基于无向图构建策略的主题句抽取

葛 斌 李芳芳 李 阜 肖卫东
(国防科技大学C4ISR技术国防科技重点实验室 长沙 410073)

摘 要 基于文档句构建无向图,将主题句的抽取问题转换为无向图中节点的权重计算问题。首先利用滑窗方法抽取主题词,构建空间向量并生成无向图,然后基于向量空间模型计算边权重,最后利用文档句相似度矩阵的权重模型对文档句权重进行建模与计算,依据压缩比得到文档的主题句。实验表明,该方法在不同的压缩比下生成的摘要质量高,主题句抽取结果接近于人工摘要,召回率和准确率综合指数较高。

关键词 主题句抽取,无向图,文档句权重,自动文摘
中图法分类号 TP311 **文献标识码** A

Subject Sentence Extraction Based on Undirected Graph Construction

GE Bin LI Fang-fang LI Fu XIAO Wei-dong
(C4ISR Technology National Defense Science and Technology Key LAB, National Univ. of Defense Technology, Changsha 410073, China)

Abstract Undirected graph based on the sentence was proposed. The problem of sentence extraction was transformed to computing undirected graph node weights. This paper first proposed sliding window-based keywords extraction algorithm, followed by the establishment of the undirected graph. The edge weights of the graph were modeled by the Vector Space Model(VSM) in turn. The node weights were computed finally by the weight model based on the similarity matrix, and the subject sentences were obtained on the ratio of compression. Experiments show that the proposed automatic summarization techniques improve the recall rate and accuracy effectively.

Keywords Subject sentence extraction, Undirected graph, Sentence weight, Automatic text summarization

主题句抽取是自动文摘系统中的一项关键技术,抽取出的主题句集合是自动文摘系统输出结果的最终展现形式,直接关系到生成摘要质量的高低。

1 相关工作

基于图的排序算法被广泛应用于引用分析、互联网的链接分析,并取得了巨大的成功。基于图进行句子抽取之所以有效,是因为它通过迭代的计算能有效地获取图的全局信息,也即文本的全局信息,从而在判断句子重要程度上更为准确。相比采用一系列特征为句子打分的方法,这种方法有更好的鲁棒性,同时它通常不需要预先设定任何参数。

Mihalcea将这种方法应用到单文档自动文摘的句子抽取中,文献[1]讨论了如何将基于图的排序算法,包括PageRank,HITS等应用到句子抽取中来,实验证明基于图进行句子抽取能够提高文摘的质量。

Erkan[2]提出了LexRank[2],该方法也是基于图进行文本处理,将文本划分为句子集合,构造以句子为顶点的图,利用图计算句子的显著度(salience),并根据显著度抽取句子。同时Erkan还实现了采用LexRank方法的自动文摘系统,该系统在DUC04的多项任务的评测中位居第一位[3],表明了基于图进行句子抽取的有效性。

但上述两种算法在进行无向图的边权重计算时,Erkan只考虑了句子与句子之间的上下文关系,并没有考虑到句子与句子间的相似度,而Mihalcea在采用相似度计算时,构造的向量空间模型维数大小与句子中所包含的词成正比,计算长文档的时间开销较大。

本文从句子间的联系及相似度计算时的维度考虑,通过构建基于文档句的无向图,对基于图的排序算法进行改进。为描述方便,先给出两个定义:

定义1(文档句) 文档中能表示为一个句子的最小语义单元。

定义2(主题句) 能反映文档主题的文档句。

本文提出的方法的基本思想是:利用滑动窗口的主题词抽取算法(Sliding Window based Extraction Algorithm,简称SWE算法)获取文档的主题词,在文档主题词基础之上,构建基于文档句的无向图,将主题句的抽取问题转换为求解无向图中节点的权重问题。采用空间向量模型对各文档句进行建模,计算句子间的相似度,确定图的边权重;通过基于句子相似度矩阵的权重模型,计算无向图的相邻节点、边以及边权重,确定图的节点权重;最后,依据压缩比得到文档的主题句。

到稿日期:2010-06-20 返修日期:2010-09-08 本文受国家自然科学基金项目(60903225,60172012)资助。
葛 斌 博士生,讲师,主要研究方向为文本分析和语义检索,E-mail:gebin1978@gmail.com;**李芳芳** 博士,讲师,主要研究方向为信息资源管理;**李 阜** 硕士生,主要研究方向为社会化网络分析;**肖卫东** 教授,博士生导师,主要研究方向为信息管理、信息系统集成。

