基于热传导模型的更新摘要算法*

杜 攀 郭嘉丰 张 瑾 程学旗 张 旭

(中国科学院计算技术研究所 网络数据科学与工程研究中心 北京 100190)

摘 要 更新摘要除了要解决传统的面向话题的多文档摘要的两个要求——话题相关性和信息多样性,还要求应对用户对信息新颖性的需求. 文中为更新摘要提出一种基于热传导模型的抽取式摘要算法——HeatSum. 该方法能够自然利用句子与话题,新句子和旧句子,以及已选句子和待选句子之间的关系,并且为更新摘要找出话题相关、信息多样且内容新颖的句子. 实验结果表明,HeatSum 与参加 TAC 09 评测的表现最好的抽取式方法性能相当,且更优于其它基准方法.

关键词 更新摘要, 面向话题的多文档摘要, 热传导模型

中图法分类号 TP 391

Update Summarization Based on Heat Conduction Model

DU Pan, GUO Jia-Feng, ZHANG Jin, CHENG Xue-Qi, ZHANG Xu

(*Research Center of Web Data Science and Engineering*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

ABSTRACT

Besides the problems of topic relevance and information diversity tackled by traditional topic-focused multi-document summarization, the update summarization is required to address the problem of information novelty as well. In this paper, HeatSum, an extractive approach based on heat conduction for update summarization, is proposed. The process can naturally make use of the relationships among the given topic, the old sentences, the new sentences, and the sentences selected and to be selected to find proper sentences for update summarization. Therefore, HeatSum is able to simultaneously address the challenging problems above for update summarization in a unified way. The experiments on benchmark of TAC 2009 are performed and the ROUGE evaluation results show that the HeatSum achieves fine performance compared to the best existing performing systems in TAC tasks and it significantly outperforms other baseline methods.

Key Words Update Summarization, Topic-Oriented Multi-Document Summarization, Heat Conduction Model

* 国家自然科学基金重点项目(No. 60933005)、国家自然科学基金项目(No. 60903139, 61003166)和国家 863 计划项目(No. 2010AA012500)资助
收稿日期:2010-10-13;修回日期:2011-03-04

作者简介 **杜攀**,男,1981 年生,博士,主要研究方向为网络挖掘. E-mail:xiaopandu@ gmail. com. **郭嘉丰**,男,1980 年生,博士,主要研究方向为社会搜索、网络挖掘、信息检索. **张瑾**,男,1978 年生,博士,主要研究方向为文本挖掘、自动文摘. **程学旗**,男,1971 年生,博士,研究员,主要研究方向为网络科学、网络搜索与数据挖掘、P2P 与分布式系统、信息安全. **张旭**,女,1983 年生,硕士,主要研究方向为视频内容分析与检索.

1 引 言

随着网络信息的不断增长,自动文本摘要技术引起人们越来越多的关注. 同时由于网络用户的参与程度越来越深,网络信息的动态变化性也大大加强,这就对传统的自动文摘技术提出新的要求——自动文摘要求能够给用户提供最新的信息,剔除那些用户已接触过的过时信息. 比如,用户对某个新闻话题感兴趣,不断追踪该话题的最新进展,他就会希望摘要能够实时提供该话题的最新进展. 然而,传统的多文档摘要技术主要针对静态文档集合操作,不能满足用户对持续更新信息的摘要需求. 因此,更新摘要技术成为自动文摘领域最近研究的热点.

更新摘要假定用户已读过关于某个话题的一些文档,要求对该话题的最新出现的文档集合进行摘要,提供给用户关于该话题的最新动态. 从这个要求出发,更新摘要需解决的主要问题如下. 1)话题相关性:摘要是针对某个话题的多个文档进行的,话题是对用户信息需求的表达(查询句,或者描述),因此摘要必须同用户关注的话题密切相关. 2)内容多样性:摘要内部必须尽可能地避免冗余信息来保持简洁. 这样才能在有限的摘要长度内,容纳尽可能丰富的信息,尽可能覆盖话题的多个侧面. 3)信息新颖性:在给定某话题和关于该话题的按时间顺序排列的两个文档集合,摘要需要关注后一个文档集(即新文档集)所包含的与前一个集和不同的新信息.

传统的面向话题的多文档摘要主要关注前两个问题,作为面向话题的多文档摘要在时间维度上的扩展,更新摘要则强调第 3 个问题,它必须解决动态环境下摘要的更新问题,给予新颖的文档内容更多的关注. 现有的大多数更新摘要方法都可归类为抽取式方法. 抽取式方法不需要深层次的自然语言处理技巧,而是从目标文档集中挑出完整的句子直接作为摘要的句子使用. 目前,已出现许多不同的抽取式方法来应对更新摘要中的信息新颖性的需求,其中包括使用时间衰减因子的方法[1-2]、句子过滤的方法[3]、子话题分析的方法[4]、正负影响扩散传播的方法[5]等. 然而,多数方法[2-5]都不能通过一个统一的模型同时处理上述三个问题,相反,通常需要额外的启发性步骤来单独解决信息多样性问题[2,5]或信息新颖性问题[3].

