标签传播算法理论及其应用研究综述*

张俊丽，常艳丽，师　文
（南京大学 信息管理学院，南京 210093）

摘　要：介绍了标签传播算法理论，分析了标签传播算法的特点，总结了其在多媒体信息检索、分类、标注、处理和社区发现等方面的应用研究，最后探讨了标签传播算法未来的研究方向。

关键词：标签传播算法；半监督学习；多媒体；社区发现
中图分类号：TP301　　**文献标志码**：A　　**文章编号**：1001-3695(2013)01-0021-05
doi:10.3969/j.issn.1001-3695.2013.01.004

Overview on label propagation algorithm and applications

ZHANG Jun-li, CHANG Yan-li, SHI Wen
(*School of Information Management*, *Nanjing University*, *Nanjing* 210093, *China*)

Abstract: This article introduced the theoretical study of label propagation algorithm, analysed its characteristics and summarized its applications in multimedia information processing, retrieval, annotation, classification and community discovery, etc. Finally, this paper proposed the future prospects and the trends of developments of the LPA algorithm.

Key words: label propagation algorithm(LPA); semi-supervised learning(SSL); multimedia; community discovery

机器学习算法可以分为有监督学习和无监督学习算法两大类。所谓有监督学习，是指从已经标注好类别的数据样本中学习；而无监督学习，是指根据数据本身的内在特点进行学习，样本事先并没有清晰的分类。半监督学习(SSL)是一种监督学习和无监督学习相结合的方法，其主要思想是：基于数据分布上的模型假设，利用少量的已标注数据进行指导并预测未标记数据的标记，然后合并到标记的数据集中[1,2]。

标签传播算法[3](LPA)是由 Zhu 等人于 2002 年提出，它是一种基于图的半监督学习方法，其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息。利用样本间的关系建立关系完全图模型，在完全图中，节点包括已标注和未标注数据，其边表示两个节点的相似度，节点的标签按相似度传递给其他节点。标签数据就像是一个源头，可以对无标签数据进行标注，节点的相似度越大，标签越容易传播。由于该算法简单易实现，算法执行时间短，复杂度低且分类效果好，引起了国内外学者的关注，并将其广泛地应用到多媒体信息分类、虚拟社区挖掘等领域中。本文利用关键字 label propagation、标签传播、标签传递、标记传播、标记传递等词作为关键词，对国内外数据库及网络资源进行了检索，结果发现，目前国内外相关文献期刊论文约有 90 篇，其中国外 82 篇，国内 8 篇，国内外硕博论文 3 篇。

1　标签传播算法基本理论

根据 LPA 算法基本理论，每个节点的标签按相似度传播给相邻节点，在节点传播的每一步，每个节点根据相邻节点的标签来更新自己的标签，与该节点相似度越大，其相邻节点对其标注的影响权值越大，相似节点的标签越趋于一致，其标签就越容易传播。在标签传播过程中，保持已标注数据的标签不变，使其像一个源头把标签传向未标注数据。最终，当迭代过程结束时，相似节点的概率分布也趋于相似，可以划分到同一个类别中，从而完成标签传播过程。

具体算法[3]如下：令$(x_1,y_1)\cdots(x_l,y_l)$是已标注数据，$Y_L=\{y_1\cdots y_l\}\in\{1\cdots C\}$是类别标签，类别数 C 已知，且均存在于标签数据中。令$(x_{l+1},y_{l+1})\cdots(x_{l+u},y_{l+u})$为未标注数据，$Y_U=\{y_{l+1}\cdots y_{l+u}\}$不可观测，$l\ll u$，令数据集 $X=\{x_1\cdots x_{l+u}\}\in R^D$。问题转换为：从数据集 X 中，利用 Y_L 的学习，为未标注数据集 Y_U 的每个数据找到对应的标签。

将所有数据作为节点(包括已标注和未标注数据)，创建一个完全连接图，其边的权重计算式如下：

$$w_{ij}=\exp\left(-\frac{d_{ij}^2}{\sigma^2}\right)=\exp\left(-\frac{\sum_{d=1}^{D}(x_i^d-x_j^d)^2}{\sigma^2}\right)\tag{1}$$

