2 节针对文档内容的动态演化性提出基于 MSAM 策略和基于 TSCM 策略的动态多文档文摘模型.第 3 节在此模型上验证新的文摘生成算法.第 4 节在 TAC 2008(Text Analysis Conference 2008)[3]数据集上对动态文摘方法的性能进行测试,并通过与会议中 Update Summarization 国际评测结果的对比,验证方法的有效性.最后给出本文的结论与展望.

1 相关工作

1.1 相关研究

所谓动态文摘就是分析历史文档和新出现文档所包含信息的关系,以历史文摘为基础,以当前文档为讨论对象,生成当前文档的文摘.方法是要提取当前文档的信息,并且过滤掉历史文摘中的信息.对动态演化的内容进行度量和建模.具体来说,就是对时序文档集中当前文档集 D_c 与历史文档集 D_h 之间的内容差异性进行建模.

动态内容的时序划分是动态文摘的基础,相关研究在新闻事件检测(news information detection,简称 NID)[4]和 TDT(topic detection and tracking)等领域[5]得到了较多的关注.时间信息在自然语言处理(natural language processing,简称 NLP)领域具有非常重要的意义[6],它是许多自然语言处理任务的基础,如多文档文摘(multi-document summarization)系统中需要按照时间顺序排列相关的信息,而在问答系统(question answering)中对“何时”问句的回答更是离不开时间信息等等.时间信息的重要作用使得时间表达识别和规范化(temporal expression recognition and normalization,简称 TERN)研究目前引起了国内外研究者们的广泛关注,国际上相关的评测有 ACE[7]中的 TERN 评测等.

新闻报道含有丰富的时间信息,所以时间信息在话题检测与跟踪研究方面起着非常重要的作用.目前,时间信息的使用以各种形式丰富了 TDT 的研究.比如,Juha Makkonen 将时间信息加入到报道的向量空间表示模型中,并尝试着将报道中的相对时间都转化为绝对时间[8];贾自艳等人提出了基于时间信息的相似度计算.Mani 等人使用时域分析方法对新闻事件的内容进行分析[9].

与传统的静态多文档文摘相比,动态多文档文摘仍然要面临文摘的内容选择和语言质量控制这两个关键问题.但不同的是,动态多文档文摘处理动态进化的相关文档集,具有很强的动态演化性.也就是说,如何在新的时序背景下去判定文摘内容的重要性、冗余性和覆盖性,以及保持文摘的语言质量,将成为问题的核心所在.

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 方法[10]、BE 方法[11]和人工评价金字塔(PYRAMID)方法[12].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、语言质量和整体反映度(综合考虑面向话题的覆盖度和流利度)进行评测.对于标准文摘的构建,官方有 8 个 NIST 评测者为各话题选择和撰写文摘,话题中的每个时间片均对应 4 个人工文摘.这样,人工文摘的质量将作为系统性能的上限,而基准系统的文摘(一般由文档中的首句构成)质量将作为系统性能的下限.文摘内容单元的选择和对比是文摘评价的两个关键问题.

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 NIS(National Institute of Standards and Technology)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 动态多文档文摘的建模方法

2.1 动态多文档文摘模型的基本概念

动态文摘的关键问题是如何对动态信息的演化性内容进行表示[13],具体来说,就是对时序文档集中当前文档集 D_i 与历史文档集 $D_1,\ldots,D_{i-1}(1\leqslant i\leqslant n)$之间的内容差异性如何建模.为了方便叙述,首先给出如下定义:

定义 1. 把时序文档序列中当前文档集 D_i 包含的信息称为当前信息(current information),用 I_c 表示.

定义 2. 把时序文档序列中历史文档集 $D_1,\ldots,D_{i-1}(1\leqslant i\leqslant n)$包含的信息称为历史信息(history information),用 I_h 表示.

定义 3. 用 f 表示从文档空间到文摘空间的映射关系,则时序文档序列中任一文档集 D_i 的文摘可记为 $f(D_i)$;同样地,历史信息 I_h 的文摘称为历史文摘,表示为 $f(I_h)$;当前信息 I_c 的文摘称为当前文摘,表示为 $f(I_c)$.

