基于事件要素加权的新闻摘要提取方法

郭艳卿[1,2] 赵 锐[1] 孔祥维[1] 付海燕[1] 蒋金平[1]
(大连理工大学信息与通信工程学院 大连 116024)[1] (国家信息中心博士后科研工作站 北京 100045)[2]

摘 要 为帮助读者从海量新闻报道中快速了解某一事件的来龙去脉,分析了新闻事件中事件要素对生成摘要的影响,结合新闻事件演变式发展的特点,提出了一种基于事件要素加权的新闻摘要提取方法。通过对事件要素的加权,对转移概率矩阵进行改进,有效地按时间顺序提取出摘要信息,使得最后生成的摘要包含更多的新闻要素细节信息,增加了输出时间轴摘要的细节性和可读性。实验结果证明了所提算法的有效性。

关键词 新闻事件,时间轴摘要,转移概率矩阵,要素加权
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.051

News-summarization Extraction Method Based on Weighted Event Elements Strategy

GUO Yan-qing[1,2] ZHAO Rui[1] KONG Xiang-wei[1] FU Hai-yan[1] JIANG Jin-ping[1]
(Institute of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China)[1]
(Post-doctoral Scientific Research Station, State Information Center, Beijing 100045, China)[2]

Abstract To facilitate the readers' fast understanding of the contexts of news events, this paper analyzed the effect of event elements on the summarization generation, and by combining the character of news evolution along the timeline proposed a news-summarization extraction method based on weighted event elements strategy. The transition probability matrix is improved by weighting the event elements, which turns out to be effective in extracting the news timeline summarization. By this way, the generated summarization contains more details of the news elements, whilst becomes more readable to readers. Experimental results demonstrate the superiority of the proposed algorithm.

Keywords News event, Timeline summarization, Transfer probability matrix, Element weighting

1 引言

随着计算机技术和互联网的不断发展,网络新闻媒体已逐步成为人们获取新闻信息的主要途径。然而新闻网站对突发新闻事件产生的海量报道,使得读者难以迅速了解整个事件的来龙去脉。因此自动生成新闻事件的摘要来帮助读者阅读十分必要。

近年来,多文档摘要技术飞速发展,一定程度上解决了上述问题。但传统的多文档摘要旨在产生一个总结来表示一组文档中的主要信息,仅仅生成每个时间点的摘要,没有考虑各时间点间的相关性,这显然是不够的[1-9]。而时间轴摘要的优势在于它很好地利用了时间信息,使得呈现在读者面前的摘要更加一目了然,这对于针对网络新闻事件进行摘要这个任务来说非常重要。相较于传统的多文档摘要方法,时间轴摘要方法起步较晚。Swan等人[10]提出通过名词短语提取和命名实体识别技术,在每个时间节点上给出了多个关键词来概括主题内容。之后,他们构建了一个时间轴摘要系统[11],这个系统综合考虑了句子的有用性和新颖性。Chieu等人[12]在对句子进行抽取时抽取被相关文本引用最多的句子,考虑的是读者感兴趣程度,针对每个重要时间节点,从文本集中抽取出一个句子来组成时间轴摘要。但以上方法都没有涉及新闻事件的演变特点。Yan等人[13]提出演变时间轴摘要方法,但该方法精炼摘要的过程非常耗时。随后他们又提出对全局和局部新闻集分别提取摘要,然后通过相应的算法将两部分融合成最终的时间轴摘要[14]。该方法从海量数据中生成新闻主题的时间轨迹,帮助读者快速浏览和了解事件,但没有考虑新闻事件中各事件要素的重要性。

通过分析新闻报道本身具有的事件要素信息,发现在摘要生成过程中,新闻报道的主体是事件,新闻事件中的时间、地点、人物等要素是人们最为关注的信息,而在新闻事件的演变发展过程中,其要素在事件演变过程中的参与度是不断变化的,突出各要素的权重会提高所生成时间轴摘要的细节性和可读性。因此本文针对新闻事件中各要素的重要性,并结合新闻报道在时间维度上为演变式发展的特点,提出了基于事件要素加权的新闻摘要提取算法。本文第2节给出时间轴摘要模型;第3节描述提出的算法;第4节给出实验结果与分析;最后总结全文。