其中：d_{ij}表示任意两个节点的欧氏距离，权重 w_{ij}受控于参数 σ。

为衡量一个节点的标注通过边传播到其他节点的概率，在此定义一个$(l+u)\times(l+u)$概率传递矩阵 T 如下所示：

$$T_{ij}=P(j\to i)=\frac{w_{ij}}{\sum_{k=1}^{l+u}w_{kj}}\tag{2}$$

其中：T_{ij}是节点 j 到 i 的传播概率。

同时定义一个$(l+u)\times C$的标注矩阵 Y，令 $Y_{ic}=\delta(y_i,c)$，它的第 i 行代表着节点 y_i 的标注概率，第 c 列代表类别，若 $Y_{ic}=1$ 则表示节点 y_i 是属于 c 类别，否则为 0。通过概率传递，使其概率分布集中于给定类别，然后通过边的权重值来传递节点标签。矩阵 Y 的初始值并不重要，但是要保证他的每行都是标准化的。算法描述如下：

输入：u 个未标注数据，l 个标注数据及其类别 C。

输出：u 个未标注数据的标注。

a)初始化，利用式(1)计算边权重矩阵 w_{ij}，得到数据间的相似度。

b)根据步骤 a)得到的 w_{ij}，利用式(2)计算节点 j 到 i 的传播概率。

收稿日期：2012-06-28；**修回日期**：2012-08-14　　**基金项目**：国家社科基金重大资助项目(10&ZD134)

作者简介：张俊丽(1982-)，女，湖北松滋人，博士，主要研究方向为多媒体信息检索与分类、机器学习(elili62@126.com)；常艳丽，博士研究生；师文，博士研究生.

c)定义一个$(l+u)\times C$维的标注矩阵Y。

d)每个节点按传播概率把它周围节点传播的标注值按权重相加,并更新自己的概率分布:

$$F_{ij}=\sum_{k=1}^{l+u}T_{ij}\quad 1\leqslant i\leqslant l+u;1\leqslant j\leqslant C \tag{3}$$

e)限定已标注数据,把已标注数据的概率分布重新赋值为初始值。重复步骤d),直到收敛。注意保持已标注数据点的标注源不变,把它的值限定为Y_l,不断地把标注从高权值传播到低权值。

$$F_{ij}=Y_{ij}\quad 1\leqslant i\leqslant l;1\leqslant j\leqslant C \tag{4}$$

2 标签传播算法的特点

a)LPA是一种半监督学习算法,具有半监督学习算法的两个假设前提:(a)邻近的样本点拥有相同的标签,根据该假设,分类时边界两侧尽可能避免选择较为密集的样本数据点,而是尽量选择稀疏数据,便于算法利用少量已标注数据指导大量的未标注数据;(b)相同流形结构上的点能够拥有相同的标签,根据该假设,未标记的样本数据能够让数据空间变得更加密集,便于充分分析局部区域特征。

b)LPA只需利用少量的训练标签指导,利用未标注数据的内在结构、分布规律和邻近数据的标记,即可预测和传播未标记数据的标签,然后合并到标记的数据集中。该算法操作简单、运算量小,适合大规模数据信息的挖掘和处理。

c)LPA可以通过相近节点之间的标签的传递来学习分类,它不受数据分布形状的局限,可以克服一些算法只能发现“类圆形”结构的缺点。只要同一类的数据在空间分布上是相近的,那么不管数据分布是什么形状,都能通过标签传播将它们分到同一个类里。因此可以处理包括音频、视频、图像及文本的标注、检索及分类,算法简单、执行速度快、可扩展性强、效果好。

3 标签传播算法的应用研究

LPA的上述特点决定了它具有较强的实用性,在多媒体信息检索与分类、网络社区发现、多媒体信息标注与处理等诸多领域都取得了令人鼓舞的成果。本文通过对国内外研究文献的整理,总结出学术界对标签传播算法的应用研究。

3.1 文本信息检索与分类

文本分类[4]是指在给定的分类模型下,将用自然语言表示的文本,根据其内容自动地确定该文本所属的一个或多个类别。信息检索主要是对用户需求和信息资源分别标引,对这两个标引结果进行匹配,并将匹配成功的资源文档作为检索结果返回给用户。实质上,文本分类和信息检索技术都是进行标引和匹配的,文本分类是将类别与文本内容进行匹配,而信息检索是将用户需求与信息资源进行匹配。在文本分类与信息检索中,其关键技术是分类算法或检索算法,该算法的性能直接影响到信息分类或者检索效果。

