一种解决“中心主题湮没问题”的基于图模型的Labeled-LDA文本分类算法

李 伟 马永征 沈 一
(中国科学院计算机网络信息中心 北京 100190)

摘 要 隐含狄利克雷分配(LDA,Latent Dirichlet Allocation)是一种用于挖掘文档集中潜在主题信息的无监督主题模型。而LDA模型的变形Labeled-LDA则可以作为有监督的多标签分类器,它建立了主题与标签的一一映射,从而学习出词与标签之间的关系。近年来,图模型在文本挖掘方面的应用取得了良好的效果,通过对文档建立图模型,为进一步分析文档的语义提供了新的途径。提出了一种利用Labeled-LDA和文档图模型进行文本分类的新算法,与传统的LDA模型方法相比,该方法的性能有较大的提高。

关键词 文本分类,图挖掘,图模型,隐含狄利克雷分配
中图法分类号 TP391.1 **文献标识码** A

Labeled-LDA Text Classification Algorithm Based on Graph Model for "Central Topic Oblivion Problem"

LI Wei MA Yong-zheng SHEN Yi
(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract Latent Dirichlet Allocation(LDA) is an unsupervised topic model used to mining potential topic information from the corpus. Labeled-LDA as a mutation of LDA can be used to do multi-classification on labeled documents, which establishes the one-to-one mapping from topic to label and learns the relationship between words and labels. Recently, the application of graph model has obtained good results in text mining, which provides a new way to analyze semantics of documents. This paper proposed a new method combining complex network theory and Labeled-LDA to do text classification. The experimental results show that our new method gets an improvement according to Macro_F1 compared to the traditional LDA model.

Keywords Text classification, Graph mining, Graph model, LDA

1 引言

近年来数据呈爆炸式增长,文本数据一直占据着非常重要的位置。如何从这些文本数据中挖掘出有用的信息一直是人们关注的热点话题。文本自动分类是文本挖掘中的研究热点,近年来得到了快速的发展[1]。

隐含狄利克雷模型(LDA,Latent Dirichlet Allocation)[5]是D. Blei等人提出的一个重要的无监督的主题模型,是完全的概率模型,它克服了PLSI[3]模型计算复杂度随文档数量线性增长的问题,并且具有行之有效的训练方法。它的变形Labeled-LDA[6]可以用作多标签分类器,它将主题与类别进行一一映射,较好地完成了多标签的分类问题。但Labeled-LDA应用于单标签分类时效果并不显著,会出现中心主题湮没问题(详述见2.4节)。

图模型[14]在文本挖掘方面的应用取得了良好的效果。传统的文本模型,例如VSM,没有考虑词与词之间的关系,通过为文档建立图模型不仅能够克服该问题,而且使得文档中词与词之间的关系有了更直观的展现,同时为挖掘文档语义层面的内容提供了新的途径[7]。

基于以上所述,为解决Labeled-LDA用于单标签分类时出现的主题湮没问题,本文通过为文档建立图模型,进一步挖掘词与词的语义关系,并通过提取词的主题成分来计算文档的中心主题,较好地克服了Labeled-LDA模型直接用于单标签分类时出现的中心主题湮没问题。实验表明该方法可以有效改进文本分类的性能。

本文第2节介绍LDA和Labeled-LDA模型,以及Labeled-LDA模型直接用于单标签分类出现的主题湮没问题;第3节阐述本文提出的文本分类新算法;第4节是实验结果展示;最后是总结和对下一步研究工作的展望。

2 LDA和Labeled-LDA在文本分类中的应用

2.1 LDA模型的基本思想

近年来概率统计方法在文本挖掘方面的应用取得了较大的成果。早期的典型代表是引入了隐性语义索引(LSI,Latent Semantic Indexing)[2]。隐性语义索引定义了语义维度的概念,将文档映射为语义空间上的一个表示,从而很好地达到

到稿日期:2013-05-03 返修日期:2013-07-05 本文受中科院十二五信息化项目“科研信息化应用推进工程”(XXH12503)资助。
李 伟(1988-),男,硕士生,主要研究方向为数据挖掘、信息检索,E-mail:liwei@cstnet.cn;马永征(1977-),男,博士,副研究员,主要研究方向为分布式计算、协同计算;沈 一(1988-),男,博士生,主要研究方向为网络协同、机器学习。