到稿日期:2014-12-26 返修日期:2015-04-21 本文受中国博士后科学基金(20110490343,2013T60090)资助。
郭艳卿(1980—),男,副教授,主要研究方向为多媒体信息处理、机器学习与数据挖掘,E-mail:guoyq@dlut.edu.cn;**赵 锐**(1989—),男,硕士,主要研究方向为多媒体信息处理;**孔祥维**(1963—),女,教授,主要研究方向为多媒体信息处理与安全;**付海燕**(1981—),女,工程师,主要研究方向为图像处理;**蒋金平**(1986—),男,硕士,主要研究方向为多媒体信息处理。

2 时间轴摘要模型

时间轴摘要是在事件几个重要的时间节点上输出可以代表该节点事件发展情况的句子集，这些句子可以较全面地概括该时间节点网络上针对该事件新闻报道的主要内容。时间轴摘要过程包括重要事件节点提取、特征句子提取、摘要句子输出3个主要部分。而对于网络上的新闻报道，其在时间维度上是演变式的发展过程，将带有时间标签的新闻文本集作为输入，得到的输出为一个时间轴摘要。

具体的时间轴摘要流程如图1所示。摘要针对的对象即文本集合1、文本集合2到文本集合 n 构成的文本总集合。摘要生成过程由局部摘要和全局摘要融合而成，局部和全局摘要均要经过两个过程：转移概率矩阵构建和多样化排序句子抽取，转移概率矩阵构建即以句子为单位，经过分词和特征提取，计算句子之间相似度的过程；多样化排序句子抽取即根据转移概率矩阵中的句子之间的关系通过排序算法进行句子选择的过程。局部和全局摘要分别计算完成后，通过两个摘要中句子的权重值融合输出最终的摘要。

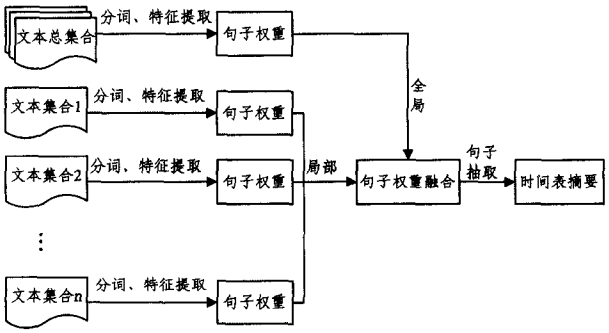

图1 基于时间演变的时间轴摘要流程

3 基于事件要素加权的新闻摘要提取方法

传统时间轴摘要方法在计算句子局部摘要转移矩阵时，使用词频统计及句子间相似性来确定矩阵的元素，没有利用新闻事件要素的重要性，导致摘要效果不够理想。针对这一问题，在文献[14]的算法的基础上，提出了针对中文的基于要素加权的新闻摘要提取方法，对文献[14]的算法进行了两点改进：(1)对局部摘要转移概率矩阵对角线上的元素进行赋值操作(2)对局部摘要转移概率矩阵中的非对角线元素进行加权处理。下面对提出的基于事件要素提取的新闻摘要提取算法进行介绍。

3.1 分词与新闻事件要素提取

本文提出的新闻摘要提取方法是基于新闻事件要素加权策略的。首先ICTCLAS分词程序可以有效地对新闻报道中的时间、地点、人物、机构团体等命名实体进行标记，如“/t”、“/nr”、“/ns”和“/nt”分别代表时间、地点、人物、机构团体4种新闻要素。经过分词后的结果包含一些无实际语义的冠词、代词、连词、助词等，这些词叫做停用词，在进行特征提取等进一步操作之前，需要将这些停用词去除；然后对去除停用词的分词结果进行进一步提取，可以得到某篇新闻报道的4个要素词组：时间要素词组 $T=\{t_1,t_2,\cdots,t_n\}$，地点要素词组 $L=\{l_1,l_2,\cdots,l_n\}$，人物要素词组 $P=\{p_1,p_2,\cdots,p_n\}$，机构团体要素数组 $NT=\{nt_1,nt_2,\cdots,nt_n\}$。4个要素词组可以统称为新闻要素词组，记为 $Pro=T\cup L\cup P\cup NT$。