本文提出一种基于热传导的抽取式更新摘要方法——HeatSum. 该方法通过句子网络上的热传导过程对句子进行排序,排序结果能同时反映句子的话题相关程度和内容新颖程度,且内容新颖的句子. 热传导过程中的源点和汇点在排序过程中能够担负着对句子的正向激励作用和负向惩罚作用. 这样,通过将话题描述设置为网络中的源点,排序过程就能对与话题描述相似的句子赋予一个正向的激励,也就是说,这些句子的排序得分得以加强,从而解决更新摘要的话题相关性的要求. 类似地通过将旧文档集中的句子设置为汇点,那么与汇点相似的句子得分就会被惩罚,从而解决更新摘要中的信息新颖性的需求. 把已选的摘要句子迭代地设置为汇点,就能保证待选句与已选句之间的差异,从而达到摘要句子之间的多样性. 这样,HeatSum 能够通过一个统一的过程同时解决更新摘要所要应对的 3 个问题.

在 TAC 2009 (http://www.nist.gov/tac/)的数据集上的实验表明,HeatSum 与 TAC 09 上表现最好的抽取式参评方法性能相当,同时较优于 TAC 提供的基准方法.

2 相关工作介绍

2.1 自动文本摘要

抽取式摘要是自动文本摘要的一个重要类型. 通过直接从目标文档集合中抽取句子作为摘要的方式,使得抽取式摘要不仅实现简单,而且语法规范,可读性好. 在面向话题的多文档摘要中得到广泛应用. 一个早期的工作是 Erkan[6]的 LexRank 方法,该方法将文本表示成有权无向图,在图上通过一个类似 PageRank 的算法来抽取重要的句子来组成摘要. 该方法的优点是能充分利用句子之间的关系获得重要且相关的句子内容,缺点是它无法表达更新摘要对于新颖性的需求,也不能反映自动文摘关于多样性的信息需求. Lin 等[7]通过使用多个特征,如句子位置、词频、话题签名等来识别重要的句子作为摘要候选句. 该方法的特点是简单快捷,缺点该方法的启发特征明显,理论贡献有限,且同样无法同时满足自动文摘的多个目标. Mihalcea 等[8]提出一种基于图的抽取式多文档摘要方法——TextRank,它实质上也是类 PageRank 的一个变种,其基本思路同 Erkan 等[6]的方法类似.

最近,又出现很多抽取式的面向话题的多文档摘要方法[9-13]. Otterbacher 等[9]提出一种基于随机行走模型的面向问题的摘要方法,该方法的优点是能够反映话题相关性,缺点同其它基于 PageRank 算法的方法一样,无法捕获新颖性和多样性. Wan 等[11]提出使用流形排序的方法来实现面向话题的多文档摘要,该方法通过基于图的正则化的方法得

到一个句子排序,其优点是能够利用句子间的全局关系进行句子打分,缺点是也需要额外的启发性步骤来实现新颖性和多样性. Zhang 等[13] 提出基于 Boosting 框架的 AdaSum 方法来实现面向话题的多文档摘要. 该方法的优点在于它能够同时获得精准的话题描述和话题相关的自动文摘,缺点也是需要启发性步骤实现文摘的新颖性和多样性. 还有很多其它面向话题的多文档摘要方法[14-15] 在评测会议 DUC 和 TAC 上提出和使用.

更新摘要是面向话题的多文档摘要在时间维度上的扩展. 它更多地关注包含在新文档集合中的不同于旧文档集合所传达信息的最新动态. 在网络的动态环境下,更新摘要显得更具现实意义. Allen 等[16] 针对某个事件的新闻报道的时序摘要的研究工作,可看作是更新摘要的一种早期形式. 最近,Wan 等[2] 提出 TimedTextRank 算法,一个加入时间因子的 PageRank 算法变种来抽取重要的和新颖的句子形成更新摘要. 它的优点是能同时处理新颖性和查询相关性,但无法实现摘要的信息多样性(需要额外步骤来实现). Boudin 等[1] 使用具有扩展性的最大边缘相关(Maximal Marginal Relevance, MMR)方法实现更新摘要. 该方法是对 MMR 方法的一种改造,MMR[3] 方法通过减小待选句子与已选句子的相似度,扩大待选句子和话题的相似度的方法,来同时面向话题多文档摘要的两个目标. 该方法是目前能够同时处理更新摘要的几个目标(新颖性,多样性,相关性)的方法之一. Li 等[5] 认为文档集合内部的句子之间有相互加强的影响,而不同文档集合间的句子具有相互削弱的影响,则可通过这种相互影响来表达句子的新颖性,因此提出正负增强排序(Ranking with Positive and Negative Reinforcement, PNR^2)的方法,同样,该模型没有特别关注摘要中的多样性问题. Steinberger 等[4] 使用迭代余量调整法(Iterative Residual Rescaling, IRR)方法来获得更准确的新颖的话题分布,来实现更新摘要. 如前所述,目前大多数方法[1-5] 都不能同时应对更新摘要的 3 个问题,常常需要额外的启发性的步骤来应对多样性问题或新颖性问题.