2 无向图构建与初始化

基于文档句的无向图构建与初始化主要包括两部分：节点及边的生成和边权重计算。

2.1 无向图节点与边的生成

同基于主题词的无向图构建过程相似，基于文档句的无向图生成策略主要分为两个步骤：

(1)建立“文档-无向图”的映射关系

将文档划分为句子的有序集合，每个文档句对应于无向图中的一个节点，文档句在文档句集合中的位置对应无向图中节点的编号，建立文档和无向图间的映射关系。

(2)确定无向图的边

为了更好地反映文挡句之间的相互关系，在每两个文档句之间均建立联系。即无向图中的节点两两之间均有一条边与对方相连。

假设文档 D 由 n 条文档句构成，则这些文档句构成的有序集合可以表示为：

$$\overrightarrow{DS}=\{S_1,S_2,\cdots,S_i,\cdots,S_n\}$$

式中，S_i 表示文档中的第 i 个句子，由此集合得到文档 D 所映射成的无向图，该无向图的节点表示为 $Vertex(i)=S_i$，无向图的边表示为 $Edge(S_i,S_j)$，且有 $Edge(S_i,S_j)=Edge(S_j,S_i)$。

2.2 基于相似度的无向图边权重计算

采用基于句子相似度的评价方法来衡量节点间的关系，并以此对无向图的边赋予权重。借用向量空间模型的思想，构造一个多维空间，将各个文档句转换成这个多维空间中的向量，把计算句子间的相似度转换成向量间的关系度量。

向量空间模型(Vector Space Mdoel，VSM)[4]是一种常见的用于文档表示的统计模型，是目前最简便有效的文本表示模型之一。向量空间模型不仅可以度量文档间的相互联系，还可以度量一篇文档中不同句子之间的相互关系。

利用分词集合构建向量空间，词语集合中的每类元素都代表向量空间中的一个维度，但该向量空间维数过于庞大，计算时间开销大。本文采用基于滑窗的主题词抽取进行特征项选择与降维处理，SWE 可以得到一个规模小得多的主题词集合，将这些主题词分别映射为向量空间中的一个维度，既降低了文档的特征项和向量空间的维数，同时也保持了文档的特征。

2.2.1 基于滑窗的主题词抽取

基于滑窗的主题词抽取的基本过程是：首先对文档进行拆分，将其中的名词抽取出来；然后通过滑动窗口来检测名词之间的相互联系，形成一个名词对的集合；最后通过在这些名词对之间建立联系，基于图的排序算法抽取出文档的主题词。相比于传统的主题词抽取算法，SWE 算法不但分析了文章中词语之间的关联关系，能够比较好地把握文章的内容和结构，而并且克服了基于理解的主题词抽取方法的知识库受限的问题。相比于基于聚类的主题词抽取算法，SWE 算法的整个运算过程不需要用户进行人工干预，属于无监督的抽取算法，因此自动化程度更高。

为更好地对文档中的主题词进行抽取，滑动窗口 σ 可以设定不同的值。文献[5]通过实验，考察滑动窗口长度对主题词抽取的影响，通过主题词权重方差和主题词权重偏移量两个指标进行分析，得到滑动窗口的长度不宜过大的结论。窗口长度 σ 值越大，所得到的主题词对的数量越多，可能会消减某些词语的重要程度，且会增加计算主题词权重的时间开销。因此，本文在进行名词对抽取时将窗口长度设为 3。

2.2.2 文档句相似度计算

基于向量空间模型的相似度计算步骤如下：

(1)构造向量空间

为了将文档中的所有句子映射到同一向量空间，将通过 SWE 算法抽取出的每个主题词作为向量空间中的一个维度。假设某文档 D_i 通过 SWE 算法抽取出的主题词集合为 $K=\{W_1,W_2,\cdots,W_l\}$，某文档句 S_i 由 m 个 K 中的元素构成，即：

$$S_i=\{\underbrace{\cdots,W_k,\cdots}_{m}\}\quad W_k\in K,m\leqslant l$$

设这 m 个元素构成的集合为 K'，则有 $K'\subseteq K$。那么 S_i 也可以表示为：

$$S_i=\{\underbrace{\cdots,W_k,\cdots}_{m},\underbrace{\cdots,W_e,\cdots}_{n}\}\quad \begin{cases}W_k\in K',W_e=0\\ m\leqslant l,m+n=l\end{cases}$$