了降维的目的，并且较好地解决了语义之间的相关性问题。随后出现的 PLSI 模型[3]更加清晰地表现了文档-语义-词项之间的概率关系。但两者的计算复杂度都是不可忽视的问题[4]。

LDA 模型是一种对语料的生成式概率模型，属于层次型贝叶斯模型[12]。它克服了 PLSI 模型计算复杂度随文档数量线性增长的问题，同时，它的模型参数的数量也是固定不变的。它的基本思想是文档可以表示为隐含主题的一个随机混合，而每一个主题则是词项上的特征化的概率分布[4]。

语料中的词项构成词典，词典长度为 V，一篇文档由 N 个词组成，标记为 $W=\{w_1,w_2,\cdots,w_N\}$。一个语料由 M 篇文档组成，标记为 $D=\{W_1,W_2,\cdots,W_N\}$。

LDA 假定对语料 D 中的每篇文档 W 的生成过程如下：

1. 选择文档长度 N，$N\sim Poisson(\xi)$；

2. 选择 θ，$\theta\sim Dirichlet(\alpha)$；

3. 对文档中 N 个词项的每个词项 w_n，有

(a)选择一个主题 z_n，$z_n\sim Multinomial(\theta)$；

(b)以 z_n 为条件的概率 $P(w_n|z_n,\beta)$选出词 w_n。

如图 1，LDA 模型是典型的概率图模型，分为“语料-文档-词”3 个层次。LDA 模型的语料层是由参数 α 和 β 定义的，向量 α 反映了隐含主题在语料中的相对强弱，而 β 反映了隐含主题在词项上的概率分布。在文档这一层隐含变量 θ 反映了一篇文档在各个主题之上的多项式分布。在词这一层的随机变量是 z 和 W，z 是每个词在隐含主题上的分配，W 是文档的向量表示。向量 θ 往往被认为是一篇文档经过 LDA 模型降维后的向量表示。由于不能预先确定各个隐含主题在语料中的分布，一般使用对称的狄利克雷先验分布，α 和 β 根据经验分别设定为 $50/K$(K 为主题数量)和 0.1[9]，在采样过程中所有的隐含主题是平等的。

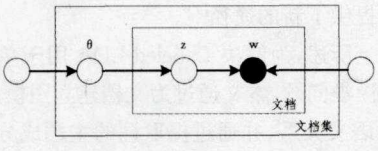


图 1　LDA 的概率图模型表示[5]

求解 LDA 主题模型一般有 3 种非精确的推导方法：1)吉布斯采样法；2)基于变分方法的 EM 求解；3)基于期望推进的方法[8]。本文实验采用的是吉布斯采样法，该方法求导过程简单快速，且推导结果较准确[9]。

LDA 模型是一个无监督的模型[13]，它尽管对于文档的隐含主题分布已经有足够的表现力，但仍然不能直接用在带标签的语料上，而且有时 LDA 会学习出一些难以解释的主题。Labeled-LDA 是 LDA 的变形，它建立起主题与类别之间一一映射的关系，将每个类别的文档映射到某一指定主题中。通过改变 LDA 的对称先验分布，达到了将标签信息融入到主题模型中的目的。利用吉布斯采样法，在训练过程中采样一个单词属于某个主题的概率为[10]：

$$P(z_{-i}=j|z_{-i},w)\propto\frac{n_{-i,j}^{(w)}+\beta}{n_{-i,j}^{(*)}+W\beta}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{d_i}+T\alpha}\tag{1}$$

式中，$n_{-i,j}^{(w)}$ 表示词 w 分配到主题 j 的数量，$n_{-i,j}^{(*)}$ 表示分配到主题 j 的词的总数，$n_{-i,j}^{(d_i)}$ 表示文本 d_i 中分配到主题 j 的词的数量，$n_{-i,\cdot}^{d_i}$ 表示文本 d_i 中词的数量。

2.2　Labeled-LDA 介绍

Labled-LDA 将类别标签融入到无监督的主题模型 LDA 中，构造一种有监督的主题模型，其运用到多标签文档分类、主题可视化、标签文档可视化等实际问题中，取得了良好的效果[6]。