3.2 转移概率矩阵要素加权

在进行转移概率矩阵要素加权之前，需要对第3.1节中得到的时间、地点、人物、团体机构名等4个新闻要素词组进行权值量化。量化过程即计算各个新闻要素权值的过程。以时间数组为例，各个时间要素的权值量化过程见式(1)。

$$W_T(t_i)=\frac{tf(t_i,T)}{|T|} \tag{1}$$

式中，$W_T(t_i)$代表时间要素 t_i 在时间数组 T 中的权重值，$tf(t_i,T)$代表 t_i 在 T 中出现的次数，$|T|$代表 T 中时间要素的总个数。

对地点、人物、机构团体要素数组进行与时间要素数组同样的权重量化过程，得到了4个要素数组的特征值，分别为 W_T、W_L、W_P 和 W_{NT}。将4个要素特征值数组用一个总的要素特征值数组表示，记为 W_{Pro}。

本文的转移概率矩阵要素加权方法是在文献[14]中转移概率矩阵构建过程的基础上进行的改进，下面对转移概率矩阵构建的各个步骤进行详细介绍。

3.2.1 句子中词的权重计算

局部摘要中词在句子中的权重计算公式见式(2)，在经典的TFIDF方法的基础上，对出现在新闻要素词组中的情况进行了相应的加权处理。

$$\pi(w,s|t)=\begin{cases}\dfrac{\delta\cdot(1+W_{Pro}(w))\cdot tf(w,s)(1+\log(\frac{|C|}{N_w}))}{\sqrt{\sum_{|S|}(\delta\cdot(1+W_{Pro}(w))\cdot tf(w,s)(1+\log(\frac{|C|}{N_w})))^2}}, \\ \qquad\qquad \text{当 } w\in Pro \text{ 时} \\ \dfrac{tf(w,s)(1+\log(\frac{|C|}{N_w}))}{\sqrt{\sum_{|S|}(tf(w,s)(1+\log(\frac{|C|}{N_w})))^2}}, \text{当 } w\notin Pro \text{ 时}\end{cases} \tag{2}$$

式中，$tf(w,s)$表示词 w 在句子 s 中出现的次数；$|C|$表示 t 时刻对应文本集中的句子个数；N_w 表示 t 时刻对应文本集中包含词 w 的句子个数；W_{Pro}代表词 w 在新闻要素权重词组中对应的权重；δ代表其对应的权值系数，δ 的调整对应着该句子中新闻要素词所占比重的调整，该系数越大，出现该词的两个句子之间的相似度越趋近于1。

式(2)对应 t 时刻局部摘要中词 w 在句子 s 中的权重，当词 w 不属于要素词时，利用TFIDF方法来求得权重值；当词 w 属于要素词时，对要素词组乘以总的权重系数来求得权重值。

3.2.2 句子之间的相似度加权计算

得出 t 时刻局部摘要中每个句子中所有词的权重值后，每个句子可以表示成由词的权重值构成的特征向量，t 时刻两个句子之间的相似度 $f(s_i\rightarrow s_j|t)$可以通过式(3)来计算，该计算公式是典型的计算两个向量之间相似度的余弦公式，即对两个句子中的词取交集，在这个交集范围内对该词在两个句子中的权重值进行乘法运算。对于两个句子同时包含某要素词的情况，若其权值系数 δ 调整得足够大，则该词对应的TFIDF值会足够大，这样就使得该词对句子之间相似度的影响远远大于非要素词，两个句子之间的相似度会无限接近1。

$$f(s_i\rightarrow s_j|t)=\sum_{w\in s_i\cap s_j}\pi(w,s_i|t)\cdot\pi(w,s_j|t) \tag{3}$$