当每个主题词都代表向量空间 V 中的一个维度时，该向量空间维度的集合可以表示为 $\{W_1,W_2,\cdots,W_l\}$，因此，任意的文档句都可以用这个向量空间中的一个向量来表示。

(2)文档句到向量空间的映射

文档句到向量空间的映射，采用对相应维度赋布尔值的方法，确定文档句在向量空间中的位置。以向量空间 V 为例，文档句 S_i 所代表的空间向量 α_i 构造算法如下所示。

算法 1 向量构造算法

```
Input:K,K',S_i,γ_i(长度为 l 的零向量)
Output:α_i
BEGIN
  foreach W_j∈S_i in K
  {
    if(W_j∈K')
        γ_i 将其第 i 维上的值置为 1;
    else
        γ_i 将其第 i 维上的值置为 0;
  }
  α_i=γ_i;END
```

(3)相似度计算

得到文档句的向量表示后，利用相似度计算向量间的相互关系，本文采用向量夹角余弦值的相似度计算方法来进行度量。即：

$$\mathrm{Sim}(S_i,S_j)=\cos\theta=\frac{\sum_{k=1}^{n}w_{ik}\cdot w_{jk}}{\sqrt{(\sum_{k=1}^{n}w_{ik}^2)(\sum_{k=1}^{n}w_{jk}^2)}}\tag{1}$$

式中，θ 表示 S_i,S_j 在 V 中的向量夹角；w_{ik} 为 S_i 在第 k 维上的取值；w_{jk} 为 S_j 在第 k 维上的取值；n 表示主题词所构建的向量空间的维数。

对文档中的句子两两计算其相似度，得到文档 D 的文档句相似度矩阵 M_sim。其中，$\mathrm{Sim}(S_i,S_j)$ 简写为 $\mathrm{Sim}_{i,j}$。

$$M\_\mathrm{sim}=\begin{bmatrix}\mathrm{Sim}_{1,1} & \mathrm{Sim}_{1,2} & \cdots & \mathrm{Sim}_{1,i} & \cdots & \mathrm{Sim}_{1,n}\\ \mathrm{Sim}_{2,1} & \mathrm{Sim}_{2,2} & \cdots & \mathrm{Sim}_{2,i} & \cdots & \mathrm{Sim}_{2,n}\\ \vdots & \vdots & \ddots & \vdots & & \vdots\\ \mathrm{Sim}_{i,1} & \mathrm{Sim}_{i,2} & \cdots & \mathrm{Sim}_{i,i} & \cdots & \mathrm{Sim}_{i,n}\\ \vdots & \vdots & & \vdots & \ddots & \vdots\\ \mathrm{Sim}_{n,1} & \mathrm{Sim}_{n,2} & \cdots & \mathrm{Sim}_{n,i} & \cdots & \mathrm{Sim}_{n,n}\end{bmatrix}$$

文档句之间的联系可以通过相似度矩阵 M_sim 得出，以此确定无向图的各条边的权重。

3 无向图节点权重计算模型

考虑无向图的带边权重，对图节点权重建模，计算各文档句节点的权重，依据压缩比抽取出文档中的主题句。

3.1 文档句权重计算方法

对无向图进行节点重要度建模时，可采用基于图的排序算法。该类算法是一种根据图的全局信息而非顶点局部信息来决定顶点的重要性的方法。如 PageRank[6]，HITS[7] 等被广泛应用于引用分析、互联网的链接分析，并取得了巨大的成功。PageRank 算法中有页面链出和链入的概念，其所构造的是一个有向图，而本文所构造的是基于文档句构建的无向图。

如果节点之间两两相连，就说明这两个节点存在联系，当计算某节点的重要度时，与之联系的节点越多，其重要度也就会越大。同时，节点的重要度不仅受与之相连的节点的影响，其本身的重要度也会对该节点的重要度产生影响。因此，图中节点的权重由与之联系的节点及其权重来决定。基于上述考虑，基于文档句的无向图节点权重模型可以表示为：

$$SW(i)=(1-d)+d*\sum_{j=1,j\neq i}^{n} M\_sim_{i,j}\cdot\frac{SW(j)}{\sum_{k=1,k\neq j}^{n} M\_sim_{j,k}} \tag{2}$$