2.2　热传导模型的应用及其特点

热传导模型在信息处理领域主要被用于 Web 上的信息推荐,特别是基于评价打分的信息推荐. 比如针对科研工作者的文献推荐及专家推荐[18-19],影视音乐推荐[20] 及社会标签推荐[19] 等.

这些推荐的共同特点是能对需要推荐的人或物构建一个关联网络,并已知道用户对其中部分人或物的喜好情况,需要推测用户对其他人或物的喜好情况. 应用热传导作为推荐模型的代表性工作是由 Zhang 等[20] 在 PRL 上首先发表的. 该工作以基于用户打分的电影推荐为例,介绍热传导模型的工作原理,着重解决网络上热传导过程的计算问题. 在电影网络中,每部电影都是网络中的一个节点,电影之间通过相关系数记录电影之间的关系. 为了预测用户对某部电影的打分,则需要设置边界条件(Dirichlet BC,即源点和汇点),该用户已打分的电影初始打分置为固定值 1(源点温度),某些未被该用户打分电影初始打分置为 0(汇点温度). 然后,在电影网络中执行热传导过程直到收敛,收敛时边界点之外的其它点的温度值,即为该用户对该电影的可能打分,从而得到向该用户推荐该电影的依据. 如前所述,热传导模型目前主要用于推荐系统中,本文尝试将热传导模型引入到自动文摘领域,即热传导模型在更新摘要中的应用.

基于热传导网络的排序模型应用的基本前提是能将应用中的部分对象(人或物)对应到热传导过程中的热源(源点)和热汇(汇点)上,并且保证热传导网络是连通的. 只有源点或汇点的网络上的热传导过程,最终将会收敛到平凡解,即全局等温. 这样的过程对于排序而言是没有意义的,即平凡的. 另外,如果热传导网络是不连通的,则热传导过程也会出现平凡解的情况,比如源点和汇点分别位于不同的连通分量中,则收敛时,网络中的温度只有两个,最高温和最低温,如果有连通分量中既不包含源点也不包含汇点,则该分量中所有节点的温度相同,对于节点排序而言也是平凡的. 因此,必须有源有汇,网络连通,热传导过程收敛时的温度分布才有排序意义,热传导模型才能适用.

对于更新摘要而言,新颖性和多样性是两个基于比较产生的性质. 通过句子之间的比较,可知道摘要中哪些句子包含的内容相对较新,哪些句子包含的内容与已选出的内容不同. 这些新颖且多样的内容才是我们希望囊括进摘要的. 相反,那些信息陈旧或跟已选摘要句表达重复内容的句子,则是需要被排除在摘要外的. 那些需要得到的句子,要距离源点(查询句)尽可能的近,而那些希望被排除的句子,则需要距离汇点(旧信息句,已选摘要句)尽可能的近. 也就是说,我们能从更新摘要任务中,自然获得基于热传导模型的方法所需要的源点和汇点,因此受 Zhang 等推荐模型的启发,通过迭代选取摘要句和设置汇点的过程,成功地将热传导模型应用在更新摘要任务上.

3 基于热传导模型的更新摘要方法

3.1 方法基本思想

本文提出一种基于热传导(扩散)过程的更新摘要方法——HeatSum. 通过迭代抽取摘要句的方法,将更新摘要的 3 个问题统一在 1 个模型里解决. 在基于网络的热传导模型[20]里,通过设置(狄利克雷)边界条件,并在网络上执行热传导过程直到平衡,非边界点上的温度就能反映出他们跟边界点之间的关系. 当达到热平衡时,或者说热传播过程收敛时,跟源点接近的节点就会有较高的温度,跟汇点接近的节点就会保持一个较低的温度. 因此,可通过将不同的节点置为边界点的方法来获得对其它节点的激励或惩罚的作用.

具体地讲,我们只需要将话题伪句置为热传导中的源点,把旧文档集合中的句子和已选出的摘要句子都置为汇点,就能通过统一的热传导过程,同时捕获更新摘要的话题相关性、内容多样性和信息新颖性的 3 个要求.

3.2 热传导模型

用 $\boldsymbol{\chi}=\{x_0,x_1,\cdots,x_n\}\in\mathbf{R}^m$ 表示句子网络上的节点,每个节点都是 m 维词空间上基于句子的 $tf*isf$ 构成的词向量,其中,tf 为句子内部词频;isf 为逆句子频率;x_0 是是一个伪句子,用来表达用户兴趣的话题描述句,将被设置为源点;$x_1,x_2,\cdots,x_s$ 表示网络上汇点的集合,包括来自历史文档的句子和新文档中已被选为摘要的句子;$x_{s+1},x_{s+2},\cdots,x_n$ 表示其它需要通过热传导过程确定与边界点关系强弱的句子. 任意两点间的连边强度通过节点间的标准余弦相似度确定,即邻接矩阵 $\boldsymbol{A}$ 中的元素 $a_{ij}=sim(x_i,x_j)$ 且 $a_{ii}=0$ 避免自环,其中 $i,j\in\{0,1,\cdots,n\}$.