3.2.3 转移概率矩阵构建

局部转移概率矩阵中对角元素的含义是句子对其本身的贡献值，所提算法将其赋值成该句子包含的新闻要素对应的权重值。局部摘要转移概率矩阵的计算过程见式(4)。

$$p(s_i \to s_j) = \begin{cases} \dfrac{f(s_i \to s_j \mid t)}{\sum\limits_{k \in |C|} f(s_i \to s_j \mid t)}, & \text{当 } i \neq j \text{ 且} \sum f \neq 0 \text{ 时} \\ f(s_i \to s_j \mid t), & \text{当 } i \neq j \text{ 且} \sum f = 0 \text{ 时} \\ 1 - \dfrac{1}{1+\delta \cdot \sum W_{Pro}(w)}, & \text{当 } i = j \text{ 时} \end{cases} \quad (4)$$

式中，δ 为权重系数，同式(2)中权值系数保持一致；$\sum W_{Pro}(w)$ 为对角线对应句子包含要素权重的总和，在局部摘要中，使用 $\sum W_{Pro}(w)$ 作为对角线元素值，$\sum W_{Pro}(w)$ 越大，对角线元素值越接近于1，即该句子对本身的贡献值越大；最终得到局部摘要转移概率矩阵 M_L。

如式(4)所示，矩阵元素对两个句子之间的相似值按行进行了归一化处理，局部摘要转移概率矩阵的计算针对某特定的时间节点文本集，第 i 行 j 列元素表示该文本集中第 i 个句子与第 j 个句子之间的相似值，归一化之后，第 i 行 j 列元素表示该文本集中第 i 个句子对第 j 个句子的贡献值。

3.3 句子权重赋值与抽取

本文提出的新闻摘要提取算法对句子权重向量的赋值参考了文献[14]中提出的 Pagerank[15] 和 Divrank[16] 相结合的方法，分别对局部概率转移概率矩阵 M_L 和由文献[14]获得的全局概率矩阵 M_G 进行迭代求解，得到全局句子权重向量 $\vec{\lambda}_G$ 和局部句子权重向量 $\vec{\lambda}_L$。

迭代过程为：首先利用 Pagerank 原理计算出权重向量 $\vec{\lambda}$，然后利用 Divrank 方法求解转移概率矩阵，重复这两个步骤直到前后两次得出的句子权重向量 $\vec{\lambda}$ 中元素值之间的最大差小于 0.0001，迭代结束。

通过式(5)对两者的权重向量完成融合，此时得出该时间节点处融合后的句子权重向量，根据权重值对权重向量从高到低进行排序，取排名靠前的句子作为摘要句子输出。

$$\vec{\lambda} = \frac{1}{1+\beta/\alpha}\vec{\lambda}_G + \frac{1}{1+\alpha/\beta}\vec{\lambda}_L \quad (5)$$

式中，α/β 代表输出时间表摘要中局部权重向量和全局权重向量所占比例，通过比较实验结果可知，当 $\alpha/\beta=10$ 时，得到的摘要效果最好，因此实验中 α/β 设置为 10。

4 实验结果与分析

本文通过调用 curl 爬虫软件接口的方式对新浪、搜狐、网易、新华网和人民网 5 个新闻网站进行了实验语料库采集，爬虫对象为社会、文化、财经、科技、体育 5 个领域近几年的新闻事件。通过分析比较两种网页去重算法[17,18]，针对新闻报道篇幅较短的特点，最终选择了文献[18]的方法对爬取的网页作了去重处理，共得到约 60000 篇无重复的新闻报道，文件大小共 1GB。将这 5 个领域的事件作为语料进行实验，并从上述 5 个领域中各取出一个新闻事件，共 5 个事件约 2500 篇报道用作结果的展示，新闻事件样本细节如表 1 所列。最后在 Visual Studio 2008 平台上使用 C# 编程语言对上述算法进行了程序实现。

