LDA语义理解研究

高 阳 杨 璐 刘晓升 严建峰
(苏州大学计算机科学与技术学院 苏州 215006)

摘 要 潜在狄利克雷分配(LDA)被广泛应用于文本的聚类。有效理解信息检索的查询和文本,被证明能提高信息检索的性能。其中吉布斯采样和置信传播是求解LDA模型的两种热门的近似推理算法。比较了两种近似推理算法在不同主题规模下对信息检索性能的影响,并比较了LDA对文本解释的两种不同方式,即用文档的主题分布来替换原查询和文本,以及用文档的单词重构来替换原查询和文本。实验结果表明,文档的主题解释以及吉布斯采样算法能够有效提高信息检索的性能。

关键词 潜在狄利克雷分配,信息检索,近似推理,文本解释
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.057

Study of Semantic Understanding by LDA

GAO Yang YANG Lu LIU Xiao-sheng YAN Jian-feng
(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract Latent Dirichlet allocation(LDA) is a popular model used in text cluster, and is proved to improve the performance of information retrieval by explaining queries and documents effectively. There are mainly two algorithms to solve the inference of LDA model: Gibbs sampling and belief propagation. This paper compared the effect of these two inference algorithms on information retrieval in different topic scales, and used two different ways to explain queries and documents. One way is representing them with document-topic distribution, the other is representing them with word refactoring. Experimental results show that document-topic distribution and Gibbs sampling inference algorithm can improve the performance of information retrieval.

Keywords Latent Dirichlet allocation, Information retrieval, Approximate inference, Textual interpretation

1 引言

随着互联网的发展,每天会出现很多新的网页,而网页的内容也以较快的速度更新。当一个用户需要使用搜索引擎来查询想要的内容时,其需要提供相对准确的链接信息,以尽可能地识别用户的意图,更好地服务于用户。

信息检索是搜索引擎的核心,相似度计算是信息检索的主要决策方式,传统的向量空间模型(Vector Space Model, VSM)为相似度提供了较为简单的计算方式。而目前较为流行的主题模型也被逐步应用到信息检索中。从最初的潜在语义索引(Latent Semantic Indexing, LSI)到概率潜在语义索引(Probabilistic Latent Semantic Indexing, PLSI),再到潜在狄利克雷分布(Latent Dirichlet Allocation, LDA),主题模型在文本上的语义挖掘已经较为成熟。文献[1]首先提出了基于聚类的信息检索模型,其使用k-means算法进行聚类分析。文献[2]使用LDA来解析信息检索中的查询以及文本,与向量空间模型相结合,得到新的模型。该模型在信息检索上的性能高于基于聚类方式的性能,在TREC提供的几个测试集上取得了较好的效果。

目前变分贝叶斯(Variational Bayes, VB)[3]、塌陷吉布斯采样(Collapsed Gibbs Sampling, GS)[4]和置信传播(Belief Propagation, BP)[5]是LDA模型的3种常用近似推理技术。本文将比较其中的置信传播(BP)和塌陷吉布斯采样(GS)在不同主题规模下对信息检索效果的影响。当数据集所占存储为GB或以上级别时,其所包含的主题数也会在10^3以上,该种情况即为大规模下的大主题。本文使用两种方式来解释文本,一种是文档的主题分布特征,一种是文档的单词重构特征。两种解释方式的区别在于,后者能够从语义角度解释原文本。本文将这两种特征与基于向量空间的方法做对比,比较其性能上的差距。实验结果表明,BP在低主题下的表现效果更好,而GS更适合于大主题下信息检索上的文本解释,并且基于文档主题解释的LDA表示方式能更好地提高检索的准确率。

2 LDA近似推理算法分析

LDA模型是基于词袋(Bag of Words, BOW)的模型,

到稿日期:2014-09-07 返修日期:2014-12-08 本文受国家自然科学基金(61373092,61033013,61272449,61202029),江苏省教育厅重大项目(12KJA520004),江苏省科技支撑计划重点项目(BE2014005),广东省重点实验室开放课题(SZU-GDPHPCL-2012-09)资助。
高 阳(1991—),女,硕士生,主要研究方向为机器学习;杨 璐(1982—),女,副教授,硕士生导师,主要研究方向为机器学习与软件工程;刘晓升(1976—),男,博士生,主要研究方向为机器学习;严建峰(1978—),男,副教授,硕士生导师,主要研究方向为机器学习,E-mail:yanjf@suda.edu.cn。