式中，$SW(i)$ 表示节点 i 的权重；d 表示阻尼系数，一般设置为 0.85[6]；$M\_sim_{i,j}$ 是相似度矩阵 M_sim 中的值，表示文档 D 中第 i 条文档句和第 j 条文档句之间的相似度。

在实际运算中，需要对 $SW(i)$ 赋予一个初始值 C，通过设置一个收敛阈值，进行迭代运算。收敛阈值记为 η（预设为 10^{-5}）。实验表明，式(2)总是在有限次的迭代后收敛，实验结果如图 1 所示。

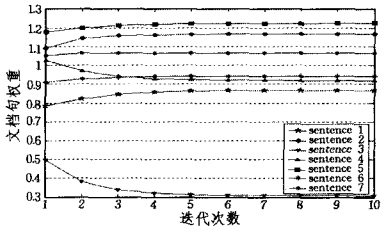

图 1 有限次迭代后无向图节点权重收敛情况

图 1 表示在给定的初始值下，利用包含 7 个文档句的文档进行权重计算时，每个文档句 S_i 的权重的变化关系。其中横坐标代表迭代的次数，纵坐标代表当前状态下文档句 S_i 的权重。可以看出，在无向图节点权重模型下，无向图节点的权重收敛十分迅速，文档句权重在经过少数几次迭代后，得到的权重值之差就远远小于 η。

3.2 基于节点权重的主题句抽取及文摘输出

计算出无向图节点权重，即得到文档句的权重。根据设定的压缩比，按文档句权重从高到低的顺序进行抽取，形成文摘并输出。

设文档 D 的文档句权重集合为 $SW=\{SW(1),SW(2),\cdots,SW(n)\}$，由于自动生成的文摘有固定的顺序，因此设生成的主题句有序集合为 $\overrightarrow{ES}$，文摘的压缩比为 pr，则主题句抽取及文摘输出算法如下所示。

算法 2 主题句抽取及文摘输出算法

```
input:SW,pr
output:ES→
BEGIN
  SW′=Rank(SW);
  while(ES′中主题句数量小于|pr*n|)
  {
    ES′.add(SW′[0]);
    SW′.remove(0);
  }
  ES→=Rank(ES′);
END
```

4 试验与结果分析

4.1 实验方案

对于输入的一篇待摘要的文章，经过分词、停用词过滤、词性过滤、主题词提取、文档句相似度计算、文档句权重计算、摘要句抽取及摘要生成等环节，最终输出一篇完整的摘要。

为验证本文所提方法的可行性，设计以下三个实验：

实验一 对大量的语料进行摘要实验，然后计算其召回率和准确率，计算其综合 F_measure 值，以验证本方法的效能；

实验二 对生成的摘要和传统的基于词频统计方法所生成的摘要进行对比分析；

实验三 在不同压缩比下进行自动摘要，检验对摘要生成的影响。

实验语料集来源于搜狗网上下载的语料库，从中选取 27 篇新闻类和 20 篇文学类文章，并从 CNKI 论文库中选取 20 篇科技论文，共 67 篇文章作为测试语料进行实验。对于每篇文章，由实验室的 3 名博士共同做出的手工文摘作为专家文摘。

实验参数设置如下：

滑动窗口的长度 σ 设定为 3。

生成摘要时的压缩比分别选定为 0.05，0.1，0.15，0.2，0.25，0.3。

4.2 实验评价指标

自动文摘评价方法是自动文摘技术研究中的一个关键部分，至今仍面临着许多争论与挑战[8,9]。结合自动文摘评测中人工评测[10]和自动评测[11]的优劣，采用将自动评测和人工主观评测相结合的评测方法，分别由自动文摘系统和人工挑选的方法产生文摘，针对两个文摘分别采用召回率、准确率以及 F-measure 3 个参数对系统进行客观评测。

(1)召回率(Recall)