LPA能够根据用户的要求,将类别标签或者用户的需求标签按照相似度将标签传播至未标注文本或网页,该相似度决定了其相邻节点间传播的概率,对类别或者用户需求与信息资源进行匹配,从而实现检索与分类、语义消歧、语料标注、信息推荐、情感分析等。这种算法的最大优点是能够利用有限的标注数据,对大量的未知信息或者网页进行标注,减少人工训练标签数据的工作量。

Yang等人[5]创新性地利用LPA算法进行英汉双语信息检索。使用OKAPI BM25作为检索模型,从初始检索结果的排序中提取出伪标注文件,采用LPA对未标注文本进行标注,并利用K-means聚类信息,然后对文档按照相关度重排。

Kim和贺松林等人[6,7]提出利用LPA进行网页分类。前者假设未标注文本容易被相同用户查询和点击,搜索引擎的点击日志里包含了相似类别网页的链接信息,利用用户的点击信息发现相似页面,构建相似文本点击图,自动扩大训练数据。其方法是通过构建训练模型——点击数据模型和查询约束模型,计算相似度,然后利用LPA对未标注信息标注,从而解决网页分类问题。而贺松林等人主要是对LPA进行改进,利用K-means和LPA进行网页分类,其主要特点是针对网页特点,使用欧氏距离构建一个带权图,利用标签传播算法将已标记节点的标签沿着边向相邻节点传播,从而将网页分类问题转为标签在图上的传播,并结合K-means聚类在保证精度的情况下降低维度,提升标签传播的效率及性能。

Blair-Goldensohn等人[8]创新性地提出将LPA应用于情感分类(也称为极性分类),他利用少量人工标签引导极性词典,基于LPA算法,充分利用上下文和用户提供的标签,对顾客购后评价进行情感分类。Rao等人[9]提出利用基于图的LPA算法进行极性分类,每个节点代表待分类的词,边的权重表示两个词之间的相似度,每个节点有两个标签,即正面的和负面的,利用WordNet和OpenOffice两个词库的英语、法语和北印度语进行实验。Speriosu等人[10]在已有研究基础上,对上述两种方法进行改进,通过增加节点和边构建一个有向的不对称关系图,直观地显示其追随者关系图,并结合词性链接,利用LPA进行极性分类。

任晓娟和郝建柏等人[11,12]利用LPA算法进行文本分类。任晓娟指出,LPA算法应用于文本分类存在噪声问题,提出LPA在标签传播的过程中,计算每个点和各个类别的相似度。如果某个节点和每个类别的相似度都比给定的阈值小,那么就把该点设定为噪声点,即不属于任何类别,剔除该噪声点后再进行信息分类。郝建柏等人则先利用模糊近邻算法,使样本与其k个近邻连接,然后利用LPA使类别标签从已标注数据向未标注数据传递,从而实现分类。

Niu等人[13]假设类似的样本拥有类似的标签,未标注样本的标签不仅受邻近标注样本的影响,也受到邻近未标注样本的影响。利用数据加权图作为向量,使LPA自动地在邻近向量之间传播标签,从而进行语义消歧。

Lansdall-Welfare等人[14]创新性地构建了一个文本稀疏数据图,将文档用顶点图表示,利用LPA沿着图的边缘传播文本数据的标签,使距离近的文本倾向于拥有相同的标签,可对不正确的文档标注及时地重分类和重标注,从而提高标注质量。

上述文献的主要作者和主要贡献如表1所示。

3.2 多媒体信息标注、检索与分类

多媒体信息的标注主要是将已标注的多媒体样本(如图像、音/视频等)看做是分类系统中的类标签,利用机器学习算法从已标注的多媒体样本集中,自动学习语义概念空间与多媒体特征空间的关系模型,从而对未标注样本生成类别标签,完成多媒体样本信息的标注。常用的标注算法有贝叶斯(Bayes)、支持向量机(support vector machine, SVM)、二维隐马尔可夫模型(2-dimensional hidden Markov model, 2D-HMM)等[15]。