这样就在可邻接矩阵 $\boldsymbol{A}$ 上构造一个传播矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{A}$,其中,$\boldsymbol{D}$ 是一个对角矩阵,对角元素

$$d_{ii}=\sum_{j=0}^{n}a_{i,j},$$

即 $\boldsymbol{A}$ 中行元素之和. 用 $H:\boldsymbol{\chi}\to\mathbf{R}$ 表示平衡时的温度函数,它为网络中的每个节点 x_i 赋予一个温度值 h_i. 可把 $\boldsymbol{H}$ 看作一个向量

$$\boldsymbol{H}=(h_0,h_1,\cdots,h_s,h_{s+1},\cdots,h_n)^{\mathrm{T}},$$

其中,源点是恒定高温点,令 $h_0=1$,汇点是恒定低温点,令 $h_i=0,i\in\{1,2,\cdots,s\}$. 我们的任务就是通过热平衡方程,来确定其他非边界点的温度. 连续拉普拉斯算子 —— ∇^2 对应的离散拉普拉斯算子在该传播网络上的定义为 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{P}$,其中 $\boldsymbol{I}$ 代表单位矩阵. 这样我们需要解决的方程形式[20]:

$$\boldsymbol{LH}=\boldsymbol{F},\tag{1}$$

其中 $\boldsymbol{F}$ 是一个外部热通量.

因为拉普拉斯算子保持热量守恒,并且从高温区域向低温区域传送热量,所以保持源点和汇点温度恒定的唯一办法就是使用外部热流,从源点流入,汇点流出. 对于其它节点,平衡条件要求它们不会有净热通量,因此 $\boldsymbol{F}$ 中除边界点外的节点对应的元素均为 0.

为了计算温度向量 $\boldsymbol{H}$,我们首先将源点和汇点对应的元素聚集成向量的一个分块 $\boldsymbol{H}_1$,其它元素构成向量的另一分块 $\boldsymbol{H}_2$,类似地,将拉普拉斯矩阵 $\boldsymbol{L}$ 也进行对应的分块操作,式(1) 则可表示为如下形式:

$$\begin{pmatrix}\boldsymbol{L}_{11}&\boldsymbol{L}_{12}\\\boldsymbol{L}_{21}&\boldsymbol{L}_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol{H}_1\\\boldsymbol{H}_2\end{pmatrix}=\begin{pmatrix}f\\0\end{pmatrix}.\tag{2}$$

因为只关心 $\boldsymbol{H}_2$ 中元素的排序,因此只需要计算:

$$\boldsymbol{L}_{21}\boldsymbol{H}_1+\boldsymbol{L}_{22}\boldsymbol{H}_2=0,\tag{3}$$

而不需要知道具体的外部热通 f. 但拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{P}$ 是不可逆的,因为它有一个 0 特征根,对应的左右特征向量分别为

$$\langle\boldsymbol{v}^0|=(\frac{d_1}{d},\frac{d_2}{d},\cdots,\frac{d_n}{d}),\ |u^0\rangle=(1,1,\cdots,1)^{\mathrm{T}},$$

其中 d_i 对应对角矩阵 $\boldsymbol{D}$ 中的对角线元素,d 为 $\boldsymbol{D}$ 所有对角元素之和. 因此式(3) 不一定能直接求得 $\boldsymbol{H}_2$,除非 $\boldsymbol{H}$ 正好位于 $\boldsymbol{L}$ 的与 0 特征根对应的特征向量正交的子空间中. 幸运地是,根据 Zhang 等[20] 的工作,以及随机电阻网络中类似工作[21-22],结合式(3),可得到 $\boldsymbol{H}_2$ 的形式如下:

$$|H_2\rangle=\frac{(\langle\boldsymbol{v}_1^0|\boldsymbol{H}_1\rangle+\langle\boldsymbol{v}_2^0|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}|\boldsymbol{H}_1\rangle)(|\boldsymbol{u}_2^0\rangle-\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}|\boldsymbol{u}_1^0\rangle)}{1-\langle\boldsymbol{v}_2^0|\boldsymbol{u}_2^0\rangle+\langle\boldsymbol{v}_2^0|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}|\boldsymbol{u}_1^0\rangle}+\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}|\boldsymbol{H}_1\rangle,\tag{4}$$

其中

$$\boldsymbol{S}=(\boldsymbol{I}-\boldsymbol{Q})^{-1},\ \boldsymbol{Q}=\boldsymbol{P}-|\boldsymbol{u}^0\rangle\langle\boldsymbol{v}^0|,$$

记 $\boldsymbol{S}_{11},\boldsymbol{S}_{12},\boldsymbol{S}_{21},\boldsymbol{S}_{22}$ 分别为 $\boldsymbol{S}$ 矩阵中根据边界点和非边界点划分出的分块矩阵,$\langle\boldsymbol{v}_1^0|,\langle\boldsymbol{v}_2^0|$ 为 $\langle\boldsymbol{v}_0|$ 中对应边界点和非边界点的分块,类似的,$|\boldsymbol{u}_1^0\rangle,|\boldsymbol{u}_2^0\rangle$ 为向量 $|\boldsymbol{u}_0\rangle$ 的分块.