召回率反映了系统正确选取的结果占所有可能正确结果的比例，计算表达式为：

$$R=\frac{D_{ik}}{D_s}$$

式中，D_{ik} 为自动文摘系统产生的文摘的句子包含在专家文摘中的数目，D_s 为专家文摘所包含的句子数目。

(2)准确率(Precision)

准确率反映了系统正确选取的结果占所有选取结果的比例，计算表达式为：

$$P=\frac{D_{ik}}{D_k}$$

式中，D_{ik} 含义同上，D_k 为文中自动文摘系统产生的文摘所包

含的句子数目。

(3)F_measure 指数

F_measure 指数是召回率和准确率的综合考虑，具体计算表达式为：

$$F\_measure=\frac{2\times P\times R}{P+R}$$

4.3 实验结果

4.3.1 摘要效能实验

本实验对选定的 67 篇语料进行摘要，分别计算出每篇摘要的召回率和准确率，然后计算 F_measure 值，部分实验结果如表 1 所列。

表 1 分类摘要评价结果

指标	新闻类	科技类	文学类
召回率(R)	0.608817	0.539133	0.486876
准确率(P)	0.61054	0.569086	0.538861
F_measure	0.609705	0.553705	0.511551
名词总数(W)	69.2963	70.05	121.15
主题词数(T)	39.37037	40.3	58.85
百分比(T/W)	57.04	52.84	47.66

通过表 1 可以看出，文摘的整体情况比较理想。新闻类文章采用本文的方法得到的摘要效果比较好，科技类的一般，文学类文章的摘要还需要进一步的修改。从表 1 的最后一行的数据，即主题词占名词集合的百分比可以发现，新闻类最高，其次是科技类文章，比例最低的是文学类。出现这种现象的原因是新闻类文章名词出现的比例较高，且分布较集中紧凑，文学类文章中词语构成比较复杂，描述性和说明性的词语比较多，名词的比例相比要小很多。

4.3.2 基于无向图构建策略和基于词频统计的摘要方法对比

本实验是对传统的基于词频统计的摘要方法和本文所提出的摘要方法进行对比，首先根据传统的基于词频统计的摘要方法对前面所选取的 67 篇语料进行摘要生成，然后计算出每篇摘要的召回率和准确率及 F_measure 值，实验结果如表 2 所列。

表 2 本文方法与基于词频统计的摘要系统的对比

类别	方法	召回率(R)	准确率(P)	F_measure 值
新闻类	词频统计	0.491585	0.510926	0.49472
	本文方法	0.608817	0.61054	0.609705
科技类	词频统计	0.46032	0.50297	0.47669
	本文方法	0.539133	0.569086	0.553705
文学类	词频统计	0.420445	0.50338	0.456105
	本文方法	0.486876	0.538861	0.511551

从表 2 可以看出，基于本系统所生成的摘要质量要优于传统的基于词频统计的摘要方法，尤其是在新闻类的文档的摘要中有着比较明显的优势。

传统的方法统计词语特征时考虑了词语的出现频率、位置等信息，在得到主题词的权重后计算句子的权重时考虑了句子包含主题词的数量以及句子的位置、线索词等信息赋予权重，最后根据压缩比按照文档句的权重由高到低抽取文摘句。与文方法相比，它没有考虑词和词之间的关系，简单的词频高低并不能完全代表词的重要程度，从文章的结构角度来说，词和词之间的联系、句子之间的关系在文章的内容上是有着关联的，和其他句子或者词关联越多就证明这个词或者句子有着更重要的地位，对其权重同样有着重要影响，所以用本文方法所生成的摘要效果比传统的方法有着明显改进。由于考虑的是名词之间的联系，对于新闻这类名词代表更多意义的文章，本方法能够较有效地得到反映文档主题的词和句子，从而生成质量较好的摘要。

4.3.3 不同压缩比生成摘要的对比

实验中对于前面所选出的 67 篇语料也分别在 0.05，0.1，0.15，0.2，0.25，0.3 的压缩比下进行了自动摘要，得到各自的召回率和准确率，然后计算在各个类别下的平均 F_measure 值，平均结果如表 3 所列。

表 3 不同压缩比下的摘要评价结果

类别	压缩比					
	0.05	0.1	0.15	0.2	0.25	0.3
新闻类	0.424	0.485	0.529	0.572	0.609	0.634
科技类	0.389	0.453	0.496	0.538	0.55	0.597
文学类	0.294	0.396	0.447	0.483	0.512	0.536