在以上定义的基础上,可以把动态文摘问题 LDynSummary$(D_i|D_1,\ldots,D_{i-1})(1\leqslant i\leqslant n)$转化为对历史信息 I_h 和当前信息 I_c 之间演化内容的差异性建模和求解.本文对历史信息 I_h 和当前信息 I_c 的演变关系进行了分析,采用文档内容过滤的方法来刻画动态演化的内容.

新信息可以通过从当前信息 I_c 中过滤掉与历史信息 I_h 相重叠的内容得到,表示为 $I_c–I_h$.然后,利用静态文摘方法生成动态文摘 $f(I_c–I_h)$.这种动态文摘模型从文档过滤的角度提取动态信息以生成文摘.考虑到文摘对文档内容的代表性,为了节省计算代价,可以将过滤对象 I_h 替换为历史文摘 $f(I_h)$,如图 1 所示.

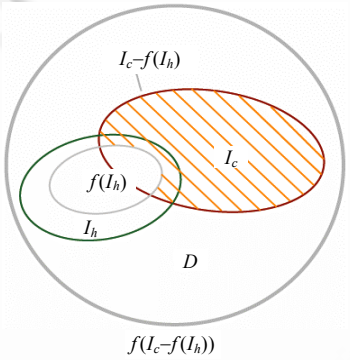


Fig.1 Document filtering model

图 1 文档过滤模型

2.2 基于矩阵子空间分析的模型

为了把握内容的动态演化趋势,分析历史信息与当前信息的相似性和差异性,可以从两个角度入手:一是过滤相似内容以刻画动态演化的差异性,二是提取差异内容以刻画动态演化的差异性.基于过滤相似内容方法已有所研究,因此,本文将从提取差异性内容角度入手,实现对动态内容的差异性建模.

在信号处理领域中,许多问题的最优化求解都可以归结为从含噪声的信号中提取某个所希望的信号,抑制其他所有干扰,如杂波或者噪声.正交投影法是解决这一问题的一个极为重要的数学工具.当两个子空间正交时,常采用正交投影法进行参数的最优估计或者信号的最优滤波,因为观测数据向量在被投影的子空间上的分量可以被提取,理论上还可以完全消掉观测数据向量在另一个正交子空间上的分量.在动态多文档文摘领域,动态文摘的主要目的是在读者已阅读相关历史文档集、了解相关内容的基础上,从相同主题的当前文档集中识别出新信息并且提取出新信息.从形式上观察,其任务与信号处理领域中的信号分解与最优化问题有相似之处,因

此可以利用信号处理的分析方法来解决动态多文档文摘的相关问题.目前在该领域中,已有不少方法可以用来把文档集抽象化为相应的数学模型,并且都取得了不错的效果.本文中,将利用线性代数中的矩阵为文档集建模.假设主题词集合中的主题词数为 m,文档集句子集合中的句子数为 n,以主题词集合为矩阵行,文档集的句子集合为矩阵列,本文把历史文档集模型化为一个 $m\times n$ 稀疏矩阵 A.文档集可被形象地描述为一个文档信息空间.

本文将历史文档集抽象为历史信息空间,将当前文档集抽象为当前信息空间.数学上,一个空间通常可以分解为两个相互正交的子空间,即正交空间 P 和正交补空间 P_T,两个子空间的直和即为整个信息空间.对于矩阵 A,因为矩阵列为文档集句子集合,所以矩阵的列空间就包含了文档集的主要信息,称为文档集的信息子空间 C.矩阵 A 中还存在一个与列空间相互正交的子空间,为列空间的正交补空间 D,也称为 W^H 的零空间.它不包含文档集的主要信息,包含的主要是一些噪声信息.本文中定义它为噪声子空间.

信息子空间 C 以及其正交补空间噪声子空间 C 是历史信息空间的正交分解,两子空间的直和就是整个历史信息空间.按如下公式可计算文档集矩阵 A 的信息子空间和噪声子空间空间.主要信息子空间 C 和噪声子空间 D,如公式(1)和公式(2)所示:

$$C=A\langle A,A\rangle^{-1}A^T \tag{1}$$

$$D=I-A\langle A,A\rangle^{-1}A^T \tag{2}$$

其中,A 是历史信息空间矩阵.