表 1 新闻摘要事件样本

新闻事件	发生时间	新闻报道数	总句子数
温州动车追尾	2011.07.23—2011.12.29	710	17044
莫言获诺贝尔	2012.08.25—2012.12.10	473	14952
三亚游客挨宰	2012.01.30—2012.04.28	323	5081
发布 iPhone5	2012.06.23—2012.09.26	466	10442
阿姆斯特朗事件	2011.01.21—2013.02.23	578	14905

目前实验应用比较普遍的是内部评价方法。内部评价方法一般将摘要系统生成的待评价摘要与“标准摘要”进行比较，待评价摘要越接近“标准摘要”，则说明其质量越高。本文的“标准摘要”由 5 位专家分别提取再整合构成，并且为了与系统生成摘要更方便地进行比较，人工生成的“标准摘要”中句子均为从新闻报道中提取的原句；同时某时间节点选取句子的个数限定在一定范围内，根据时间节点的重要性灵活选取。算法的性能采用内部评价方法中的召回率、准确率和 F 值 3 个指标来评价。其中召回率和准确率分别体现了待评价摘要的覆盖率和正确率；F 值为召回率和准确率两个评价指的综合考虑。如果用 N 表示待测摘要中符合“标准摘要”(人工提取)的句子个数，用 N_p 表示“标准摘要”的句子总数，用 N_r 表示待测摘要的句子总数，则召回率、准确率和 F 值分别为：

$$Recall = \frac{N}{N_p} \quad (6)$$

$$Precision = \frac{N}{N_r} \quad (7)$$

$$F\text{-}Measure = \frac{2 * R * P}{R+P} \quad (8)$$

针对摘要方法设计的实验包括：

(1)式(2)和式(4)中权值系数 δ 的设置；

(2)基于事件要素提取的新闻摘要提取方法与基于时间演变的时间轴摘要方法的实验结果比较。

实验(2)中本文提出的算法生成摘要的过程中，使用的权值系数 δ 为经过实验(1)后确定的最佳值。下面对这两个实验分别进行介绍。

4.1 句子相似度计算中权值系数 δ 的设置实验

式(2)中 δ 的设置关系着要素权重在摘要生成过程中所占的权重，5 个新闻事件采用本文摘要提取方法的 F 值随着 δ 不同设置的变化情况如图 2 所示。

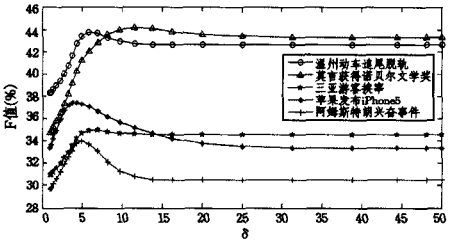

图 2 随参数 δ 调整的 F 值(%)

从实验结果可以看出，5 类事件 F 值都是先随着 δ 的增大而增大，在 $\delta=5$ 附近达到最大值，然后又随着 δ 的增大而减小，最后达到一个稳定值，这个稳定值为该实验的最优解。分析其原因，$\delta=5$ 与 $\delta=1$ 相比，可以更有效提高同时包含要素词组的两个句子之间的相似度，这时摘要抽取的句子中包含更多新闻要素细节信息，对整个新闻报道更具有代表性；而当 $\delta>5$ 时，同时包含要素词组的两个句子之间的相似度会逼近于 1，这时摘要抽取句子时过分依赖新闻的细节信息，导致

其他部分重要信息的丢失。

从实验结果中还发现，当 δ 取 20 和 50 时，5 个摘要的 F 值保持某个稳定值不变，这是因为 δ 大于某值时，同时包含要素词组的句子之间相似度已非常接近 1，随着 δ 增大，其对应 F 值不再发生变化。

目前本文关于每类事件最优权重系数的选取方法是：对实验语料进行分类，由于同一类事件中各个事件具有相似性，最优权值接近，训练得到每一类事件的最优权重系数。实验过程中先对每个事件进行分类，选择对应类别的权重系数完成摘要的自动生成。