从结果中可以看出，随着压缩比的增大，摘要的 F_measure值随之变大，摘要的准确率和对原文要点的覆盖率逐渐提高，更能满足对原始文档内容的替代，可以提高在信息检索中的准确率。对于固定长度的一篇文档，当压缩比较小时，只抽取较少的句子，所生成的摘要内容比较粗陋，会漏掉文章的一些内容。而随着压缩比的增大，抽取句子的数目增多，所生成的摘要也逐渐变得比较完整，更全面地反映了文章的内容。但高压缩比和时间开销成正比，也会占用大量的空间资源，且摘要过长就失去了摘要的意义，因此实际中要综合考虑摘要的质量和时空开销，压缩比一般设为 0.25 或者 0.3。

4.3.4 时间复杂度分析

采用基于词频统计的方法计算句子相似度，对文档句中的每个词，统计其在文档中的出现频率，若文档中有 n 个待统计的词，则此类系统生成摘要的时间复杂度为 $O(n^2)$。

本文所提自动摘要方法的时间复杂度受节点重要度模型、相似度计算、空间向量模型和主题句权重模型等因素的影响。在进行文档句之间的相似度计算时，构造空间向量所需的时间复杂度为 $O(k^n)$，其中 k 为文档中句子的数量，而句子间相似度计算的时间复杂度为 $O(k^2)$。由式(2)可知，在利用节点重要度模型进行计算时，其时间复杂度为 $O(n^2)$。对于通常意义上的文档来说，句子数量一般不大于文档中词语的数量，即有 $k\leqslant n$。因此，本文方法时间复杂度仍为 $O(n^2)$，在提高摘要召回率、准确率和 F_measure 值的前提下，并没有降低算法的时间复杂度。

结束语 基于文档句构建无向图，在向量空间中度量节点间的相似度，基于相似度矩阵对节点权重建模，通过计算图中边与节点的权重，确定文档句之间的相互联系及文档句权重，最终得到文档的自动摘要。实验证明该方法是有效可行的。在该方法的基础上，可研究基于语义分析的主题词抽取，此外，文档句权重计算可采用无向图边剪枝策略、降维等方法进一步降低计算的时空复杂度。

参 考 文 献

[1] Milialcea R. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization[C]// Proc. of the 42nd Meeting of the Association of Computation Linguistics(ACL). 2004

[2] Erkan G, Radev D R. LexRank: Graph-based Lexical Centrality

as Salience in Text Summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479

[3] Passonneau R J. Applying the Pyramid Method in the 2006 Document Understanding Conference[C]// Proceedings of the 2006 Document Understanding Conference. Brooklyn, NY, 2006

[4] Salton G, McGill M J. Introduction to Modern Information Retrieval [M]. New York: McGraw-Hill Book Company, 1983: 400

[5] 李阜. 基于滑窗取词的单文档自动摘要技术研究[D]. 长沙: 国防科技大学, 2009

[6] Brin S. The Anatomy of a Large-scale Hypertextual Web Search Engine[J]. Computer Networks and ISDN Systems, 1998, 30: 1-7

[7] Kleinberg J M. Authoritative Sources in a Hyperlinked Environment[J]. Journal of the ACM, 1998, 46(5): 604-632

[8] Mani I. Summarization Evaluation: An Overview[C]// Proc. of the NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. Tokyo: National Institute of Informatics. 2001

[9] Radev D, Teufel S, Saggion H, et al. Evaluation Challenges in Large Scale Multi-Document Summarization[C]// ACL 2003. Sapporo, Japan, 2003: 7-12

[10] Saggion H, Teufel S, Radev D, et al. Meta-evaluation of Summarization In a Cross-lingual Environment Using Content-based Metrics[C]// Proc. of the 19th COLING. TaiPei, 2002

[11] Lin C Y, Hovy E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics[C]// Proc. of Language Technology Conference. Edmonton, Canada, 2003

(上接第 137 页)

图 4 描述了基于 Random Walk 数据集，每处理一个数据流元素，两种算法的运算时间。从图中可以看出，运算时间随采样率的提高而增长，由于 CluReservoir 算法在 ReservoirSample 的基础上进行了聚类运算，因此运算时间相对较长，但仍然表现出时间线性的特征，说明算法适用于数据流处理场景。图 5 显示的基于大规模实际数据的实验结果也验证了这一点。