如图 2 所示，Labled-LDA 利用向量 Λ 将类别与主题一一映射，例如，假设一个语料共有 4 个类别标签，其中一个文档 d 有两个类别标签，对于 $\Lambda^{(d)}=\{0,1,1,0\}$，表明文档 d 对应 2、3 标签，即主题。基于此，Labeled-LDA 在吉布斯采样过程中变换概率计算式(1)为式(2)：

$$P(z_{-i}=j|z_{-i},w)\propto\frac{n_{-i,j}^{(w)}+\beta}{n_{-i,j}^{(*)}+W\beta}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha_j}{n_{-i,\cdot}^{d_i}+T\alpha}\tag{2}$$

式(2)将 α_j 与类别标签联系起来，当 $\Lambda_j=0$ 时，$\alpha_j=0$，从而改变了狄利克雷先验，使得文档 d 中词的分配倾向于 2、3 标签，从而实现了标签与类别之间的映射。由此可以看出，Labeled-LDA 相较于 LDA 更适用于有监督的文本分类。

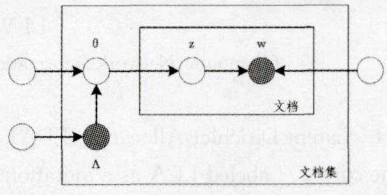


图 2　Labeled-LDA 的概率图模型表示[6]

2.3　Labeled-LDA 应用于文本分类

我们将 Labeled-LDA 模型用于训练过程，而推断过程仍然使用 LDA 方法，即训练过程中的采样方法使用式(2)，推断过程的采样公式为式(1)。

当使用 Labeled-LDA 推断一篇文档在隐含主题上的概率分布时，我们自然地可以认为该文档在某个主题上的词分配越多，它属于该主题对应类别的概率就越大。对于单标签分类，我们需要做的就是将正确的分类标签对应的主题概率最大化。一般的做法是将 Labeled-LDA 训练后得到的 θ 向量(即文档的主题概率分布)作为文档在主题层面上的表示，这样就起到了降维的目的。我们可以直接得到最大概率的主题对应的标签，也可以通过 SVM 进行训练，得到新文档的类别标签，以此达到分类的目的。但是有些文档通过 Labeled-LDA 训练得到的概率最高的主题与类别标签对应主题并不一致。作为示例，图 3 给出了一则经过 Labeled-LDA 词分配后的以健康为中心主题的新闻稿。

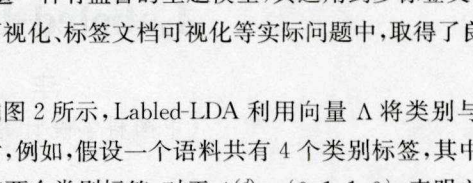


图 3　一则经过 Labeled-LDA 词分配后的新闻稿

由图 3 分析统计得到，分配到“健康”主题的词数小于分

配到“招聘”和“军事”主题的词数，如果直接以上述方法，该新闻就有很大概率被分到“招聘”类别中去。由此我们可以看出，直接以经过 Labeled-LDA 模型推断后词分配的数量作为分类依据，会造成某些文档误分类的现象，我们称之为“中心主题湮没”。

2.4 中心主题湮没问题

Labeled-LDA 推断过程中，某些词会以相近的概率被分配到几个甚至所有的主题中。例如训练文档集经过 Labeled-LDA 训练后，“商家”一词在 5 个主题中的分配数量如表 1 所列。

表 1 “商家”一词在 5 个主题中的分配数量

主题	健康	体育	招聘	文化	军事
数量	24	1	24	8	0

从表 1 可以看出，“商家”一词在“健康”和“招聘”主题中分配的数量最大，而且相等。根据式(2)，容易推出在对测试文档集进行推断的过程中，该词会以较高概率分配到“健康”或者“招聘”主题中。对于这种会以相近概率被分配到多个主题的词，我们称为非主题词，而那些高概率只隶属某一主题的词，我们称为主题词。当一个主题词分配到对应主题中(例如“健康”一词分配到一篇以“健康”为中心主题的文档中)时，我们认为该主题词对体现文档的主题的作用“较强”；相对地，如果一个非主题词分配到某一篇文档中(例如“商家”一词被分配到一篇以“健康”为中心主题的文档中)，我们认为该非主题词对体现文档的主题的作用“较弱”。