表 1　LPA 应用于文本信息检索与分类的主要研究者及主要贡献

主要作者	主要贡献
Yang 等人[5]	将 LPA 应用于英汉双语信息检索
Kim 等人[6]	通过构建相似文本点击图进行网页分类
贺松林等人[7]	构建相似度带权图,基于 K-means 和 LPA 进行网页分类
Blair-Goldensohn 等人[8]	利用少量人工标签引导极性词典进行情感分类
Rao 等人[9]	利用 LPA 进行极性分类,在 WordNet 和 OpenOffice 词库的英、法和北印度语进行实验
Speriosu 等人[10]	构建有向的不对称关系图,结合词性链接,利用 LPA 进行极性分类
任晓娟[11]	LPA 在标签传播过程中计算每个点和各个类别的相似度,剔除噪声点后再进行文本分类
郝建柏等人[12]	LPA 结合模糊近邻算法,进行文本分类
Niu 等人[13]	利用数据加权图作为向量,使 LPA 在邻近向量之间传播标签,进行语义消歧
Lansdall-Welfare 等人[14]	构建一个文本稀疏数据图,提高文本标注质量

多媒体信息的检索和分类问题,究其根本也是多媒体信息的标注与匹配问题,其基本过程是先建立多媒体信息内容的描述(如用户需求描述或者多媒体类别描述),然后利用机器学习方法学习多媒体信息类别或用户需求信息并进行标注,最后利用学习得到的模型对未知信息进行检索、分类和匹配。有实验结果表明,LPA 在解决此类问题时,能够充分利用有限的标注数据,减少训练标签数据的人工量,不需要监督学习即能准确地完成信息标注,分类或检索效果较好,很适合复杂的多媒体信息处理。

随着网络多媒体视频数据的快速增长,有效的音/视频标注、分割和分类日益重要。很多学者将 LPA 应用于音/视频信息标注中,并针对传统算法中存在的不同问题各自提出了其解决方案。其主要思路是在视频序列的两端——开始帧和结束帧,分配人工标签,然后 LPA 利用较少的多媒体信息标签作为样本,将标签传播至视频序列的其他帧中,获得视频序列的像素级标签,从而进行音/视频标注、检索和分类。

Badrinarayanan 等人[16]为提高传统算法中长视频信息标注的准确度,提出了一个概率图模型,在视频序列的开始帧和结束帧中给定手工标签,将视频视为多个静态的图片,利用隐马尔可夫模型(HMM)挖掘图像像素、图片补丁或者语义区域。在视频标注过程中,引入一个时间序列模型,然后基于 LPA 将视频两端的标签传播至未标注帧中,实现视频序列的分类和分割等。Budvytis 等人[17]通过手工标注视频的开始帧和结束帧构建有向图,利用 LPA 和变分期望最大化算法(variational expectation maximization algorithm,VEM)将标签传播至视频的每个像素,每个像素被分配不同类别或者空标签,将分类器估计的像素标签反馈到 Bayes 网络中,构建原始序列及其时间反转序列的叠加传播。倪煜等人[18]指出传统方法不能解决视频在时间上的延续性,未能考虑连续多帧的相互关系,提出利用 LPA 对镜头连续帧进行时—空分析,挖掘镜头视觉本征特性,针对不同镜头变换设置不同初始标签,利用视频流中连续多帧之间的相关性将给定的初始状态标签通过相关性进行传播,从而获取镜头边界特征,并利用 SVM 进行分类。Tang 等人[19]为了解决语义分布中局部线性限制问题,提出将核线性近邻传播(kernel linear neighborhood propagation,KLNP)应用到视频标注,该算法将 LPA 结合一致性假设和局部线性嵌入算法(local linear embedding, LLE)构建一个非线性映射内核空间,利用非线性核映射空间来优化重构系数, 从而改进了 LPA。

利用 LPA 进行图像标注的基本思想是[20]:利用已标注图像集或其他可获得的信息自动学习图像的语义概念空间与视觉特征空间的潜在关联或者映射关系,预测未知图像的标注。此类一般先将图像分割成子图像块,提取出其颜色特征、纹理特征等局部视觉信息,找出图像间的关系模型,计算分段子图像块的相似性,然后再对其进行信息标注和分类。