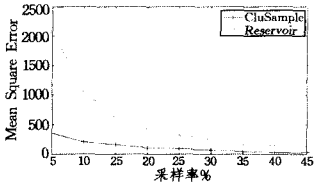

图 2 采样质量与采样率关系（合成数据，$\omega=256$）

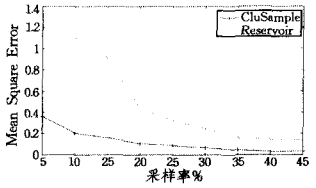

图 3 采样质量与采样率关系（真实数据，$\omega=10240$）

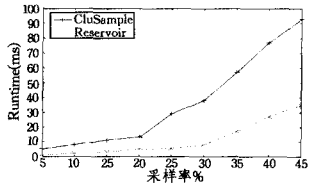

图 4 处理时间与采样率关系（真实数据，$\omega=256$）

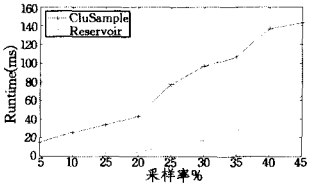

图 5 处理时间与采样率关系（真实数据，$\omega=10240$）

结束语 本文提出一种基于局部聚类的轨迹数据流摘要构造算法。该算法基于“偏倚采样”的思想，利用轨迹数据流局部连续性的特征，首先将指定时间范围内的数据流聚类划分成若干大小不一的数据窗口，并针对数据窗口大小定义偏倚系数，实现对窗口内数据的偏倚采样，从而构造数据流在滑动窗口内的摘要。由于算法充分考虑了数据的分布特征，对重要轨迹数据会自动提高采样率，因此，在平均采样率较低的情况下，其采样质量明显优于传统的蓄水池均匀采样。

参 考 文 献

[1] Sergio I, Eduardo M, Arantza I. Location-dependent query processing: Where we are and where we are heading[J]. ACM Comput. Surv., 2010, 42(3): 1-73

[2] Brian B, Shivnath B, Mayur D, et al. Models and issues in data stream systems[C]// Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Madison, Wisconsin, ACM, 2002

[3] Jeffrey S V. Random sampling with a reservoir[J]. ACM Trans. Math. Softw., 1985, 11(1): 37-57

[4] Charu C A. On biased reservoir sampling in the presence of stream evolution[C]// Proceedings of the 32nd international conference on very large data bases. Seoul, Korea, VLDB Endowment, 2006

[5] Kun-Ta C, Hung-Leng C, Ming-Syan C. Feature-preserved sampling over streaming data[J]. ACM Trans. Knowl. Discov. Data, 2009, 2(4): 1-45

[6] 张春阳，周继恩，钱权，等. 抽样在数据挖掘中的应用研究[J]. 计算机科学，2004，31(2)：126-128

[7] Arvind A, Singh M G. Approximate counts and quantiles over sliding windows[C]// Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Paris, France, ACM, 2004

[8] Panagiotis K, Nikos M. One-pass wavelet synopses for maximum-error metrics[C]// Proceedings of the 31st international conference on Very large data bases. Trondheim, Norway, VLDB Endowment, 2005

[9] Dimitris S, Antonios D, Timos S. Hierarchically compressed wavelet synopses[J]. The VLDB Journal, 2009, 18(1): 203-231

[10] Edith C, Haim K. Summarizing data using bottom-k sketches[C]// Proceedings of the twenty-sixth annual ACM symposium on principles of distributed computing. Portland, Oregon, USA, ACM, 2007

[11] Vladimir B, Rafail O, Carlo Z. Optimal sampling from sliding windows[C]// Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. Providence, Rhode Island, USA, ACM, 2009

[12] Christopher R P, Christos F. Density biased sampling: an improved method for data mining and clustering[C]// Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas, Texas, United States, ACM, 2000

[13] 余波，朱东华，刘嵩，等. 密度偏差抽样技术在聚类算法中的应用研究[J]. 计算机科学，2009，36(2)：207-209

[14] Yingyi B, Lei C, Wai-Chee F A, et al. Efficient anomaly monitoring over moving object trajectory streams[C]// Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris, France, ACM, 2009