可看到,$|\boldsymbol{H}_2\rangle$ 计算复杂度主要取决于 $\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}$. 当需要设置一个新的源点或汇点时,只需要对 $\boldsymbol{S}$ 中

对应的行和列和 $\boldsymbol{H}$ 中对应元素进行位置调整,并重新计算 $\boldsymbol{S}_{11}^{-1}$ 即可.

基于上述热传导模型,最终的 HeatSum 算法如下所示.

step 1　对于任意两个句子 $x_i, x_j \in \boldsymbol{\chi}$,计算其余弦相似度 $sim(x_i, x_j)$,包括新文档集合 $\boldsymbol{\chi}_n$ 中的句子、旧文档集合 χ_o 中的句子以及话题描述句 x_t.

step 2　对于句子 x_i 和 x_j,如果有 $sim(x_i, x_j) > 0$,则向网络中添加边 (x_i, x_j),从而得到网络的邻接矩阵 $\boldsymbol{A}$,其中元素 $a_{ij} = sim(x_i, x_j)$,令 $a_{ii} = 0$ 以避免自环.

step 3　初始化边界条件,即句子节点的温度向量 $\boldsymbol{H}$,将话题描述 x_t 设置为源点,旧文档集合的句子初始化为汇点. 即 $\boldsymbol{H}(x_t) = 1, \boldsymbol{H}(x_o) = 0$,其中, $x_o \in \chi_o$. 其他节点 $\boldsymbol{H}(x_n)$ 为待计算变量.

step 4　将温度向量 $\boldsymbol{H}$ 按照边界点和非边界点进行分块,即 $\boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2)^{\mathrm{T}}$,其中,$\boldsymbol{H}_1$ 代表边界点对应的温度常量,$\boldsymbol{H}_2$ 代表非边界点对应的温度变量. 类似地,将邻接矩阵分块,得

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}.$$

step 5　将 $\boldsymbol{A}$ 按行规范化得到矩阵 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$,其中

$$d_{ii} = \sum_{j=0}^{n} a_{i,j}.$$

step 6　根据

$$\boldsymbol{S} = (\boldsymbol{I} - \boldsymbol{Q})^{-1} = (\boldsymbol{I} - \boldsymbol{P} + |\boldsymbol{v}^0\rangle\langle\boldsymbol{u}^0|)^{-1}$$

计算 $\boldsymbol{S}$,其中

$$\langle \boldsymbol{v}^0| = (\frac{d_1}{d}, \frac{d_2}{d}, \cdots, \frac{d_n}{d}), |\boldsymbol{u}^0\rangle = (1, 1, \cdots, 1)^{\mathrm{T}}.$$

step 7　根据式(4) 计算 $|\boldsymbol{H}_2\rangle$.

step 8　取 $|\boldsymbol{H}_2\rangle$ 中的最大元素对应的句子作为摘要句子,并将该元素赋值为零,移动到 $|\boldsymbol{H}_1\rangle$ 中,修改 $\boldsymbol{S}$ 矩阵对应的行、列以及分块.

step 9　判断摘要是否达到预定的摘要长度,如果没有,则转到 step 7 继续执行.

基于热传导的更新摘要算法,通过构建句子间的相似关系网络,并将查询句子置为源点,历史句子和已选句子置为汇点,在网络上进行热传导直到收敛的过程,来获得各个句子的温度打分作为句子的排序值. 该排序能同时反映出句子与查询的紧密程度,与历史句子的不相似程度,以及与已选摘要句的差异. 从而同时捕获更新摘要的 3 个基本目标. 配合一个迭代的句子选取框架,HeatSum 能在一个统一的框架内获得更新摘要在 3 个目标的综合性能上的较好的表现.

4　实验及评价结果

4.1　数据集

更新摘要是由 NIST(http://www.nist.gov) 组织的 TAC(http://www.nist.gov/tac/2009/Summarization) 评测会议 3 个主要任务之一. 他们为更新摘要提供标准的评测数据集. 在 TAC09 上实验验证本文方法的性能. NIST 为 TAC09 的更新摘要任务共提供 44 个话题,每个话题都由 AQUAINT-2 的新闻文档集中抽取的 20 篇文档组成. 数据的详细统计信息如下:话题个数为 44,文档数量为 880,单篇文档的平均句子个数为 22.8,单个句子的平均词汇个数为 23.1,数据来源为 AQUAINT-2,摘要长度为 100. 这 20 篇文档被分为 2 个集合 A 和 B,每个集合包含 10 篇文档,集合 A 中的文档在时间上都要早于集合 B 中的文档.