Zhong 等人[21]基于模糊语境的 LPA 进行图像标注,分配图像语义区域。利用模糊表示和模糊逻辑描述空间不变量的语境信息,标签在图像间采用基于视觉特征双向层稀疏编码进行传播,根据空间不变量之间分段子图像块的相似性计算,利用 LPA 使标签在图像内传播形成图像内部的语义区域关系,最后利用基于 K 近邻的模糊 C-均值对图像分类。Liu 等人[22]利用双层稀疏编码,将每一层稀疏编码产生的图像标签分配给那些共享图像标签的子图像块和候选区域,利用 LPA 传播标签,当所有候选区域的双层稀疏编码结果被融合,并赋予一个区域标签,这样可将图像标签分配到相应的语义区域标签内。Wang 等人[23]利用能量最小化的 HMM,结合 LPA 实现交互式多标签图像/视频分割。通过图像点阵像素,给定一个预先标注好的种子样本,反复传播该种子标签,从而为未标注数据进行标注,并允许用户根据分割结果调整初始标记的种子。

Yang 等人[24]针对传统的社会化标签准确度不高的问题,提出先利用 LPA 挖掘音乐的内容相似性,然后利用主成分分析法寻找到标签空间的低维结构,从而探索音乐路径之间内容的相似性以及路径—标签矩阵中的语义冗余,计算基于音乐特征向量与人工标注之间的相似度,然后重标注社会化标签,剔除社会性标签中的噪声标签。

上述文献的主要作者和主要贡献如表 2 所示。

表 2　LPA 应用于多媒体信息检索与分类的主要研究者及主要贡献

主要作者	主要贡献
Badrinarayanan 等人[16]	为提高传统算法中长视频标注的准确度,提出概率图模型,利用隐马尔可夫模型挖掘图像信息,引入时间序列模型进行视频标注,主要应用于视频序列的分类分割等
Budvytis 等人[17]	构建有向图,利用 LPA 和 VEM 将标签传播至视频的每个像素,构建原始序列和它的时间反转序列的叠加传播
倪煜等人[18]	利用 LPA 对镜头连续帧进行时—空分析及视频流中连续多帧之间的相关性,将给定的初始状态标签通过相关性进行传播,从而获取镜头边界特征
Tang 等人[19]	为解决语义分布中局部线性限制问题,提出将 KLNP 应用到视频标注
Zhong 等人[21]	基于模糊语境的标签传播算法进行图像标注
Liu 等人[22]	利用双层稀疏编码,将图像标签分配到相应的语义区域标签内
Wang 等人[23]	利用能量最小化的 HMM,结合 LPA,实现交互式多标签图像/视频分割
Yang 等人[24]	探索音乐路径之间内容的相似性及路径—标签矩阵的语义冗余,剔除社会性标签的噪声标签

3.3　社区发现

社区发现能够在大型复杂网络中自动搜寻或发现社区,具有重要的实际应用价值[25],如社会网络中的社区代表的是根据兴趣或背景而形成的真实的社会团体,引文网络中的社区主要是代表针对同一主题的相关论文,万维网中的社区大多表示讨论相关主题的若干网站,而生物化学网络或者电子电路网络中的社区可能就是某一类功能单元等。目前常用的社区发现算法包括基于完全二分图核的社区发现算法、PageRank、HITs、基于最大流的社区发现算法等。Symeon 等人[26]通过实验比

较了17种常见的社区检测算法的复杂度和适应社区规模的大小,比较了一般矩阵和稀疏矩阵两种情况下算法的复杂度,结果发现,LPA的复杂度分别为 $O(n^2)$ 和 $O(n)$ (n 为社区节点数),较之其他算法,LPA复杂度最低,能够很好地适应大规模社区的监测(106~109个节点),经过5次迭代后开始收敛,既不需要优化预定义的目标函数,也不需要关于社区的数量和规模等先验信息,对社区的大小也没有限制。

2007年,Raghavan等人[27]最早提出将LPA应用于社区发现,该算法被简称为RAK算法。其主要思想是利用网络结构作为指导来探测社区结构,每个节点被初始化为一个独特的标签,并在每一步迭代中,其节点标签选择为其大多数邻近节点的标签,在这个反复的过程中,密集连接的节点组从一个独特标签变成了一个具有共识的社区节点,那些具有相同标签的节点便组合成了同一个社区。该算法的流程包括初始化,标签更新,最后添加具有相同标签的顶点到同一个社区;通过在空手道俱乐部网和美国大学橄榄球网的实验结果表明,其社区检测效果良好。此后,很多研究人员都展开了相关方面的研究,不断改进RAK算法,使之更好地应用于社区发现。