当一篇文档中绝大多数词语是隶属于某一主题的主题词，我们就可以认为该篇文档隶属于该主题的概率较大，分类结果也会很明确；但如果一篇文档由大量非主题词构成，并且由于采样带有一定的随机性，我们往往难以保证将这类文档归入正确的类别。

由此可见，如果我们将一篇文档中的非主题词和主题词同等对待，就会因分配的随机性导致中心主题不突出，甚至被湮没的问题(如图 3 所示)。

针对这个问题，本文提出了基于文档语义图的文本分类方法，通过建立文档图模型提取主题子图，计算各个主题子图的主题成分，减少非主题词对分类的影响，从而避免中心主题湮没问题。

3 基于文档语义图的文本分类方法

3.1 文档语义图和主题子图

对于一篇文档，我们可以将短语或者单词看作结点，用连线代表结点之间的关系，来构造文档语义图。本节将给出一种基于 Labeled-LDA 的文档语义图的构造方法。

在一篇经过 Labeled-LDA 分配后的文档中，每个词都会被分配到一个主题中，我们将一个词作为一个结点，本文文档图构造方法基于以下 3 条规则：1)相邻的词建立一条边，本文认为文档中上下单词之间有承接的语义关系；2)隶属于同一主题的相邻词建立一条边；3)两个结点之间只能有一条边。本文认为隶属于同一主题的单词在文档中有近似的语义关系。依此规则对 2.3 节中的新闻稿构造语义图，如图 4 所示。

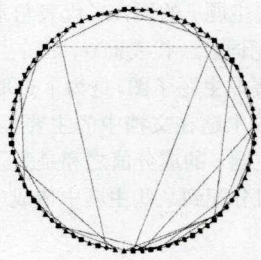


图 4 新闻稿语义图

如图 4 所示，不同形状的结点代表隶属于不同主题的词。词通过语义承接关系或是语义相近关系建立连接。我们选择隶属于某一主题的所有的词以及与这些词直接相连的其他主题词来提取主题子图，如图 5 所示。

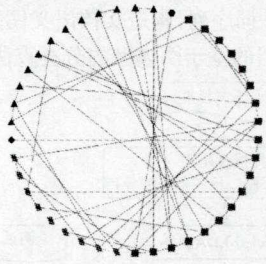


图 5 健康主题子图

通过提取主题子图，我们获得与主题语义最相近的词序列，为执行文本分类算法做准备。

3.2 词的主题成分提取

在经典的 tf-idf 模型[15]中，为了能够更好地衡量一个词在语料或一篇文档中的重要性，往往认为在文档中出现次数越多的词，它的重要性越高，而出现在越多文档中的词，它的重要性越低。我们参考这种方法，对经过 Labeled-LDA 模型训练后的训练集词的主题成分进行提取，提取方法如式(3)所示：

$$I_{ij}=\left(\frac{N_{ij}}{N_i}\right)^2 \tag{3}$$

式中，I_{ij} 代表词 i 在类别 j 上的成分，N_{ij} 代表在类别 j 中所出现词 i 的文档数目，N_i 则代表出现词 i 的文档数目。由此，图 3 中的新闻文档中“健康”一词的主题成分如图 6 所示，容易看出其属于健康主题的成分最多。

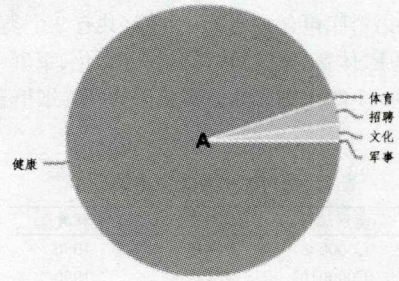


图 6 “健康”一词的主题成分图

3.3 基于图模型的文本分类算法

本文假设，文档中某一主题的成分越高，文档属于该主题对应类别的概率就越大。基于图模型的文本分类算法通过构造文档语义图，提取文档的主题子图，并计算所有主题子图中结点相应的主题成分之和，最大值所对应的主题就是该文档所属主题即类别。给定文档 W，计算方法如式(4)所示：

$$\hat{J}(w)=\arg\max_j\left(\sum I_{w_j j}+\sum I_{w'_j j}\right) \tag{4}$$

式中，w_j 代表隶属主题 j 的词，w_j' 代表与隶属主题 j 的词直接相连的其它主题词，$I_{w_j j}$ 代表词 w_j 的 j 主题成分。

基于图模型提取主题子图，是为了选取与该主题最相近关系的词来计算该主题在文档中的主要程度；计算 j 主题子图中各个结点的主题 j 的成分值之和是为了避免某些主题成分均衡的词因随机分配到某些主题中造成中心主题湮没的情况。