不考虑文档与文档之间的顺序，同时也不考虑每篇文档里的单词顺序。如图1所示，LDA模型假设一篇文档是一些主题的分布，而一个主题是单词表上单词的分布。则一篇文档的生成过程如下所示，其中 Dir 代表狄利克雷分布：

$$\theta_d \sim Dir(\alpha), \phi_k \sim Dir(\beta), z_i \sim \theta_d, x_i \sim \phi_{z_i} \quad (1)$$

首先从一个基于 α 的狄利克雷先验中获得一篇文档 d 的分布 θ_d，从一个基于 β 的狄利克雷先验中获得每个主题 k 的分布 ϕ_k，从 θ_d 中获得一个主题 z_i，再从主题单词分布 ϕ_{z_i} 中获得一个单词 x_i，重复这样的过程直到得到所有的文档。表1给出了本文中与LDA模型相关的一些参数。

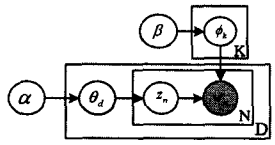

图1 LDA图模型

表1 符号标签

符号	意义
$1\leqslant d\leqslant D$	语料库文本索引
$1\leqslant w\leqslant W$	单词表中单词索引
$1\leqslant k\leqslant K$	主题索引
$x_{w,d}$	索引为{w,d}的单词的数目
$z_{-w,d}$	文本d中除w外所属的主题
$z_{w,-d}$	单词w除文本d外所属的主题
$\mu(z_{w,d}=k)$	将w分配给主题k的概率
$\mu(z_{\cdot,d})$	$\sum_w \mu(z_{w,d})$
$\mu(z_{w,\cdot})$	$\sum_d \mu(z_{w,d})$
$\theta_{k\|d}$	文本d在主题上的分布
$\phi_{w\|k}$	单词w的因子
$n_{k\|d}$	被分配给主题k的文档d的个数
$n_{w\|k}$	被分配给主题k的单词w的个数
ϕ_k	$\sum_w \phi_{w\|k}$
α,β	狄利克雷超参

变分贝叶斯是最原始的用来推理LDA后验概率的算法，其相关改进算法[6]与BP类似，而VB在速度和精度上都不如GS和BP。考虑到实验效率的问题，本文将主要介绍并比较基于GS和BP的两种LDA近似推理算法。

2.1 SparseLDA算法

SparseLDA[7]是一种改进版的GS算法，在速度以及内存消耗方面都做出了改进，其精度与GS保持一致。在GS中，给定一个文档 d 里的一个单词 w，该单词属于主题 k 的概率计算如下：

$$P(z=k|w) \propto (\alpha+n_{k|d})\frac{\beta+n_{w|k}}{\beta W+n_{\cdot|k}} \quad (2)$$

其中，$n_{w|k}$ 代表已分配给主题 k 的单词 w 的个数，$n_{k|d}$ 代表已分配给主题 k 的文档 d 的个数，而 $n_{\cdot|k}=\sum_w n_{w|k}$。SparseLDA将该概率分解成3部分：

$$P(z=k|w) \propto \frac{\alpha\beta}{\beta W+n_{\cdot|k}}+\frac{n_{k|d}\beta}{\beta W+n_{\cdot|k}}+\frac{(\alpha+n_{k|d})n_{w|k}}{\beta W+n_{\cdot|k}} \quad (3)$$

其中，第一项对所有文档都是常量，第二项独立于当前的单词 w。以这种方式切分计算，可以大幅度减少计算时间，速度最高可达到GS的20倍。而SparseLDA使用了稀疏编码，用一个32位的整型来同时存储二元组 $(k,n_{w|k})$，由于GS算法的稀疏性，该方法可以有效降低对内存的需要。而相比而言，另一种加速GS算法FastLDA[8]的速度最高可达到GS的8倍，而内存消耗上没有改进，效果没有SparseLDA明显。

2.2 BP算法

文献[5]首次用BP算法来求解LDA模型，使用马尔科夫随机场概念将LDA模型转换为树状因子图，用消息传递的机制来解释LDA。如图2所示，一篇文档 d 的一个单词 w 受到同一篇文档中不同的单词对 w 的影响记为 $x_{-w,d}\mu(z_{-w,d}=k)$，以及受到不同文档中的同一单词对 w 的影响记为 $x_{w,-d}\mu(z_{w,-d}=k)$。其中 $x_{w,d}$ 是指文档 d 中单词 w 出现的个数，$-w$ 是指除 w 以外的其他单词，$-d$ 是指除 d 以外的其他文档。