Barber等人[28]为了避免LPA中所有顶点都分配到同一社区,改进RAK算法,提出了一种模块化标签传播算法(modularity-specialized label propagation algorithm, LPAm),即基于约束的LPA监测网络社区。给定一个目标函数,使得LPA受到约束,引入一个变量使其社区的模块度值最大化,将社区发现问题转换为目标函数最优化的求解问题,在具有相同标签相互连接的顶点个数基础上,定义一个目标函数 H,利用LPA算法发现 H 函数的局部最优值。

Liu等人[29]发现上述LPAm易在模块空间中陷入局部极大值,从而导致社区检测不准确的问题,提出将LPAm与多步贪婪凝聚算法(multi-step greedy agglomerative algorithm, MSG)融合,设计了一种模块化专业化的标签传播算法(modularity-specialized label propagation algorithm, LPAm+),该算法利用MSG同时合并多个相似社区,能够避免陷入局部最大值,更加精准地检测网络社区。

Gregory[30]对RAK算法进行了扩充,使之能够检测重叠社区,提出了一种挖掘重叠社区结构的算法COPRA,每个节点可以保留若干个社区标签,从而使传播过程中包括多个社区的信息。实验证明该算法能够有效地检测重叠社区,但每次迭代时间有所增加,且当混合或重叠社区太多时,可能会导致不正确的标签随机选择,导致性能下降。因此该算法对规模较小的重叠社区发现较为有效。

金弟等人[31]认为传统方法或因时间复杂度较高、或因搜索能力偏弱,不能对大规模复杂网络进行有效聚类,提出了基于遗传算法的复杂网络社区探测方法,在分析网络模块性函数 Q 的局部单调性的基础上,给出一种快速、有效的局部搜索变异策略——局部搜索的遗传算法,利用LPA作为初始种群的生成方法,从而提供高精度和多样性的初始种群,然后将标签传播至未标注节点。

有的学者从不同角度改善了LPA在虚拟社区发现中的执行速度。如Xie等人[32]通过研究实验发现LPA在经过五次迭代后,95%的节点已正确地聚集;后面的迭代主要是对社区内节点的更新,是不必要的。因此改进了LPA的更新准则和迭代规则,减少了原LPA中不必要的更新和迭代,通过记录现有社区的边界节点信息来节约算法迭代时间,将节点分为沉默节点和活跃节点两种,当所有节点都变成沉默节点时,即达到了算法收敛。Cordasco等人[33]提出了一种半同步的LPA,通过对网络顶点并行着色,使任何两个相邻的顶点都不共享相同的颜色,分步同时传播标签。该算法结合了同步和异步模式的优势,能够克服任何网络的振荡问题,可以收敛到一个稳定的标签,每一步的传播都并行作业,从而改善了LPA的执行速度。Leung等人[34]指出模块最大化方法不是一个无标度区间测度方法,仅依靠它检测社区不可行,提出了一种扩展的LPA用于实时社区监测,采用启发式方法提高其平均检测性能和适应性,利用同步和异步LPA检测网络社区,提高社区监测速度,通过简单的参数调整使算法具有一定的可扩展性,可适应用于不同规模的网络。

上述文献的主要作者和主要贡献如表3所示。

表3 LPA应用于社区发现的主要研究者及主要贡献

主要作者	主要贡献
Symeon等人[26]	通过实验比较了17种社区发现算法的执行效率和效果
Raghavan等人[27]	将LPA应用于社区发现中
Barber等人[28]	避免LPA中所有顶点都分配到同一社区,在RAK基础上提出了LPAm算法
Liu等人[29]	解决LPAm易陷入局部极大值问题,将LPAm与MSG融合,提出LPAm+算法
Gregory[30]	提出了一种挖掘重叠社区结构的算法COPRA,扩展RAK算法,检测重叠社区
金弟等人[31]	提出了基于遗传算法的复杂网络社区探测方法,以解决大规模复杂网络的有效聚类问题。
Xie等人[32]	改进了LPA的更新准则和迭代规则,避免算法不必要的迭代,优化检测速度
Cordasco等人[33]	提出了一种半同步的LPA,改善利用LPA进行社区发现的速度
Leung等人[34]	指出模块最大化方法不是一个无标度区间测度方法,仅依靠它检测社区不可行;提出了一种扩展的LPA用于实时社区监测,提高RAK算法检测速度