同样地,把当前文档集模型化为一个 $m\times n$ 的矩阵,用 B 表示,称为当前信息空间.它包含当前文档集中的所有信息,其中包括新颖性信息和历史信息.历史信息是当前文档集中包含的重复性信息,读者对其已有了解,因此它对当前文档集来说是垃圾信息.为了减少信息获取者获取新信息的时间,应使文摘中包含尽可能少的冗余信息,使读者以最少的时间获取最有价值的信息.动态多文档文摘的主要任务是以历史文档集所含信息为基础,从中识别并抽取出最新的信息.完成该任务的第 1 步是要正确地识别新信息,如果不能正确地识别新信息,也就无所谓抽取信息形成文摘了.在信号处理中,解决最优滤波问题的最佳方法是用子空间方法计算复合信号到某一特定子空间的正交投影,然后加以提取.受此启发,我们提出用子空间方法来计算当前信息空间 B 到历史信息空间的噪声子空间 C 上的投影,投影中将只包含新信息,不包含历史信息,达到了识别新信息的目的.然后,提取此新信息,运用相关算法对其进行组织,即可生成当前文档集的动态文摘.

2.3 基于文本相似度累加的模型

文本相似度计算是自然语言处理、智能检索、文档聚类、文档分类、自动应答、词义排歧和机器翻译等很多领域的基础研究课题.本文研究的动态多文档文摘的动态性体现在分析历史文档和当前文档间的关系上,因此在文档内容过滤的差异性描述方面,可以利用文档集合中句子级的文本相似度累加运算.找出当前文档集合中与历史文档集合内容差异最大的句子集合,对当前文档过滤后的内容进行重排序,留下的差异度大的句子集最能反映同一主题下信息演化的动态特性,生成的多文档文摘也是反映当前时序阶段内容的侧重点.受矩阵子空间分析方法的启发,实验了基于文本相似度累加的动态文摘模型.

3 动态多文档文摘的建模方法

3.1 基于矩阵子空间分析的方法

方法的基本思想是:根据各类的训练样本,由原始模式特征空间产生各类对应的子空间,所求出的这些子空间的基本矢量分别反映了各类模式分布结构信息,每个子空间与每个类别一一对应.本文的动态文摘用基于代数的方法和基于统计迭代学习的方法建立子空间.

子空间法实质上是每类分别处理,动态文摘的文档过滤原理也是分类处理,对原始数据进行结构分析,将每类最重要的特征提取出来,实现特征提取、数据压缩,高维空间向低维空间的线性映射.子空间法由于加大了各类的表示差别,因此对动态文档内容差异性的识别效率更高.在子空间法中,通常采用内积运算,这可使计算量大为减少.因此,考虑用矩阵子空间的方法来提高文档内容过滤的质量和动态性,具体实现过程如算法 1 所示.

算法 1. 基于 MSAM 的多文档文摘算法.

1. 生成历史文摘,提取历史文摘的指定数量的主题词集合 A,提取当前文档集的指定数量的主题词集合 B,对集合 A 和 B 进行合并,形成并集 C,从集合 C 中取出指定数目主题词形成主题词集合 D.
2. 以集合 D 中的主题词为矩阵列,以历史文摘中的句子集合为矩阵行形成历史信息矩阵 X;以集合 D 中的主题词为矩阵列,以当前文档集中的每篇文档的句子集合为矩阵行形成文档集中文档数目的当前信息矩阵 Y_n.
3. 对历史信息矩阵 X 进行子空间分解,用公式 $P=X\langle X,X\rangle^{-1}X^T$ 和公式 $P_\perp=I-X\langle X,X\rangle^{-1}X^T$ 分别计算矩阵 X 的正交空间和正交补空间.
4. 对 n 个当前文档矩阵分别计算其在历史信息矩阵 X 上的正交投影,即 $Z_n=P_\perp X_n$.
5. 然后根据 Z_n 的行与 X_n 的行的相似度进行取舍,当 Z_n 中的行与 X_n 中的对应行的相似度小于某一设定值时,在与此矩阵对应的文档中删除矩阵行对应的句子.
6. 用自动多文档文摘生成方法对处理过的文档集进行处理,生成动态文摘.