TAC 的更新摘要任务要求为每个集合都提供一个 100 字的摘要. 其中对集合 B 的摘要是在假设用户已读过集合 A 中所有文档的前提下完成,因此,应提供给用户出现在 B 中而没有出现在 A 中的话题相关的新信息. 本文只关心更新摘要的性能,因此,没有给出集合 A 的摘要.

4.2　评价方法

ROUGE(http://haydn.isi.edu/ROUGE) 已成为最常用的自动文摘评价工具之一[23],使用 ROUGE 评测的得分与人工评价得分具有较好的相关性. ROUGE 通过计数人工摘要和自动摘要中共同出现的词序列、词对或 N 元词串来定量地衡量自动摘要的好坏. ROUGE-N 是基于 N 元词串的召回率的评价指标,计算方法如下:

$$ROUGE\text{-}N = \frac{\sum_{S \in \{Refs\}} \sum_{gram_n \in S} Cnt_{match}(gram_n)}{\sum_{S \in \{Refs\}} \sum_{gram_n \in S} Cnt(gram_n)},$$

其中,n 代表 N 元词串的长度,$Cnt_{match}(gram_n)$ 是人工文摘(参考文摘) 和自动文摘中共同出现的 N-元词串的最大个数. $Cnt(gram_n)$ 是人工文摘中出现的 N-元词串的总个数.

在评价中,使用 ROUGE-1、ROUGE-2 和 ROUGE-SU4 (ROUGE-2 的扩展版本)3 个自动评价指标进行评价. 它们是 TAC 评测会议最常使用的自动评价指标. 评价结果由 ROUGE 工具 1.5.5 版,使用 TAC09 评测会议使用的参数设置得到.

4.3 基准方法介绍

为了评价本文方法 HeatSum 的有效性,把该方法与 4 个基准方法进行对比.

第 1 个基准是参加 TAC09 评测的抽取式更新摘要方法中在 ROUGE-2 指标上取得最好成绩的方法,即 S24. 该系统是由 Long 等[24]基于条件信息距离理论构造的更新摘要框架. 该方法认为在给定历史文档集合的条件下,最好的更新摘要应具有同新文档集合最短的条件信息距离. 然而构成候选摘要的句子集合是原始文档集合中句子集合的子集,因此摘要子集个数是指数级别的,因此计算代价是个值得考虑的问题.

第 2 个基准方法是代表性的方法——可扩展的最大边缘相关法(Scalable Maximal Marginal Relevance, SMMR). 该方法是基于 MMR 方法改造而成的方法. MMR 方法通过将相关性(句子与查询之间的相似度)和多样性(句子同已选句子之间的不相似性)的线性组合,构造出一个能同时度量句子相关性和多样性的测量标准. 而 SMMR 通过向 MMR 引入时间衰减因子,给时间戳较新的句子的 MMR 值修正成一个较高得分,从而体现句子的新颖性.

在目前已知的更新摘要方法中,SMMR[1]是试图在统一的框架下解决更新摘要的 3 个问题的代表性方法之一,它使用一个能同时反映话题相关性和内容新颖性的度量指标,结合一个能解决信息多样性的迭代过程来完成更新摘要的 3 个目标. 因此我们实现 SMMR 方法作为另外一个基准方法. 但容易看出,SMMR 方法带有明显的启发性特征,理论依据不够充分.

另外两个基准分别是 BaseLine-L、BaseLine-U. BaseLine-L 和 BaseLine-U 是由 TAC09 提供的两个基准. BaseLine-L 抽取最新文档中的最前面的句子构成 100 字的摘要. 它提供一个抽取式摘要所能达到的性能的下界[25]. BaseLine-U 是由来自蒙特利尔大学的 5 人摘要小组从文档集中抽取句子构成摘要. 它为抽取式的摘要提供一个近似的上界.

4.4 评价结果

在 TAC09 数据集上的性能比较结果如表 1 所示.

表 1 中系统 S24 代表参评 TAC09 的实际抽取式方法中表现最好的方法. 该方法借助条件信息距离的方法实现. 它假设在给定历史文档集合的条件下,最好的更新摘要应具有同新文档集合最短的条件信息距离. 这个假设导致该方法的摘要句子集的搜索空间是原始文档集合中的句子集合的所有子集. 虽然摘要有长度限制,这个搜索空间的指数级的规模也需要相当大的计算代价. HeatSum 的计算代价为 $O(KN^2)$,其中,K 为摘要句的个数,通常为 3 ~ 5;N 为文档集合中的所有句子的个数,在摘要句较少的情况下,数量级接近为文档句子总数的平方,与原始矩阵求逆运算代价相比,降低一个数量级. 从表 1 可看到,本文方法与 S24 有非常接近的性能表现. 另外,该方法借助很多自然语言处理的技巧来提高摘要的质量,因此它不能算作是纯粹的抽取式摘要. 通常采用自然语言处理的技巧来对抽取的句子进行后处理,能较大提高自动文摘的质量. 因此,同 S24 相比 HeatSum 方法的性能还有很大的提升空间.