4 结束语

LPA能够利用少量已标注节点预测未标记节点的标签信息,将标签传播至未标注节点。该算法的特点决定了它是一种通用的半监督学习新方法,在理论和实际应用中表现出很多优越的性能。LPA的优越性使得它在检索与分类、多媒体信息标注及处理、社区发现等领域都得到了长足的发展和广泛应用,是一种很有研究前途的计算方法,为半监督学习和数据挖掘提供了有效的手段。虽然LPA的研究已取得很大进展,但由于针对LPA的集中研究只有五年左右的时间,仍存在较大的研究空间。未来的研究工作可从以下两个方面展开:

a)加强算法的理论研究。(a)从计算复杂性和收敛性角度,深入分析算法的性能;(b)标签传播算法的可适用性在很大程度上依赖于算法的相似度、判别概率和迭代算法等,可从不同角度改善该算法的性能,完善算法结构,提高算法执行效率;(c)LPA以无噪声干扰的数据为研究对象,然而实际应用中的数据很容易受到噪声干扰,“纯净”样本往往难以获得,可将抗干扰技术引入标签传播算法中,有利于提升LPA的鲁棒性;(d)LPA常用于处理大规模的音/视频图像等多媒体数据,有效地降维能够提高该算法的普适性。

b)探索LPA新的应用领域。目前对于该算法的应用只有比较有限的实验研究成果,大多属于仿真和对比实验,缺乏LPA在实际生活中的应用研究。本文认为可以应用于如下几个领域:

(a)应用于图书、博物和档案的多媒体信息服务。图书、

博物和档案的保存、收集与应用方式等方面已经实现从传统方式向现代化数字方式的快速转变,图博档数字信息资源的共建共享迫在眉睫,现代多媒体技术使图博档中不同的信息资源形式具备了相互融合的可能和趋势[35]。可以利用 LPA 算法检索图博档多媒体信息资源,分析图书、博物、档案多媒体信息的利用率,根据用户的检索要求,主动推送相关的多媒体信息,深入挖掘图书、博物、档案多媒体资源的内在关系,提供个性化的多媒体信息服务。

(b)应用于农村信息服务,如检索、分类有关三农的多媒体信息资源,从而为农民提供多样化的、高质量的信息。

(c)应用于电子商务,如挖掘用户的商务行为,主动推送相关的商品和服务。

参考文献:

[1] ZHU Xiao-jin, GHAHRAMANI Z, LAFFERTY J. Semi-supervised learning using Gaussian fields and harmonic functions [C]//Proc of the 20th International Conference on Machine Learning. 2003:328-335.

[2] OLIVIER C, BERNHARD S. Semi-supervised learning[M]. Cambridge: MIT Press,2006:1-53.

[3] ZHU Xiao-jin, GHAHRAMANI Z. Learning from labeled and unlabeled data with label propagation, CMU-CALD-02-107[R]. Pittsburghers:Carnegie Mellon University, 2002.

[4] 张俊丽. 文本分类中的关键技术研究[D]. 武汉:华中师范大学,2008.

[5] YANG Ling-peng, JI Dong-hong, NIE Yu. Information retrieval using label propagation based ranking[C]//Proc of the 6th NTCIR Workshop. 2007:140-144.

[6] KIM S M, PANTEL P, DUAN Lei, *et al*. Improving Web page classification by label propagation over click graphs[C]//Proc of the 18th ACM Conference on Information and Knowledge Management. New York:ACM Press,2009:1077-1086.

[7] 贺松林,张晖. 基于 K-means 和 label propagation 的半监督网页分类[J]. 软件导刊,2011,10(2):49-51.

[8] BLAIR-GOLDENSOHN, HANNAN K, McDONALD R, *et al*. Building a sentiment summarizer for local service reviews [EB/OL]: (2008-04-22)[2012-05-22]. http://www.dejanseo.com.au/research/google134368.pdf.

[9] RAO D, RAVICHANDRAN D. Semi-supervised polarity lexicon induction[C]//Proc of the 12th Conference of the European Chapter of the ACL. 2009: 675-682.

[10] SPERIOSU M, SUDAN N, UPADHYAY S, *et al*. Twitter polarity classification with label propagation over lexical links and the follower graph[C]//Proc of the 1st Workshop on Unsupervised Learning in NLP. 2011: 53-63.

[11] 任晓娟. 基于改进标注传播算法的半监督资源分类[D]. 吉林:吉林大学,2008.