3.4 文本分类处理流程

基于以上所述，整合文本分类过程，描述如下：

根据图 7 所示，首先对训练文档集和测试文档集进行预处理(包括分词、停用词去除、过滤词频过少的词)，再使用 Labeled-LDA 对预处理后的测试文档集进行训练获得主题模型，并计算词典中各个词的主题成分。然后对测试文档集进行推断，获得主题-词序列，构造文档语义图并提取主题子图，再使用 3.3 节介绍的基于图模型的分类算法进行分类。

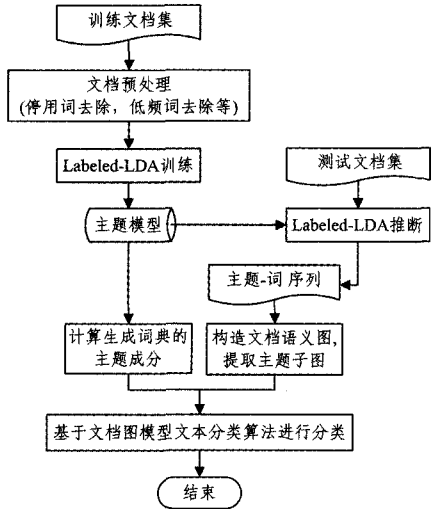


图 7 文本分类流程图

4 实验结果和分析

4.1 实验语料

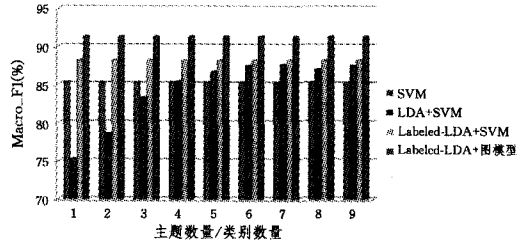
实验使用的文档集是搜狗实验室的新闻语料库(精简版 http://www.sogou.com/labs/dl/c.html)，源于 Sohu 新闻网站保存的大量经过编辑手工整理与分类的新闻语料与对应的分类信息。语料库包含文档 17910 篇，共有 9 个类别，分别为财经、IT、健康、体育、旅游、教育、招聘、文化、军事。训练集和测试集按照 5:1 的比例划分。实验语料库详细信息在表 2 中给出。

表 2 语料库中各类文本的分布表

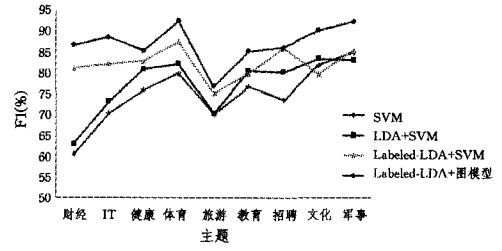
类别编号	类别名称	文档数量
C000008	财经	1990
C000010	IT	1990
C000013	健康	1990
C000014	体育	1990
C000016	旅游	1990
C000020	教育	1990
C000022	招聘	1990
C000023	文化	1990
C000024	军事	1990

4.2 实验结果以及分析

对比实验采用 Labeled-LDA＋SVM、LDA＋SVM 和 VSM＋SVM 的方法，这两种方法使用与本文算法相同的文本预处理方法，使用 LibSVM[11] 执行 SVM 算法，使用线性核函数。实验的评估方法采用 $F1$ 值和 $Macro\_F1$ 值。$F1$ 值用来分别评价各类别的分类性能，$Macro\_F1$ 值用来评价综合性能。由于 Labeled-LDA 将主题与类别一一映射，因此主题数量 $K=9$(类别数量)，根据经验设定 $\alpha=50/K$，$\beta=0.1$。实验结果如图 8 所示。