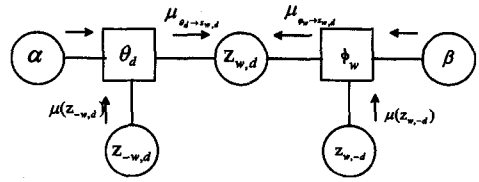

图2 基于因子图的置信传播

则一篇文档 d 的一个单词 w 被分配给第 k 个主题的概率为：

$$\mu(z_{w,d}=k) \propto \frac{\tilde{\mu}(z_{-w,d}=k)+\alpha}{\sum_k[\tilde{\mu}(z_{-w,d}=k)+\alpha]} \times \frac{\tilde{\mu}(z_{w,-d}=k)+\beta}{\sum_w[\tilde{\mu}(z_{w,-d}=k)+\beta]} \quad (4)$$

其中，

$$\tilde{\mu}(z_{-w,d}=k)=\sum_{-w} x_{-w,d}\mu(z_{-w,d}=k) \quad (5)$$

$$\tilde{\mu}(z_{w,-d}=k)=\sum_{-d} x_{w,-d}\mu(z_{w,-d}=k) \quad (6)$$

由此得到的文档主题分布以及主题单词分布的计算公式为：

$$\theta_{k|d}=\frac{\tilde{\mu}(z_{\cdot,d}=k)+\alpha}{\sum_k[\tilde{\mu}(z_{\cdot,d}=k)+\alpha]}$$
$$\phi_{w|k}=\frac{\tilde{\mu}(z_{w,\cdot}=k)+\beta}{\sum_w[\tilde{\mu}(z_{w,\cdot}=k)+\beta]} \quad (7)$$

2.3 算法复杂度分析

如表2所列，BP算法的时间复杂度为 $O(KDW_dT)$，空间复杂度为 $O(K*(NNZ+W+D))$，其中 K 是主题数目，D 是输入文档-单词共现矩阵的列数，W_d 是共现矩阵每一列中非零单元 $x_{w,d}\neq 0$ 的个数，T 是迭代次数，NNZ 为单词-文档共现矩阵中非0元素的总数目。SparseLDA算法的时间复杂度为 $O(\hat{K}DW_dT)$，空间复杂度为 $O(\hat{K}*(W+D)+ntokens)$，其中 $\hat{K}$ 远小于 K，$ntokens$ 表示所有文档的单词总数。从两种算法的空间复杂度看，差距在于 $ntokens$ 与 $K*NNZ$，一般情况下 $ntokens$ 是 NNZ 的5～20倍左右，由此可以看出，在主题数目 K 较小的情况下，两种算法的内存消耗差距是不明显的。而当 K 较小时，$\hat{K}$ 与 K 的差距也不大，所以此时两种算法的训练速度也是差不多的。

表2 SparseLDA和BP的复杂度比较

	时间复杂度	空间复杂度
SparseLDA	$O(\hat{K}DW_dT)$	$O(\hat{K}*(W+D)+ntokens)$
BP	$O(KDW_dT)$	$O(K*(NNZ+W+D))$

3 信息检索中LDA语义解释

信息检索包含很多带有结构的多媒体文档、有意义的文

本内容和其他媒体等。常见的信息媒体包括图片、视频、音频等。较为广泛的搜索情景是某一用户向搜索引擎输入一个查询，搜索引擎通过一系列的处理，反馈给用户一个经过排序的文档列表。而在当今能够处理数十亿网页的商业化网络搜索引擎时代，一个关键问题就是相似度。相似度[9-11]是信息检索中的基本概念，一个相关文档是指一个用户把查询发给搜索引擎后得到的信息。

3.1 相似度度量

文本相似度计算的主要方法有余弦法、内积法、Jaccard系数以及Dice系数。本文主要使用余弦法来计算两个向量的相似度。余弦相似度的几何意义在于，用向量空间中的两个向量夹角的余弦值来衡量两个个体之间的差异大小。余弦值越大，夹角越小，两个向量越相似。