4.2 所提方法和文献[14]时间轴摘要方法的实验对比

本文对文献[14]中的时间表摘要算法进行了程序实现，在网络新闻语料库的基础上与提出的基于要素提取的时间表摘要方法进行了实验对比。该实验中，所提算法中的权值参数 δ 设置为 5，对比实验结果如表 2 所列。

表 2 两摘要方法实验比较(%)

新闻事件	召回率		准确率		F 值	
	文献[14]	本文方法	文献[14]	本文方法	文献[14]	本文方法
温州动车追尾	41.33	**49.91**	31.05	**39.81**	35.45	**44.29**
莫言获诺贝尔奖	43.75	**49.25**	24.97	**36.33**	31.79	**41.81**
三亚游客挨宰	32.72	**44.32**	25.83	**29.17**	28.87	**35.18**
发布 iPhone5	34.81	**38.27**	27.88	**36.62**	30.96	**37.43**
阿姆斯特朗事件	27.37	**40.46**	20.14	**30.29**	23.20	**34.64**

从实验结果可以看出，提出的新闻摘要提取方法在 5 个新闻事件上均优于文献[14]中的方法，其中 4 个新闻事件对应的 F 值提高了 6%～10%左右，F 值提高最多的是“阿姆斯特朗”事件，本文方法比文献[14]的方法提高了 11.44%，证明本文提出的方法有效提高了时间轴摘要的质量。通过对实验数据分析，得出以下结论：(1)对于时间跨度比较大的新闻事件，本文方法和文献[14]方法的时间轴摘要效果最差。5 个事件中“阿姆斯特朗”时间跨度最大，达到了 3 年 6 个月，时间节点比较分散，每个时间节点上新闻报道数量较少，导致了全局摘要效果较差，从而影响了整个摘要的质量。(2)新闻报道句子数越多的事件对应的时间轴摘要质量越高，“三亚游客挨宰”事件句子数最少，效果最差；相反，“温州动车追尾脱轨”事件句子数最多，效果最好。(3)包含要素信息较多的新闻事件对应的摘要质量提高较大，“阿姆斯特朗”、“温州动车”、“莫言”3 个事件包含新闻要素信息比较丰富，采用本文摘要方法所得的 F 值提高了 9%～11%；“三亚游客”、“苹果”事件包含新闻要素信息相对较少，采用本文摘要方法所得的 F 值提高了 6%～7%，这也充分说明了提出的新闻摘要提取方法中新闻要素加权的重要性。

4.3 所提方法与文献[14]时间轴摘要方法的时间效率对比

在相同的实验环境下，对 5 个不同的新闻事件分别进行 20 次实验，平均运行时间如表 3 所列。

表 3 两摘要方法的时间比较

新闻事件	总句子数	文献[14]方法(s)	本文方法(s)
温州动车追尾	17044	220	242
莫言获诺贝尔奖	14952	185	201
三亚游客挨宰	5081	72	80
发布 iPhone5	10442	146	154
阿姆斯特朗事件	14905	182	198

实验结果表明，生成摘要的总时间随每个新闻事件报道句数的增加而增加。对每个新闻事件而言，本文方法的运行时间比文献[14]稍长，这是由于本文方法需要对其进行初始分类以及加权等操作，但两种方法运行时间的差值与生成摘要的总时间相比，仍在一个可以接受的范围内。

结束语 在传统时间轴摘要框架的基础上，结合新闻事件自身的特点，提出基于事件要素加权的新闻摘要提取方法，将新闻报道中出现事件要素信息的句子进行了相应加权处理，通过对转移概率矩阵进行改进，最终得到便于浏览的整个新闻事件的演变过程。实验表明，本文方法优于文献[14]中提出的时间轴摘要算法。

参 考 文 献

[1] Goldstein J, Kantrowitz M, Mittal V, et al. Summarizing text documents: sentence selection and evaluation metrics[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, 1999: 121-128

[2] Radev D R, Jing H, Sty M. Centroid - based summarization of multiple documents[J]. Information Processing and Management, 2004, 40(6): 919-938