(a) $Macro\_F1$ 结果



(b) $F1$ 结果

图 8 实验结果展示

根据图 8 所示，Labeled-LDA＋SVM 算法的性能要高于传统的 LDA＋SVM 算法和 VSM＋SVM 算法的，而 Labeled-LDA 结合文档图模型的算法相较于 Labeled-LDA＋SVM 算法的 $Macro\_F1$ 值又约有 5.2%的提升。其次，相较于传统的 LDA 分类算法，本算法采用固定的主题数量，在主题维度更低的情况下得到了更好的分类效果，避免了最优主题数量的选取，在一定程度上节省了运算时间，提高了分类效率。

结束语 本文提出了一种结合 Labeled-LDA 和文档语义图的文本分类方法。首先，本文通过介绍 LDA 主题模型和 Labeled-LDA 模型，分析得出 Labeled-LDA 模型更适合应用于带标签的文本分类中；然后分析了利用 Labeled-LDA 对文档进行主题分配后出现的中心主题湮没问题，提出了一种基于文档图模型的文本分类算法。实验证明，本算法的文本分类性能与传统的 LDA＋SVM 方法和 VSM＋SVM 方法相比有所提高。

下一步我们将重点研究使用复杂网络方法挖掘更深层次的文档语义关系，提高文档语义图的表现力，进一步将本算法应用到更加复杂的文本挖掘场景中。

参 考 文 献

[1] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17：1848-1859

[2] Chen L，Tokuda N，Nagai A. A new differential LSI space-based probabilistic document classifier[J]. Information Processing Letters，2003，88(5)：203-212

[3] Hofmann T. Probabilistic Latent Semantic Indexing [C]// SIGIR. 1999：50-57

[4] 李文波，孙乐，张大鲲. 基于 Labeled-LDA 模型文本分类新算法[J]. 计算机学报，2008，31：620-627

[5] Blei D，Ng A，Jordan M. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research，2003，3：993-1022

[6] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. August 2009: 248-256

[7] 黄云平，孙乐，李文波. 基于上下文图模型文本表示的文本分类研究[C]// 第四届全国信息检索与内容安全学术会议论文集(上). 2008

[8] 赵鑫，李晓明. 主题模型在文本挖掘中的应用[R]. PKU-CS-NCIS-TR2011XX. June 2011

[9] Griffiths T L, Steyvers M. Finding scientific topics[C]// Proceedings of the National Academy of Sciences. April 2004, 101: 5228-5235

[10] Griffiths T. Gibbs sampling in the generative model of Latent Dirichlet Allocation[OL]. http://people.cs.umass.edu/~wallach/courses/s11/cmpsci791ss/readings/griffithso2gibbs.pdf

[11] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines[OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

[12] Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55: 77-84

[13] Blei D M, McAuliffe J D. Supervised topic models[C]// NIPS. 2007

[14] Cancho R F I, Sole R V. The small world of human language[J]. Proceedings of The Royal Society of London B: Biological Sciences, 2001, 268(1482): 2261-2265

[15] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523

(上接第 204 页)

的准确率。对比平均跃迁时间可以看出，改进 KAL-RFID 方法最少，因此，清洗效果较其他方法好。

表 1 跃迁响应表

	跃迁响应次数	跃迁响应判定准确率	跃迁响应时间	平均跃迁响应时间
WIN-5	28	83.3%	119	4.96
改进 SMURF	24	100%	37	1.54
KAL-RFID	22	91.7%	43	1.79
改进 KAL-RFID	24	100%	11	0.46

本文将标签的运动速度引入数据清洗过程，根据标签的运动速度动态设置置信度 δ。将阅读器的通信区域分成 N 个子区域，变量 S 表示一个子区域内标签被阅读器读到所需的最小读写周期数，可得到：

$$f=R/(V*T_{epoch})=S*N \tag{11}$$

其中，S 越小表示标签运动速度越快，S 越大表示标签运动的速度越慢。S 取 0 到 250，各算法清洗出错率效果图如图 5 所示。图 5 表明，标签移动越慢各算法的清洗效果越好，在标签移动速度非常快的情况下，各算法的清洗出错率仍较高。本文提出算法的出错率则较其他算法有较大的改善。