文本相似度计算即在将输入文本转化为特征向量之后，使用cosine距离来计算查询与文档的相似度，q代表查询，d代表文档，$\vec{V}(q)$和$\vec{V}(d)$分别是查询和文档的表示方式；然后按照相关程度从高到低排序，选出最相关的前N个文档。

$$sim(q,d)=\frac{\vec{V}(q)\cdot\vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|} \tag{8}$$

3.2 检索精度评价标准

评估搜索引擎精度的指标有平均精度均值(Mean Average Precision，MAP)和准确度Precision@10即P@10。MAP指如果与q_j相关的文档是$d_1,d_2,\cdots,d_{m_j}$，m_j是与第j个查询相关的文档数目，而R_{jt}是排名最靠前的第d_t篇文档的相似度的结果，那么

$$MAP(Q)=\frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{t=1}^{m_j}precision(R_{jt}) \tag{9}$$

而P@10指标关注搜索结果排名最靠前的文档的结果质量，用于评估在搜索结果排名最靠前的10个文档中有多大比例是相关的。本文使用TREC提供的软件来计算两个评价标准。

3.3 LDA语义解释方式

3.3.1 向量空间模型

原始的向量空间模型是基于TF-IDF的，主要利用了文本的词频信息。VSM文本相似度计算方法是使用最广泛的文本相似度计算方法，这种方法以词在文本中出现的频率以及在文本集中出现的频率来表征词的权重，通过计算向量之间的余弦相似度来计算文本的相似度。该模型只考虑了词频信息，无法从语义的角度来理解查询或文本。VSM从词频和字符串比较两方面对信息检索中的查询和文本进行匹配，却忽略了词与词之间的关系，例如近义词、同义词以及有的词语通常成对出现等情况。

3.3.2 LDA语义模型

为了更好地理解用户所要表达的意图，本文使用LDA模型来解释查询与文本，从大量的查询中训练得到LDA模型。选择查询作为训练集的原因在于，精简短小的查询很好地表达了用户的意图，并且包含的语义比较全面。

首先给出文本上的主题解释，记为$p(z|d)$，来替换原始的VSM表示方法。文本上的主题解释即通过LDA来解释原始的文本，把一篇文本变成几个主题概率的组成。例如，一个查询主要包含了3个主题，每个主题的概率由大到小排序分别是0.5，0.2，0.1，剩下的0.2的概率由其他主题组成。

另一种使用LDA来解释文本的方法是文本上的单词表示，记为$p(w|d)$，即$p(w|z)p(z|d)$。该方法从一篇文本的主题概率分布中选择概率最大的前n个主题，再从每个主题的单词概率分布中选择概率最大的前m个单词，将这些单词的概率向量作为原文本的特征向量，构成文本的单词重构。

相比文本上的主题解释方法$p(z|d)$，该$p(w|d)$文本解释方法可以获取原文本与新文本的单词组成，可以直观地看出原文本与新文本的差别。而由于LDA是聚类算法，其主题解释无法知道每个类具体是什么样的标签，无法知道具体的含义，即无法从$p(z|d)$的概率分布中知道原始文本的实际意义，而$p(w|d)$能够具体地表示。

4 实验分析

本文使用的数据来自于中国第三大搜索引擎soso。本文使用1000万个查询来训练LDA模型，使用soso人工标注的测试集来评价其性能的好坏。soso人工标注的测试集包含922个查询以及58853篇相关的文档，每个查询与文档的相似度由3名编辑来标记，并取平均值。其中每个查询的长度较短，平均有5个单词。以下实验是在多核130GB内存的服务器上运行，有足够的内存空间，满足BP时内存的需求。实验使用对称超参，固定$\alpha=5/K$，$\beta=0.01$。

对于LDA模型质量的好坏，使用混淆度(perplexity)来初步计算，越低的混淆度值代表越好的泛化性能：

$$Perp=\exp\{-\frac{\sum_{w,d}x_{w,d}\log[\sum_k\theta_d(k)\phi_w(k)]}{\sum_{w,d}x_{w,d}}\} \tag{10}$$

4.1 模型精度与效率

首先，使用1000万个查询作为训练集，分别用SparseLDA算法和BP算法对训练集进行参数学习。表3列出了主题数目$K=\{50,100,150,200\}$时，SparseLDA算法与BP算法每次迭代的训练时间的对比数据。从表3可以看出，BP算法的计算时间是SparseLDA算法的5～8倍，且随着主题数目的增加而变长。而SparseLDA算法的计算时间对主题数目不敏感，主题数目的增加对其计算时间影响不大。该实验结果与前面所述的时间复杂度分析保持一致。