3.2 基于文本相似度累加的方法

从当前信息 I_c 中过滤掉与历史信息 I_h 中相似的信息内容,过滤剩下的句子集合就是动态文摘系统的文档集合.其核心思想是,计算句子级的相似度累加,以区别于以往的词级相似度.由此度量内容过滤的准确性,提高了文摘的语义理解性能和可读性.首先,动态地对历史文档集合和当前文档集合做句子相似度计算,通过句子相似度值的累加计算,过滤掉重复的信息,对当前文档句子集合做重新排序,以抽取出相似度值最低的句子集合.设置一个阈值 n,n 表示过滤掉的句子的数目.$count(S^c)$表示当前文档句子数,$count(S^c)–n$ 表示当前文档中包含新信息的句子数.

具体实现过程如算法 2 所示.

算法 2. 基于 TSCM 的多文档文摘算法.

1. 设历史信息句子集合为 S^h,当前信息句子集合为 S^c,统计 S^h 中每个句子的长度.
2. 对 S^c 中的每个句子 S_i^c ($0<i<count(S^c)$)计算 S_i^c 与 S_i^h ($0<j<count(S^h)$)相似度值 Sim_i^j,应用新提出的动态 TF-IDF-ISF 方法进行相似度计算(具体详见第 3.3.3 节),如公式(3):

$$Sim(senti,sentj)=\frac{\sum_{k=0}^{count}TF\times IDF\times ISF(tk)}{Length(sentj)} \tag{3}$$

其中,$Sim(senti,sentj)$表示句子 $senti$ 与 $sentj$ 的相似度;$TF\times IDF\times ISF(tk)$表示词语 tk 的权重,其中,tk 表示 $senti$ 和 $sentj$ 中共同出现的词语,k 表示 $senti$ 和 $sentj$ 中共同出现的词语的数量;$Length(sentj)$表示 $sentj$ 中词语的数量.

3. 计算相似度的累加迭代的值,按公式(4)计算:

$$Weight(S_i^c)^{(j)}=Weight(S_i^c)^{(j-1)}+Sim_i^j \tag{4}$$

其中,$Weight(S_i^c)^{(j)}$表示句子 S_i^c 与 S^h 中前 j($0<j<count(S^h)$)个句子相似度值的累加.

4. 设置一个阈值 n,n 表示过滤掉的句子的数目.根据当前文档句子集合 S^c 中的句子 S_i^c ($0<i<count(S^c)$)的 $Weight(S_i^c)^{(j)}$($j=count(S^h)$)值,按照从高到底的顺序排列,删除当前文档句子集合 S^c 中的前 n 个句子.$count(S^c)$代表当前文档集合中句子的数目.剩下的 $count(S^c)–n$ 个句子就是当前文档中包含新信息的句子.

3.3 动态多文档文摘句子加权方法

本文提出的 3 种句子加权方法,分别是短语的信息粒度表示方法、动态 TF-IDF-ISF 的句子加权方法和句子相似度计算的句子加权方法.对基于矩阵子空间分析的动态文摘模型而言,其第 6 步是运用相应的静态文摘方法对处理后的句子集合进行处理以生成文摘.本文在该模型中实验了下述 3 种加权方法,并对实验结果进行

测试,验证了该动态模型和相应句子加权方法的有效性.同时,在第 2 种动态模型方法上也实验了这 3 种句子加权方法,依然达到了良好的效果.

3.3.1 基于短语的信息粒度表示方法

信息粒度(information granularity)[14]是反映信息详细程度的概念.为了适应不同子系统信息需求的详细程度不同而设置不同的粒度,以描述使用该知识对论域划分的分类情况.信息粒度是指信息单元的相对大小或粗糙程度,对文摘来说,内容的信息粒度可以是篇章、段落、句子、事件、短语、关键词、子话题等.

本文提出的基于短语的信息粒度表示方法如公式(5)所示:

$$Weight(senti)=\sum_{j=1}^{lenth(senti)} Phrase\_Weight(j)+SentLength\_Weight(senti) \tag{5}$$

其中,*Phrase_Weight*(*j*)是句子中短语的权重,计算方法如公式(6)所示:

$$Phrase\_Weight(j)=\frac{FR(Phrase)}{\mathrm{Max}FR} \tag{6}$$

其中,*FR*(*Phrase*)是短语的词频,Max*FR* 是词频最大短语的词频,*SentLength_Weight*(*senti*)是句子长度权重.