[12] 郝建柏,陈贤富,黄双福,等. 一种基于模糊近邻标签传递的半监督分类算法[J]. 微电子学与计算机, 2010,27(2):30-33.

[13] NIU Zheng-yu, JI Dong-hong, TAN C L. Word sense disambiguation using label propagation based semi-supervised learning[C]//Proc of the 43rd Annual Meeting on Association for Computational Linguistics. 2005: 238-241.

[14] LANSDALL-WELFARE T, FLAOUNAS L, CRISTIANINI N. Scalable corpus annotation by graph construction and label propagation[C]//Proc of the 1st International Conference on Pattern Recognition Applications and Methods. 2012: 25-34.

[15] TSAI C F, HUNG C. Automatically annotating images with keywords: a review of image annotation systems[J]. Recent Patents on Computer Science,2008,1(1): 55-68.

[16] BADRINARAYANAN V, GALASSO F, CIPOLLA R. Label propagation in video sequences[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2010:3265-3272.

[17] BUDVYTIS I, BADRINARAYANAN V, CIPOLLA R. Label propagation in complex video sequences using semisupervised learning [EB/OL]. [2012-03-15]. http://mi.eng.cam.ac.uk/~cipolla/pub lications/inproceedings/2010-BMVC -label-propagation.pdf.

[18] 倪煜,赵耀,朱振峰. 结合标签传递的镜头边界检测与分类[J]. 中国图象图形学报,2011,16(6):995-1001.

[19] TANG Jin-hui, HUA Xian-sheng, QI Guo-jun, *et al*. Video annotation based on kernel linear neighborhood propagation[J]. IEEE Trans on Multimedia,2008,10(4):620-628.

[20] ISMAIL M. Image annotation and retrieval based on multi-modal feature clustering and similarity propagation[D]. Louisville:University of Louisville,2011.

[21] ZHONG Sheng-hua, LIU Yan, LIU Yang, *et al*. Region level annotation by fuzzy based contextual cueing label propagation[J]. Multimedia Tools and Applications,2012(1):1-23.

[22] LIU Xiao-bai, CHENG Bin, TANG Jin-hui, *et al*. Label to region by bi-layer sparsity priors[C]//Proc of the 17th ACM International Conference on Multimedia. New York:ACM Press,2009:115-124.

[23] WANG Fei, WANG Xin, LI Tao. Efficient label propagation for interactive image segmentation[C]//Proc of the 6th International Conference on Machine Learning and Applications. 2007:136-141.

[24] YANG Y H, BOGDANOV D, HERRERA P, *et al*. Music retagging using label propagation and robust principal component analysis [C]//Proc of the 21st International Conference on Companion World Wide Web. New York:ACM Press,2012:869-876.

[25] 胡健,董跃华,杨炳儒. 大型复杂网络中的社区结构发现算法[J]. 计算机工程, 2008,34(19):92-93,100.

[26] SYMEON P, YIANNIS K, ATHENA V, *et al*. Community detection in social media performance and application considerations[J]. Data Mining and Knowledge Discovery,2012,24(3):515-554.

[27] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E,2007,76(3):1-12.

[28] BARBER M J, CLARK J W. Detecting network communities by propagating labels under constraint[J]. Physical Review E, 2009, 80(2):129-139.

[29] LIU Xin, MURATA T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks [EB/OL]. (2010-03-19)[2012-05-01]. http://arxiv.org/pdf/0910.1154.pdf.

[30] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics,2010(12):1-26.

[31] 金弟,刘杰,杨博,等. 局部搜索与遗传算法结合的大规模复杂网络社区探测[J]. 自动化学报,2011,37(7):873-882.

[32] XIE Jie-rui, SZYMANSKI B K. Community detection using a neighborhood strength driven label propagation algorithm[C]//Proc of IEEE Network Science Workshop. 2011:188-195.

[33] CORDASCO G, GARGANO L. Community detection via semi-synchronous label propagation algorithms [EB/OL]. (2011-04-23)[2012-03-25]. http://arxiv.org/abs/1103.45-50.

[34] LEUNG X Y, HUI Pan, LIO P, *et al*. Towards real-time community detection in large networks[EB/OL]. (2009-06-22)[2012-03-18]. http://www.social-nets.eu/publications/phyE.pdf.

[35] 朱学芳. 图博档信息资源数字化建设及服务融合探讨[J]. 情报资料工作, 2011(5):57-60.