表3 不同主题数下，SparseLDA和BP每次迭代的时间(s)

主题数	50	100	150	200
SparseLDA	2.58	3.23	4.17	5.31
BP	11.34	22.43	33.22	44.12

虽然BP算法在计算时间方面有所不足，但从混淆度的角度来比较，BP算法的优势较为明显。图3和图4分别是在soso查询和soso文档上得到的混淆度结果。在soso查询上，BP算法得到的混淆度与SparseLDA算法的差异随着主题的增大而增大。在soso文档上，BP算法混淆度的下降幅度与SparseLDA差距不大，但是总体的混淆度依然比SparseLDA低6%～8%。

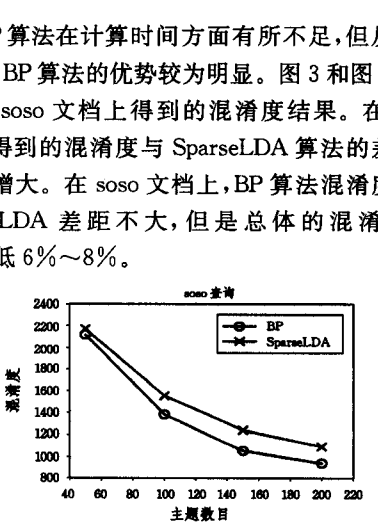

图3 $K=\{50,100,150,200\}$时，查询混淆度的变化

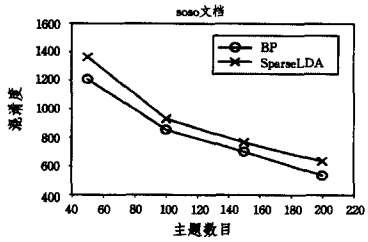

图 4　$K=\{50,100,150,200\}$时，文档混淆度的变化

4.2　相似度评测

使用 LDA 的两种语义解释方法：文档主题分布 $p(z|d)$ 和文档单词重构 $p(w|d)$来替换 VSM，在对 soso 查询和 soso 文档分布进行语义解释后，再比较查询向量与各个文档向量的相似度。

图 5 和图 6 分别是 BP 算法和 SparseLDA 算法在 $p(z|d)$以及 $p(w|d)$两种表示方式下与 VSM 表示方式做对比，得到的 MAP 结果以及 P@10 结果。从两幅图中都可以看到：$p(z|d)$表示方式得到的结果普遍高于 $p(w|d)$的。

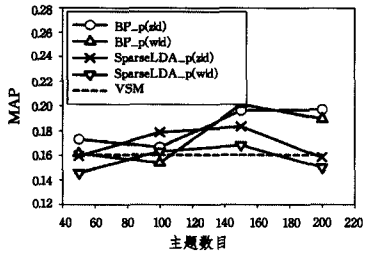

图 5　$K=\{50,100,150,200\}$时的 MAP 值

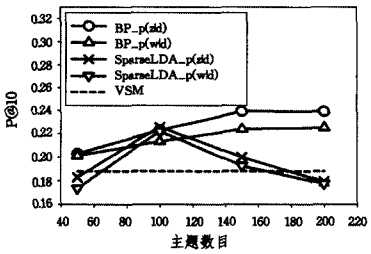

图 6　$K=\{50,100,150,200\}$时的 P@10 值

对于 BP 算法，其 MAP 以及 P@10 的值都随着主题数目的增加而增加。其基于 $p(z|d)$表示方法，在主题数目为 150 时，MAP 值相比 VSM 的 MAP 值提高了 25%，其 P@10 值相比 VSM 的 P@10 值提高了 33%。而 SparseLDA 算法与主题数目的关系不明显，但是在主题数目为 100 和主题 150 时的 MAP 以及 P@10 都高于由 VSM 得到的值，当主题数目为 100 时，其 MAP 值相比 VSM 的 MAP 值提高了 12.5%，P@10值提高了 22%。由以上实验可知，由 $p(z|d)$表示方式得到的结果，在主题数小于 200 时，BP 算法结果的整体趋势优于 SparseLDA 算法。为了研究更多主题下的情况，给出了 $K=\{100,500,1000,1500\}$时，基于 $p(z|d)$表示方式和 $p(w|d)$表示方式的几组实验，实验结果如表 4 所列。SparseLDA 算法随着主题数目的进一步增大，其 MAP 值也在不断地上升，而 BP 算法的 MAP 值只是上下略有波动，没有显著的上升趋势；并且，在更多的主题下，SparseLDA 算法所消耗的内存比 BP 算法要小得多。基于此分析，SparseLDA 算法更加适合多主题的稀疏性比较强的文本训练。