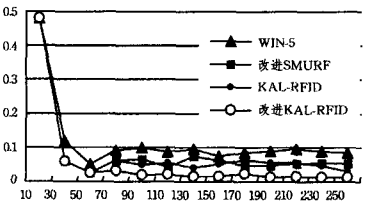


图 5 不同速度下算法的清洗出错率

4.4 空间代价分析

WIN-5 需要的存储空间最少，其算法的复杂度最低。改进 SMURF 算法滑动窗口的大小是可适应性变化的，清洗过程需要不断验证完整性条件及检测标签状态变化条件，因此输入输出数据量较 WIN-5 大，所需的空间代价相对 WIN-5 较高。KAL-RFID 通过时间更新方程和测量更新方程进行自回归逼近真实值，算法复杂度高，需要大量的输入输出数据，消耗的存储空间也较高。若不进行处理，改进的 KAL-RFID 方法与 KAL-RFID 有相当的复杂度，本算法在标签没有发生跃迁的情况下减少系统采样率，即扩大滑动窗口的大小，来减少数据的输入输出量；在检测到标签发生跃迁的时间段里通过减小滑动窗口保证预测值的准确性；因此在消耗空间代价及整体处理速度上较原始 KAL-RFID 方法有较大的改善。

结束语 本文针对现有清洗算法的不足，根据 RFID 数据的特点，提出了基于改进卡尔曼滤波的数据清洗算法。算法引入了标签的动态属性，根据标签运动情况动态调整滑动窗口大小，并将其应用到整个卡尔曼滤波的处理过程。算法动态控制采样率，及时调整阅读器读写率，使得卡尔曼预测过程使用的样本数据及更新过程作出的最近估计值都更接近真实数据。测试表明，改进 KAL-RFID 算法能很好地应用于标签频繁移动的场景。本文提出的算法主要解决数据清洗框架中单阅读器数据清洗层的问题，下一步的工作将致力于解决多阅读器数据清洗层的数据冗余问题以及判定标签的归属问题。

参 考 文 献

[1] 王霞，玄丽娟，夏秀峰. 基于时序关系的 RFID 不确定数据清洗算法[J]. 辽宁大学学报，2012，39(2)：174-178

[2] 马岩，张延园，尹方鸣. 基于滑动窗口的 RFID 数据流多标签清洗算法[J]. 科学技术与工程，2009，9(5)：1165-1171

[3] 李晓静，谷峪，吕燕飞，等. 基于动态事件概率模型的高效 RFID 数据清洗算法[J]. 计算机研究与发展，2008，45(Suppl.)：8-12

[4] 杨梦宁，赵鹏，张小洪，等. 一种基于总线模型的数据清洗方法[J]. 计算机科学，2010，37(4)：224-226

[5] Jeffrey R, Alonso G, Franklin M, et al. A pipelined framework for on line cleaning of sensor data streams[C]// Proc of ICDE, 2006. Washington: IEEE Computer Society, 2006: 773-778

[6] Bai Yi-jian, Wang Fu-sheng, Lin Pei-ya. Efficiently Filtering RFID data streams[C]// Proceedings of the 1st International VLDB Workshop on Clean Database, 2006. Seoul: Morgan Kaufmann Publishers, 2006: 50-57

[7] Jeffery S R, Garofalakis M, Franklin M J. Adaptive cleaning for RFID data streams[C]// Proceedings of the 32nd International Conference on Very Large Data Bases, 2006. Seoul: ACM, 2006: 163-174

[8] Li Xing, Fu Wen-xiu. Efficient RFID Data Cleaning Method[J]. TELKOMNIKA, 2013, 11(3): 1707-1713

[9] Meng Ling-yong, Yu Feng-qi. RFID Data Cleaning Based on Adaptive Window[C]// Proc of the 2nd International Conference on Future Computer and Communication, 2010. Wuhan, China, IEEE, 2010: 746-749

[10] Wang Yan, Song Bao-yan. Cleaning Method of RFID Data Stream Based on Kalman Filter[J]. Journal of Chinese Computer Systems, 2011, 32(9): 1794-1799

[11] Wang Fu-sheng, Liu Shao-rong, Liu Pei-ya. Complex RFID Event Processing[J]. VLDB Journal, 2009, 18: 913-931