在实际计算过程中,如果 *SentLength*<*x*,则舍掉此句,不参与运算.

如果 *SentLength*≥*x*,则 *SentLength_Weight*(*senti*)=0;*x* 是动态的句子长度阈值,根据实际需要进行调整.

3.3.2 基于动态 TF-IDF-ISF 的句子加权方法

TF-IDF(term frequency-inverse document frequency)[15]的概念被公认为信息检索中最重要的发明.TF-IDF 是一种常用的、有效的词汇加权算法,在 TF-IDF 词汇权重计算中,TF(term frequency)称为词频,用于计算该词汇描述文档内容的能力;IDF(inverse document frequency)称为反文档频率,用于计算词汇区分文档的能力.

本文提出了基于动态 TF-IDF-ISF 的句子加权方法,如公式(7)所示:

$$Score(senti)=\alpha\times fWordWgt+\beta\times fPosWgt+\gamma\times TimeWgt \tag{7}$$

其中,*Score*(*senti*)表示句子 *senti* 的打分; $fWordWgt=\sum_{k=0}^{count(senti)} TF\times IDF\times ISF$, 表示句子 *senti* 的词语权重,其中,*TF*(*w*)表示词 *w* 在文档中的出现频率,*IDF*(*w*)表示词 *w* 的反文档频率,*ISF*(*w*)表示词 *w* 的反句子频率,*f*(*w*)是频率统计函数,*DF*(*w*)为整个文档集合中包含词 *w* 的文档数,*SF*(*w*)为整个文档集中包含词 *w* 的句子数.

这里,*TF*=*f*(*w*),$IDF=\frac{1}{DF(w)}$,$ISF=\frac{1}{SF(w)}$.

ISF(inverse sentence frequency)用于计算词汇区分句子的能力,称为反句子频率.在一个文档集中,词汇的频率越大,出现的文档数就越少(包含这个词的文档数);同时,出现的句子数越少(出现这个词的句子),说明此词汇在同一句子中出现的次数越多,那么它就是被强调的词语,它是文档集主题词的可能性将非常大.

fPosWgt 表示 *senti* 的位置权重, $Position\_Weight(senti)=\frac{1}{i}$;

$TimeWgt=\frac{time}{count(D)}$ 表示句子 *senti* 的时间信息值,其中,*time* 表示句子 *senti* 所属文档的出版时间在文档集合中的排序值,*count*(*D*)表示文档集合中文档的数量.

在实际计算过程中,当 *SentLength_Weight*(*senti*)>*x* 时,*SentLength_Weight*(*senti*)=0;

当 *SentLength_Weight*(*senti*)<*x* 时,舍掉此句.

3.3.3 基于句子相似度计算的句子加权方法

相似度是一个很复杂的概念,在语义学、哲学和信息理论共同体中被广泛地讨论.在不同的具体应用中,相似度的含义有所不同.例如,在基于实例的机器翻译中,相似度主要用于衡量文本中词语的可替换程度;在信息检索中,相似度更多地是反映文本与用户查询在意义上的符合程度;在自动问答中,相似度反映的是问题与答案的匹配程度;而在多文档文摘系统中,相似度可以反映出局部主题信息的拟合程度.

基于句子相似度计算的句子加权方法如公式(8)所示:

$$Weight(S_i)=\sum_{j=1}^{count} Sim_{ij} \tag{8}$$

其中,$Weight(S_i)$(0<i<$count$)表示句子 S_i 的权重;$count$ 表示文档集合中句子的总数;$Sim_{ij}=\frac{Sim\_Length(S_i,S_j)}{Length(S_j)}$,表示句子 S_i 与 S_j 的相似度值.

4 实验结果与分析

4.1 实验数据及评价

在 DUC 2007 中,Update Summarization 作为一个先导任务,测试语料从主要任务的 45 个话题中挑选 10 个话题,每个话题由 3 个连续进化的时间片组成,即每个话题有 3 个时序相关的文档集 A,B,C,分别包含 10,8,7 个文档.假定读者对先前的文档有一定的了解,Update Summarization 任务的目的就是针对每一个时间片给出 100 字的文摘,该文摘反映沿着时间线的内容进化.可见,文摘的发展方向在发生变化.