表 4　$K=\{100,500,1000,1500\}$时，SparseLDA 和 BP 的 MAP 值

主题数	100	500	1000	1500
SparseLDA_p(z|d)	0.1788	0.1844	0.1986	0.2122
SparseLDA_p(w|d)	0.1728	0.1784	0.1906	0.2042
BP_p(z|d)	0.1667	0.1784	0.1725	0.1668
BP_p(w|d)	0.1627	0.1764	0.1675	0.1648

在此，借助 $p(w|d)$表示方式来分析结果。通过输出 LDA 对查询和文档的解释，即输出一篇文档里概率最大的前几个主题的前几个单词，可以发现，BP 算法无法区分两个主题相似但略有区别的短文本，在主题数较少的情况下，这两个主题是合为一个主题的，但当主题数目较大，需要区分这两个主题时，BP 无法区分。例如一个查询包含“时尚 衣服”，一个查询包含“新装 衣服”，通过 BP 算法得到的这两个查询的文本解释的单词基本都是一样的，而通过 SparseLDA 得到的在前 5 个单词中就能区分开来，而这种区分能够更精准地识别用户的意图，结果如表 5 所列。经初步分析，在主题较大的情况下，“时尚 衣服”和“新装 衣服”应该被区分开，在主题较小的情况下，这两个查询可以表示为同样的主题分布。

表 5　$p(w|d)$单词重构展示

查询	SparseLDA	BP
时尚 衣服	时尚，正品，流行，女装，韩	新款，正品，时尚，女装，韩
新装 衣服	新款，正品，时尚，女装，韩	新款，正品，时尚，女装，韩

BP 算法在主题大与主题小的情况下得到的结果类似，而 SparseLDA 的随机性带来了相似语义之间的差异。另一个可能原因是 BP 不适合查询这样的短文本。相关长文本的研究将作为后续工作进行扩展。

结束语　本文分析对比了 LDA 主题模型的吉布斯采样和置信传播的优缺点，并将这两种推理算法的 LDA 模型应用到信息检索中。通过与传统的基于 TF-IDF 的向量空间模型对比，本文提出的基于 LDA 模型语义向量的两种文本解释方式在 MAP 以及 P@10 上取得了较大的提高，MAP 值最高可提升 25%，P@10 值最高可提升 33%。置信传播算法的稳定性以及精度效果都比 SparseLDA 的更好，在低主题规模下，其改进后的 MAP 以及 P@10 比 SparseLDA 的高，但在大主题规模下，在内存方面以及 MAP 值方面没有 SparseLDA 好，SparseLDA 更适合于大规模主题下的文本分析。

参 考 文 献

[1] Liu X Y, Croft W B. Cluster-based retrieval using language models[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrievel. 2004:186-193

[2] Wei X, Croft W B. Lda-based document models for ad-hoc retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrievel. 2006:178-185

[3] Blei D M, Ng A, Jordan M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1):993-1022

[4] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of USA, 2004, 101(1):5228-5235

[5] Zeng Jia, Cheung W K, Liu Ji-ming. Learning topic models by belief Propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 33(5):1121-1134

（下转第 304 页）

参 考 文 献

[1] 王亮,胡卫明,谭铁牛. 人运动的视觉分析综述[J]计算机学报,2002,25(3):225-237
Wang Liang, Hu Wei-ming, Tan Tie-niu. A survey of visual analysis of human motion[J]. Chinese Journal of Computers, 2002,25(3):225-237

[2] Stauffer C, Grimson W E L. Adaptive background mixture models for real-time tracking[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Fort Collins, USA, 1999:23-25

[3] Wren C, Azarbayejani A, Darrell T, et al. Pfinder: realtime tracking of the human body[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7):780-785

[4] Tuzel O, Porikli F, Meer P. A Bayesian approach to background modeling[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC, USA: IEEE, 2005:58-65