[3] Canhasi E, Kononenko I. Multi-document summarization via Archetypal Analysis of the content-graph joint model[J]. Knowledge and Information Systems, 2014, 41(3): 821-842

[4] Cai Xiao-yan, Li Wen-jie. Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(5): 1597-1607

[5] Ferreira R, de Souza Cabral L, 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

[6] Xu Yong-dong, Zhang Xiao-dong, Quan Guang-ri, et al. MRS for multi-document summarization by sentence extraction[J]. Telecommunication Systems, 2013, 53(1): 91-98

[7] Luo Yi-hui, Xiong Shu-chu. A combination scheme for distributed multi-document summarization[J]. Journal of Intelligence, 2013, 32(11): 133-136(in Chinese)
罗毅辉，熊曙初. 一种集成框架下的分布式多文档自动摘要方法[J]. 情报杂志，2013，32(11)：133-136

[8] Wang Hong-ling, Zhang Ming-hui, Zhou Guo-dong. Chinese multi-document summarization system based on topic information[J]. Computer Engineering and Applications, 2012, 48(25): 132-136(in Chinese)
王红玲，张明慧，周国栋. 主题信息的中文多文档自动文摘系统[J]. 计算机工程与应用，2012，48(25)：132-136

[9] Canhasi E, Kononenko I. Weighted archetypal analysis of the multielement graph for query-focused multi-document summarization[J]. Expert Systems with Applications, 2014, 41(2): 535-543

[10] Swan R, Allan J. Automatic generation of overview timelines[C]// ACM SIGIR. Athens, 2000: 49-56

[11] Allan J, Gupta R, Khandelwal V. Temporal summaries of new topics[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, 2001: 10-18

[12] Chieu H L, Lee Y K. Query based event extraction along a timeline[C]// ACM SIGIR. Sheffield, 2004: 425-432

[13] Rui Yan, Wan Xiao-jun, Otterbacher J, et al. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution[C]// Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, 2011: 745-754

[14] Rui Yan, Liang Kong, Huang Cong-rui, et al. Timeline generation through evolutionary trans-temporal summarization[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, 2011: 433-443

[15] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[R]. Stanford Digital Library Technologies Project, 1999: 1-17

[16] Mei, Qiao zhu, Jian Guo, et al. DivKank: the Inte-rplay of Prestigo and Diversity in information Networks[C]// Special Interested Group on Knowledge Discovery in Databases. Washington, United States, 2010: 1009-1018

[17] Chen Ji-li, Niu Qin-zhou. Duplicated webpages deletion based on feature code[J]. Microcomputer Information, 2006(3): 113-115 (in Chinese)
陈基漓，牛秦洲. 基于特征码的网页去重[J]. 微计算机信息，2006(3): 113-115

[18] Xiong Zhong-yang, Ya Man, Zhang Yu-fang. Detection and climination of similar Web pages based on text structure and string of feature code[J]. Journal of Computer Applications, 2013, 33(2): 554-557(in Chinese)
熊忠阳，牙漫，张玉芳. 基于网页正文结构和特征串的相似网页去重算法[J]. 计算机应用，2013, 33(2): 554-557

(上接第 236 页)

然而，与 D 修复不同，I 修复或 U 修改的结果中可能包含 I 中不存在的元组属性值，插入元组一般根据指定的包含约束规则操作；修改元组的属性值一般用属性域内的常数或变量替代，有些 U 修改算法仅产生一个修复结果，有些产生多种修复结果[11]。因此，I 修复或 U 修改的不确定性较大。

总之，不管采用 D 修复方式还是 I 修复或 U 修复方式，我们都可以采用统一的查询语言结构以及修复系统模型。

结束语 本文将数据修复与一致性查询处理相结合，研究了基于 D 修复、满足多种类型约束的一致性查询方法。虽然 D 修复可能删去了原本正确的元组，但该方法能产生多种结果，使得所删去的元组可能在其他结果中出现。此外，扩展了标准 SQL 查询语言，采用蕴含关系表达多种类型的约束，定义了新的约束创建语句、一致性查询语句，提出了查询与修复系统模型。下一步将继续优化查询与修复算法，实现查询与修复系统功能。