在 TAC 2008 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 2 个时间片,且均由 10 个文档组成.例如,话题“D0801A”由两个时间片“D0801A-A”和“D0801A-B”组成,二者内部的 10 个文档分别由各自的 id 来表示,话题本身由〈title〉和〈narrative〉来描述.

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.本文在 TAC 2008 的 Update Summarization 测试数据上,将动态文摘结果的 ROUGE-2(R-2)和 ROUGE-SU4*(R-SU4*)得分与 TAC 2008 Update 实际系统的得分进行了对比,结果表明,本文的动态多文档文摘方法具有良好的性能.

4.2 实验结果

本文做了 4 组实验:第 1 组实验是在子空间分析策略模型下,3 种不同加权方法生成的动态多文档文摘评测结果比较;第 2 组实验是在文本相似度累加动态多文档文摘模型下,3 种不同句子加权方法生成的动态多文档文摘评测结果比较;第 3 组实验测试文本相似度累加模型中,文档过滤的句子数对动态多文档文摘的影响;第 4 组实验是与 TAC 2008 测试系统的性能进行对比.

实验 1 是在子空间分析策略模型下,3 种不同加权算法生成的动态文摘的 ROUGE 评价结果.表 1 中 3 种加权算法的 ROUGE 得分的对比可以看出,MSAM_2 的 ROUGE_2 和 ROUGE_SU4*分数比基础系统 MSAM_1 的打分要高,说明基于 TF-IDF 的加权方法的性能比基于相似度累加的加权方法的性能要好.MSAM_3 两项打分均提高很多,说明基于短语级的句子加权方法的性能较好.因此,基于短语的句子加权方法比上述两种方法都好.因此,它能够较为有效地进行有用句子的提取,从而获得较好的文摘质量.

Table 1 Abstract performance comparison of three algorithms in MSAM

表 1 MSAM 的 3 种算法的文摘性能对比

系统标号	形态	句子加权方法	R-2	R-SU4*
MSAM_1	动态子空间模型	句子相似度累加	0.042 22	0.090 33
MSAM_2	动态子空间模型	TF-IDF	0.052 27	0.104 2
MSAM_3	动态子空间模型	短语级加权方法	**0.052 27**	**0.121 54**

实验 2 是在文本相似度累加动态多文档文摘模型下,3 种不同加权算法生成的动态文摘的 ROUGE 评价结果.由表 2 的对比可以看出,文本相似度累加模型中应用短语的信息粒度表示方法进行句子加权,ROUGE_SU4*效果较好,应用动态 TF-IDF-ISF 的句子加权方法 ROUGE_2 得到了很好的成绩.但是,实验过程中发现,TSCM 模型的系统文档处理速度很快,可以应用到大数据量的实用多文档文摘系统中,因此在物理性能上比 MSAM 模型系统要好很多.

Table 2 Abstract performance comparison of three algorithms in TSCM

表 2 TSCM 的 3 种算法的文摘性能对比

系统标号	形态模型	句子加权方法	R-2	R-SU4*
TSCM_1	动态文本相似度累加模型	句子相似度累加	0.063 32	0.115 3
TSCM_2	动态文本相似度累加模型	TF-IDF-ISF	**0.092 35**	0.109 7
TSCM_3	动态文本相似度累加模型	短语级加权方法	0.050 13	**0.138 6**

实验 3,为了研究文本相似度模型中文档过滤的句子数对动态文摘的影响,我们对受句子长度影响最大的 TSCM_3 进行分析.实验在 10~110 的取值范围内对句子个数进行调整,并用 ROUGE 指标对其生成的动态文摘进行评价,实验结果如图 2 所示.

文摘长度阈值选取测试结果如图 2 所示.