[5] Kim H, Sakamoto R, Kitahara I, et al. Background subtraction using generalised Gaussian family model[J]. IEEE Electronics Letters, 2008, 44(3):189-190

[6] Elgammal A, Duraisewami R, Harwood D, et al. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance[J]. Proceedings of IEEE, 2002, 90(7):1151-1163

[7] Elgammal A, Harwood D, Davis L S. nonparametric model for background subtracting [C]// Proceedings of 6th European Conference on Computer Vision. Dublin, 2000:751-767

[8] Monnet A, Mittal A, Paragios N, et al. Background modeling and subtraction of dynamic scenes[C]// Proceedings of the 9th International Conference on Computer Vision. Washington DC, USA: IEEE, 2003:1305-1312

[9] Ojala T, Pietikainen M, Maenpaa T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7):971-987

[10] Mason M, Duric Z. Using histograms to detect and track objects in color video[C]// Proceedings of 30th Applied Imagery Pattern Recognition Workshop. Washington DC, USA: IEEE, 2001:154-159

[11] Matsuyama T, Ohya T, Habe H. Background subtraction for nonstationary scenes[C]// Proceedings of Asian Conference on Computer Vision. Taipei, China: IEEE, 2000:622-667

[12] Heikkila M, Pietikainen M. A texture-based method for modeling the background and detecting moving objects[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4):657-662

[13] 徐剑,丁晓青,王生进,等. 一种融合局部纹理和颜色信息的背景减除方法[J]. 自动化学报,2009,35(9):1145-1150
Xu Jian, Ding Xiao-qing, Wang Sheng-jin, et al. Background Subtraction Based on a Combination of Local Texture and Color[J]. Acta Automatica Sinica, 2009, 35(9):1145-1150

[14] 李志焕. 改进型的 LBP 算法及其在运动目标检测中的应用[D]. 广州:中山大学,2009
Li Zhi-huan. Improved LBP Algorithm and its Application in Motion Detection [D]. Guangzhou: Sun Yat-sen University, 2009

[15] 陈雨丝. 基于背景差分的光照鲁棒性运动目标检测与跟踪技术研究[D]. 成都:西南交通大学,2011
Chen Yu-si. Research on Moving Objects Detection and Tracking Based on Background Subtraction with Illumination Robustness[D]. Chengdu: Southwest Jiaotong University, 2011

[16] 邓宇,李振波,李华. 图切割支持的融合颜色和梯度特征的实时背景减除方法[J]. 计算机辅助设计与图形学学报,2006,18(11):1741-1747
Deng Yu, Li Zhen-bo, Li Hua. A Fusing Color and Gradient Features Approach to Real-time Subtraction using Graph Cuts[J]. Journal of Computer-Aided Design & Computer Graphics, 2006, 18(11):1741-1747

[17] Koller D, Weber J, Huang T, et al. Towards robust automatic traffic scene analysis in real-time[C]// The 12th International Computer Vision and Image Processing Conference. 1994:126-131

[18] 张旗,梁德群,樊鑫,等,基于小波域的图像噪声类型识别与估计[J]. 红外与毫米波学报,2004,23(4):281-285
Zhang Qi, Liang De-qun, Fan Xin, et al. Identifying of Noise Types and Estimation of Noise Level for a Noisy Image in Wavelet Domain[J]. Journal of Infrared and Millim Waves, 2004, 23(4):281-285

(上接第 282 页)

[6] Asuncion A U, Welling M, Smyth P, et al. On smoothing and inference for topic models[C]// Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. 2009:27-34

[7] Yao L, Mimno D M, McCallum A. Efficient methods for topic model inference on streaming document collections[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009:937-946

[8] Porteous I, Newman D, Ihler A T, et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Knowledge Discovery and Data Mining. 2008:569-577

[9] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. England: Cambridge University Press, 2008

[10] 李峰,李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报,2007,21(3):99-105
Li Feng, Li Fang. An New Approach Measuring Semantic Similarity in Hownet 2000[J]. Journal of Chinese Information Processing, 2007, 21(3):99-105

[11] 江敏,肖诗斌,等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报,2008,22(5):84-90
Jiang Min, Xiao Shi-bin, et al. An Improved Word Similarity Computing Method Based on Hownet[J]. Journal of Chinese Information Processing, 2008, 22(5):84-90