参 考 文 献

[1] Greco S, Molinaro C. Querying and repairing inconsistent databases under three-valued semantics[C]// ICLP 2007. Springer Berlin Heidelberg, 2007: 149-164

[2] Arenas M, Bertossi L, Chomicki J. Consistent query answers in inconsistent databases[C]// Proceedings of the ACM Symposium on Principles of Database Systems. New York: ACM Press, 1999: 68-79

[3] Andrea R M, Bertossi L, Marileo M C. Consistent query answering under spatial semantic constraints[J]. Inf. Syst., 2013, 38(2): 244-263

[4] Bertossi L, Kolahi S, Lakshmanan Laks V S. Data cleaning and query answering with matching dependencies and matching functions[J]. Theory Comput. Syst., 2013, 52(3): 441-482

[5] Arenas M, Gottlob G, Pieris A. Expressive languages for querying the semantic Web[C]// Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems(PODS 2014). 2014: 14-26

[6] Fuxman A D, Miller R J. First-Order Query Rewriting for Inconsistent Databases[J]. Journal of Computer and System Sciences, 2007, 73(4): 610-635

[7] Wolf G, Kalavagattu A, Khatri H, et al. Query processing over incomplete autonomous databases: query rewriting using learned data dependencies[J]. The VLDB Journal, Springer Berlin Heidelberg, 2009, 18(5): 1167-1190

[8] Caroprese L, Greco S. Active integrity Constraints for Database Consistency Maintenance[J]. IEEE transaction on Knowledge and Data Engineering, 2009, 21(7): 1042-1058

[9] Wijsen J. Database repairing using updates[J]. ACM Transactions on Database Systems, 2005, 30(3): 722-768

[10] Kolahi S, Lakshmanan Laks V S. On approximating optimum repairs for functional dependency violations[C]// ICDT 2009. ACM Publisher, 2009: 53-62

[11] Beskales G, Ilyas Ihab F, Golab L. Sampling the Repairs of Functional Dependency Violations under Hard Constraints[J]. Proceedings of the VLDB Endowment, 2010, 3(1): 197-207

[12] Hu Yan-li, Zhang Wei-ming, Luo Xu-hui, et al. Dependencies Theory and its Application for Repairing Inconsistent Data[J]. Computer Science, 2009, 36(10): 11-15(in Chinese)
胡艳丽，张维明，罗旭辉，等. 基于数据依赖的数据修复研究进展[J]. 计算机科学，2009, 36(10): 11-15

[13] Cheng Lu-qing. Conditional functional dependency and data quality control[J]. Information System Engineering, 2009(11): 106-108(in Chinese)
程录庆. 条件函数依赖与数据质量控制[J]. 信息系统工程，2009(11): 106-108

[14] Geng Yin-rong, Liu Bo. Conditional functional dependencies for detecting data inconsistencies[J]. Computer Engineering and Applications, 2012, 48(3): 122-125(in Chinese)
耿寅融，刘波. 基于条件函数依赖的数据库一致性检测研究[J]. 计算机工程与应用，2012, 48(3): 122-125

[15] Beskales G, Ilyas I F, Golab L, et al. Sampling from repairs of conditional functional dependency violations[J]. The VLDB Journal, 2014, 23(1): 103-128

[16] Neehar C, Krishna T V S. Inconsistent relational data cleaning by detecting conditional functional dependencies[J]. International Journal of Computer Science and Information Technology & Security (IJCSITS), 2013, 3(1): 120-125

[17] Antova L, Koch C, Olteanu D. From complete to incomplete information and back[C]// SIGMOD'07. ACM, 2007: 713-724

[18] Chandel A, Hassanzadeh O, Srivastava D. Benchmarking Declarative Approximate Selection Predicates[C]// SIGMOD'07. ACM, 2007: 353-364