表 1 TAC09 数据集上的性能对比

Table 1 Performance comparison on dataset TAC09

方法	ROUGE-1	ROUGE-2	ROUGE-SU4
BaseLine-U	0.37068	0.09820	0.13631
S24	0.36758	0.09615	0.13520
HeatSum	0.36761	0.09455	0.13329
SMMR	0.35784	0.08931	0.12945
BaseLine-L	0.28613	0.05142	0.09091

SMMR 方法是试图在一个统一的框架下解决更新摘要的 3 个目标的代表性方法. 从表 1 可看出,HeatSum 的性能表现在各个评价指标上都一致地、显著地优于 SMMR. SMMR 方法通过查询相似性,历史句不相似性的线性组合来获取相关性和信息多样性,又通过句子的时间戳引入一个得分上的时间衰减因子来矫正得分,获得信息新颖性.

不难看出,SMMR 方法具有明显的启发性特征,相比之下,基于热传导模型的更新摘要具有更强的理论基础. 但 SMMR 方法的实现方法简单,计算效率较高,不失为实际应用的有效方法之一. 我们将 HeatSum 和 SMMR 在 44 个话题上进行详细对比,对比结果如图 1 所示,图 1 中,x 坐标是 44 个话题序号,y 坐标是 HeatSum 方法和 SMMR 方法在 ROUGE-1 得分上的差值. 两种方法在 ROUGE-1 得分上的差值大部分落在点划线(0 水平线)上方. 对两种方法在 44 个话题上的 ROUGE-1 得分的 *T-test* 结果也表明本文方法(p-*value* = 0.05)水平上优于 SMMR. 与 SMMR 相比,HeatSum 通过充分利用句子之间,句子与查询之间,新句子与旧句子之间的局部关系,通过热传导过程来揭示句子之间在整体上的特定关系——新颖性,多样性和相关性. 尽管 HeatSum 的计算方法前人已进行优化,但在大规模的网络应用中,计算仍需要更加深入研究.

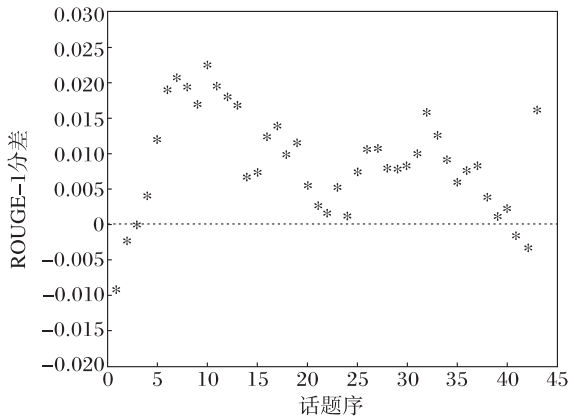

图 1　方法 HeatSum 和 SMMR 在数据集 TAC09 上的对比

Fig. 1　Comparison of HeatSum and SMMR on dataset TAC09

表 1 中的另外两个基准分别是 BaseLine-U 和 BaseLine-L. BaseLine-U 是 TAC09 提供的由人工抽取句子构成的摘要,该方法所能达到的摘要性能被认为是抽取式摘要方法所能达到的近似性能上界. BaseLine-L 也是由 TAC09 提供,它从最新的文档中截取前 100 词构成抽取式摘要. 该方法过于简单,但鉴于经验表明,摘要句是首句的可能性较大,该方法获得的摘要不一定是最差的抽取式摘要. 该方法的性能表现被认为是抽取式摘要的性能下界. 从表 1 可看出,HeatSum 显著地优于 SMMR 方法和 TAC09 提供的性能下界 BaseLine-L,但比性能上界 BaseLine-U 的表现要差. 同时跟参评系统中表现最好的抽取式方法 S24 相比,具有相当的性能表现. 鉴于 HeatSum 没有采用复杂的自然语言处理技巧进行摘要的后处理,有理由相信,HeatSum 的性能还有很大的提升空间.

5　结 束 语

本文提出一种基于热传导模型的更新摘要方法. 通过热传导模型中边界条件的设定,自然利用热传导机制,对话题相关的句子进行激励,对陈旧的信息进行惩罚. 通过将已选出的句子作为新的恒定低温的汇点,使得与已选摘要句较为相似的候选摘要句也得到惩罚,从而能通过一个统一的迭代选句的过程,同时解决更新摘要的 3 个问题——话题相关性、内容多样性、内容新颖性. 在 NIST 提供的TAC09 关于更新摘要的数据集上的实验表明,本文方法能取得与最好的系统相当的表现,同时接近抽取式摘要方法所能达到的性能上限,该方法显著优于 SMMR 及其它基准方法.

理论和实验表明,热传导模型不仅适用于各种各样的推荐系统,也能较好处理自动文摘这一同时需要满足话题相关性、内容多样性和新颖性的综合性排序目标的网络应用. 未来我们将尝试将该方法应用在其它需要满足多个目标的排序应用中,比如搜索引擎的查询推荐、商品评论综述、微博综述等网络应用中.