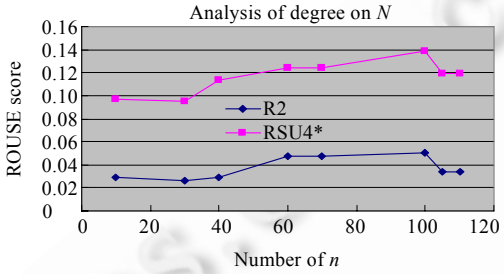


Fig.2 TSCM_3 model threshold analysis of abstracts length

图 2 TSCM_3 模型文摘长度阈值分析

由图 2 可以看出,TSCM_3 模型随着句子个数改变的变化趋势,在句子数为 100 时达到最优性能.此时,在 ROUGE-2 和 ROUGE-SU4*上的得分分别为 0.05013 和 0.13860,在 ROUGE-2 上进一步缩小了与第一名的差距,同时在 ROUGE-SU4*上更加凸显了 TSCM_3 模型的优越性.

实验 4 是 MSAM 模型和 TSCM 模型的最好性能与参加 TAC 2008 Update 任务评测前 3 名的实际系统的性能对比,见表 3.其中,MSAM_3 模型整天性能接近最好成绩;TSCM_3 中,短语信息粒度的句子加权方法的 R-2 比第一名的系统稍差,但是 R-SU4*的分数已经超过了第一名.TSCM_2 模型中,基于动态 TF-IDF-ISF 的句子加权方法 R-2 分数处于前列,与第三名一样,整体分数进入了所有评测系统的前 5 名.说明这两种动态文摘模型在不同侧重点上具有很好的性能和潜力.

Table 3 Comparison with TAC 2008

表 3 与 TAC 2008 比较

系统	R-2	R-SU4*
MSAM_3	0.052	0.121
TSCM_2	0.092	0.109
TSCM_3	0.050	0.138
Rank 1	0.101	0.137
Rank 2	0.097	0.134
Rank 3	0.092	0.132

4.3 实验分析

由以上实验结果可以看出,本文提出的两种模型在不同程度上体现了各自的优势.综合分析,基于矩阵子空间分析方法的动态文摘模型的打分与 TAC 2008 官方发表的 73 个系统结果相比排在中上,说明该模型是性能较好的模型.该模型有其独特的优势.此模型将历史文档集和当前文档集分别模型化为历史信息空间和当前信息空间,从信息空间的角度来分析问题.首先对历史信息空间正交分解,将其分解为主要信息空间和噪声空间;然后计算当前信息空间在噪声空间中的投影,以识别出其中的新信息;最后,提取新信息构成文摘.模型算法始终

以整个文档集作为研究对象,从全局出发分析问题.它生成的文摘概括面较广.

MSAM 始终在全局层面分析问题,其文摘具有很强的全面性,不会遗漏任何重要性信息,适用于在全面性要求高、不允许有信息遗漏的场合为文档集生成文摘.ROUGE 评测工具的 ROUGE-2 打分时衡量文摘显著性(文摘信息对文档集主要信息的代表性)的指标,ROUGE-SU4*打分是衡量文摘动态性(文摘包含新信息的比例)的指标.

基于文本相似度累加模型实验打分位于 TAC 2008 发布的所有系统结果的中上位置.当模型算法第 6 步所使用的加权方法是以短语为信息粒度的加权方法时,其 ROUGE-SU4*打分已高出了 TAC 2008 中第一名的系统.可见,该模型是一种动态性很强的模型,能够很好地对文档演化性内容的差异性建模,很好地识别出文档集中的历史信息,进而使生成的文摘具有较强的动态性.TSCM 适用于文档信息更新很快的领域.它能最大限度地为读者识别新信息,降低读者获取信息的代价.此外,使用基于动态 TF-IDF-ISF 的句子加权方法时,总的 ROUGE-2 打分仅次于第三名的系统,表明通过修改加权方法也可使该模型生成的文摘具有较强的显著性.SCM 使用不同的加权方法时,可使模型分别生成具有动态性强和显著性强的文摘.模型的灵活性高,算法调整可适用于有不同要求的领域,是一种适应性很强的动态文摘模型.

上述两种动态文摘模型都能够较为有效地过滤冗余的历史信息,提取有用的动态信息生成动态多文档文摘,从而给出较好的动态文摘.本文提出的 3 种句子加权方法均在评测分数上有良好表现,因此促进了动态多文档文摘模型和方法的发展,具有很重要的研究价值.

5 结 论

动态多文档文摘模型的研究是近年来新兴的研究热点,目前正处于起步阶段.本文在认真研究国内外多文档文摘领域最新发展的基础上,对动态内容的演化关系进行了差异性分析,采用内容过滤的方法刻画演化内容,从而提出了两种动态文摘模型.本文还在句子加权方法上做了改进,应用了基于短语信息粒度的加权方法,在两种模型上均验证了 3 种动态多文档文摘生成方法.在 TAC 2008 的 Update Summarization 测试数据上进行的实验表明了本文所提出的动态多文档文摘模型及生成方法的有效性.下一步将结合内容差异性和主题相关性进行研究,抽取各个模型的优点进行整合,进一步提高静态文摘算法的性能,使动态多文档文摘向基于理解的文摘领域靠近.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是哈尔滨工业大学计算机科学技术学院郑德权副教授、赵铁军教授的指导表示感谢.

References:

[1] Mani I. Automatic Summarization. John Benjamins Publishing Company, 2001.

[2] Zhang S, Zhao TJ, Yu H, Zhao H. The research on the influence of the types of document sets on multi-document summarization. Journal of Computational Information Systems, 2007,3(3):1201−1206.

[3] Dang HT, Owczarzak K. Overview of the TAC 2008 Update Summarization Task. In: Proc. of the Text Analysis Conf. 2008.

[4] Allan J, Jin H, Rajman M, Wayne C, Gildea D, Lavrenko V, Hoberman R, Caputo D. Topic-Based novelty detection. Technical Report, ws99, Baltimore: Center for Language and Speech Processing, Johns Hopkins University, 1999.

[5] Allan J, Papka R, Lavrenko V. On-Line new event detection and tracking. In: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Melbourne, 1998. 37–45. [doi: 10.1145/290941.290954]

[6] Mani I. Recent developments in temporal information extraction. In: Nicolov N, Mitkov R, eds. Proc. of the RANLP. 2004.

[7] http://projects.ldc.upenn.edu/ace/intro.html

[8] Makkonen J. Investigations on event evolution in TDT. In: Proc. of the Student Workshop of Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics. Edmonton, 2003. 43–48. [doi: 10.3115/1073416.1073424]

[9] Mani I, Wilson G. Robust temporal processing of news. In: Proc. of the 38th Annual Meeting on Association for Computational Linguistics. Hong Kong, 2000. 69−76. [doi: 10.3115/1075218.1075228]

[10] Lin CY, Hovy E. Automatic evaluation of summaries using *N*-gram cooccurrence statistics. In: Proc. of the 2003 Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003). Morristown: Association for Computational Linguistics, 2003. 71−78. [doi: 10.3115/1073445.1073465]

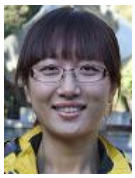
[11] Lin CY, Hovy E. The automated acquisition of topic signatures for text summarization. In: Proc. of the 18th COLING Conf. Saarbrucken, 2000.

[12] Nenkova A, Passonneau R, McKeown K. The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. on Speech and Language Processing, 2007,4(2):Article 4. [doi: 10.1145/1233912.1233913]

[13] Zhang J, Cheng XQ, Xu HB. GSPSummary: A graph-based sub-topic partition algorithm for summarization. In: Proc. of the 4th Asia Information Retrieval Symp. (AIRS 2008). Harbin, 2008. [doi: 10.1007/978-3-540-68636-1_31]

[14] Li WJ, Wu ML, Lu Q, Xu W, Yuan CF. Extractive summarization using inter-and intra-event relevance. In: Proc. of the 44th Annual Meeting of the Association for Computational Linguistics, Vol.44. 2006. 369−372. [doi: 10.3115/1220175.1220222]

[15] Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. In: Proc. of the Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 412−420.



刘美玲(1981－),女,黑龙江佳木斯人,博士生,讲师,CCF 会员,主要研究领域为信息检索,信息处理和自动文摘技术.

赵铁军(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,人工智能.

郑德权(1968－),男,博士,副教授,主要研究领域为自然语言处理,知识工程,信息检索.

于洋(1977－),男,博士生,助理研究员,主要研究领域为人工智能,地理信息系统,信息检索.