参　考　文　献

[1] Boudin F, El-Bμèze M, Torres-Moreno J M. A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization // Proc of the 22nd International Conference on Computing Linguistics. Manchester, UK, 2008: 23-26

[2] Wan Xiaojun. Timedtextrank: Adding the Temporal Dimension to Multi-Document Summarization // Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, Netherlands, 2007: 867-868

[3] Zhang Jin, Cheng Xueqi, Xu Hongbo, *et al*. ICTCAS's ICTGrasper at TAC 2008: Summarizing Dynamic Information with Signature Terms Based Content Filtering [EB/OL]. [2010-08-13]. http://www.nist.gov/tac/publications/2008/participant.papers/ICTCAS.proceedings.pdf

[4] Steinberger J, Ježek K. Update Summarization Based on Novel Topic Distribution // Proc of the 9th ACM Symposium on Document Engineering. Munich, Germany, 2009: 205-213

[5] Li Wenjie, Wei Furu, Lu Qin, *et al*. PNR 2: Ranking Sentences with Positive and Negative Reinforcement for Query-Oriented Update Summarization // Proc of the 22nd International Conference on Computational Linguistics. Manchester, UK, 2008: 489-496

[6] Erkan G, Radev D R. LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 2004, 22(1): 457-479

[7] Lin C Y, Hovy E. Manual and Automatic Evaluation of Summaries // Proc of the ACL-02 Workshop on Automatic Summarization. Morristown, USA, 2002: 45-51

[8] Mihalcea R, Tarau P. Textrank: Bringing Order into Texts // Proc of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 404-411

[9] Otterbacher J, Erkan G, Radev D. Using Random Walks for Question-Focused Sentence Retrieval // Proc of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada, 2005: 915-922

[10] Saggion H, Bontcheva K, Cunningham H. Robust Generic and Query-Based Summarization // Proc of the 10th Conference on European Chapter of the Association for Computational Linguistics. Stroudsburg, USA, 2003: 235-238

[11] Wan Xiaojun, Yang Jianwu, Xiao Jianguo. Manifold-Ranking Based Topic-Focused Multi-Document Summarization // Proc of the 20th International Joint Conference on Artificial Intelligence.

Hyderabad, India, 2007: 2903-2908

[12] Wei Furu, Li Wenjie, Lu Qin, *et al*. Query-Sensitive Mutual Reinforcement Chain and Its Application in Query-Oriented Multi-Document Summarization // Proc of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, Singapore, 2008: 283-290

[13] Zhang Jin, Cheng Xueqi, Wu Gaowei, *et al*. Adasum: An Adaptive Model for Summarization // Proc of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, USA, 2008: 901-910

[14] Conroy J M, Schlesinger J D. Classy Query-Based Multidocument Summarization [EB/OL]. [2010-08-13]. http://www.nlpic.nist.gov/tac/projects/duc/pubs/2005papers/ida.conroy.pdf

[15] Hovy E, Lin C Y, Zhou L, *et al*. Automated Summarization Evaluation with Basic Elements // Proc of the 5th Conference on Language Resources and Evaluation. Genvoa, Italy, 2006: 899-902

[16] Allan J, Gupta R, Khandelwal V. Temporal Summaries of New Topics // Proc of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, USA, 2001: 10-18

[17] Carbonell J, Goldstein J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries // Proc of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 335-336

[18] Zhou Tao, Kuscsik Z, Liu Jianguo, *et al*. Solving the Apparent Diversity-Accuracy Dilemma of Recommender Systems. Proc of the National Academy of Sciences of the United States of America, 2010, 107(10): 4511-4515

[19] Zhang Zike, Zhou Tao, Zhang Yicheng. Personalized Recommendation via Integrated Diffusion on User-Item-Tag Tripartite Graphs. Physica A: Statistical Mechanics and Its Applications, 2010, 389(1): 179-186

[20] Zhang Y C, Blattner M, Yu Y K. Heat Conduction Process on Community Networks as a Recommendation Model. Physical Review Letters, 2007, 99(15): 154301

[21] Korniss G, Hastings M, Bassler K, *et al*. Scaling in Small-World Resistor Networks. Physics Letters A, 2006, 35(5/6): 324-330

[22] Wu F Y. Theory of Resistor Networks: The Two-Point Resistance. Journal of Physics A: Mathematical and General, 2004, 37(26): 6653-6673

[23] Lin C Y. Rouge: A Package for Automatic Evaluation of Summaries // Proc of the ACL-04 Workshop on Text Meaning and Interpretation. Barcelona, Spain, 2004: 74-81

[24] Long Chong, Huang Minlie, Zhu Xiaoyan. Tsinghua University at TAC 2009: Summarizing Multi-Documents by Information Distance [EB/OL]. [2010-08-13]. http://www.nist.gov/tac/publications/2009/participant.paper/THVSVM.proceedings.pdf

[25] Dang H T, Owczarzak K. Overview of the Tac 2009 Summarization Track (draft) [EB/OL]. [2010-08-13]. http://www.nist.gov/publications/2009/presentations/TAC2009_Sum_overview.pdf