doi:10.3969/j.issn.1000-436x.2014.08.006

基于查询—文档异构信息网络的半监督学习

刘钰峰[1]，李仁发[1,2]
(1. 湖南大学 信息科学与工程学院，湖南 长沙 410082；2. 湖南大学 嵌入式系统与网络实验室，湖南 长沙 410082)

摘 要：基于图的半监督学习近年来得到了广泛的研究，然而，现有的半监督学习算法大都只能应用于同构网络。根据查询及文档自身的内容特征和点击关系构建查询—文档异构信息网络，并引入样本的判别信息强化网络结构。提出了查询—文档异构信息网络上半监督聚类的正则化框架和迭代算法，在正则化框架中，基于流形假设构造了异构信息网络上的代价函数，并得到该函数的封闭解，以此预测未标记查询和文档的类别标记。在大规模商业搜索引擎查询日志上的实验表明本方法优于传统的半监督学习方法。

关键词：异构信息网络；半监督学习；信息检索；点击日志
中图分类号：TP391 **文献标识码**：A **文章编号**：1000-436X(2014)08-0040-08

Semi-supervised learning by constructing query-document heterogeneous information network

LIU Yu-feng[1], LI Ren-fa[1,2]
(1. School of Information Science and Engineering, Hunan University, Changsha 410082,China;
2. Embedded System and Networking Laboratory, Hunan University, Changsha 410082,China)

Abstract: Various graph-based algorithms for semi-supervised learning have been proposed in recent literatures. However, although classification on homogeneous networks has been studied for decades, classification on heterogeneous networks has not been explored until recently. The semi-supervised classification problem on query-document heterogeneous information network which incorporate the bipartite graph with the content information from both sides is considered. In order to strengthen the network structure, class information of sample nodes is introduced. A semi-supervised learning algorithm based on two frameworks including the novel graph-based regularization framework and the iterative framework is investigated. In the regularization framework, a new cost function to consider the direct relationship between two entity sets and the content information from both sides which leads to a significant improvement over the baseline methods is developed. Experimental results demonstrate that proposed method achieves the best performance with consistent and promising improvements.

Key words: heterogeneous information networks; semi-supervised learning; information retrieval; click-through data

1 引言

现实世界中存在着许多包含大规模数据的信息网络，例如基于链接信息的互联网网页，基于学术论文的引文网络等。从大型信息网络中抽取知识已经吸引了大量研究[1,2]。在一些应用场景中，可以获得少量的标记数据，学习器对标记数据进行学习，借此对数据进行分类有助于研究者发现信息网络中隐藏的结构，并进一步理解不同数据节点的作用，这一方法称为半监督学习[3,4]。

信息检索技术的飞速发展吸引了大量学者对互联网网页进行研究。由于链接信息中存在着大量的噪声链接，人们认为有必要寻找新的可靠的数据源作为相关性评价的依据。近年来，挖掘用户的使用行为改进检索系统的性能逐渐成为研究热点。当把半监督学习算法应用到 Web 日志上时，一个自然的思路是构

收稿日期：2013-05-12；**修回日期**：2013-09-11
基金项目：国家自然科学基金资助项目(61173036)
Foundation Item: The National Natural Science Foundation of China (61173036)

造查询—文档二部图。然而，现有的经典算法如基于图的半监督学习算法、基于半正定规划的半监督学习算法以及半监督支持向量机(SVM, semi-supervised)等只适用于同构节点[5]。这导致只能分别对二部图的一端收集相应的特征信息进行分类，其缺点在于忽略了不同类型节点之间的联系，丢失了一些具有重要区分能力的信息。另一种方案是只考虑查询和文档的点击关系，把查询和文档放在同一个二部图中进行分析，这在本质上把查询和文档视为同一种类型的节点进行分析，这种方式的缺点在于只考虑了成对的约束信息，却忽略了不同节点本身的特征信息。因此，研究者试图融合节点本身的特征信息及异构节点相互之间的联系提升应用的性能。例如，文献[6]把查询和文档这 2 个不同特征空间的节点映射到同一个潜在的特征空间，然后在此空间内计算查询和查询、查询和文档以及文档和文档之间的相似度。文献[7]提出了同时基于异构数据及其成对约束进行非负矩阵分解的半监督聚类的框架(SS-NMF, semi supervised non-negative matrix factorization)。近年来，学术论文数据集(DBLP)引起了诸多研究者的重视，在 DBLP 中存在着作者、论文、主题词以及会议等诸多异构数据。文献[1]分析了星形异构网络上的半监督聚类，并把它应用到 DBLP 数据上；文献[8]建立了异构网络上的主题传播模型 TMBP，该模型能同时利用异构网络及网络节点上的内容信息对作者、论文以及会议进行聚类；文献[9]在此基础上进一步分析异构网络上的排序问题用于专家推荐。

然而，在已知研究中还没有针对查询和文档的内容特征以及点击关系构造异构信息网络以及统一分析进行半监督学习的框架。本文的目标是在给定少量查询和文档类别标记的情况下预测未标记查询和文档的类别。根据查询和文档自身内容特征分别构造基于特征的相似图，同时基于查询和文档之间的点击关系构建查询—文档二部图，并引入样本标记的判别结构信息，同一类别节点之间的关系可以通过查询—文档二部图进行传播并增强。提出了查询—文档异构信息网络上半监督聚类的正则化框架和迭代算法。分析了正则化框架和迭代算法的时间复杂度，结果表明在大规模稀疏网络中本迭代算法更有优势。在 DBLP 数据集上的实验表明本文方法优于传统的半监督学习方法。

2　基于查询—文档异构信息网络的半监督学习

近年来，半监督学习技术大多基于平滑假设，即认为相似样本的标记也应该相似，在此基础上发展的流形假设认为“处于一个很小的局部邻域内的样本具有相似的性质[10]”。本节首先讨论如何基于 Web 日志构造查询—文档异构信息网络，然后基于流形假设讨论在该网络上的正则化框架以及半监督学习迭代算法。

2.1　构建查询—文档异构信息网络

由不同对象构建的异构信息网络可以表示为 $G=(V,E)$，其中，V 可以表示为不同类型的顶点的集合，即 $V=V_1\cup V_2\cup\cdots\cup V_T$，$E$ 为连接顶点的边的集合。如果 $T=1$，则该图称为同构网络，如果 $T>1$，则称为异构网络。

本文中，考虑 $V=Q\cup D$，其中，Q 为查询的集合，D 为文档的集合。$E=E_{QQ}\cup E_{QD}\cup E_{DD}$，其中，$E_{QQ}=Q\times Q$，$E_{QD}=Q\times D$，$E_{DD}=D\times D$。令 $G_Q=(Q,E_{QQ})$，$G_{QD}=(Q,D,E_{QD})$，$G_D=(D,E_{DD})$，则 $G=G_{QQ}\cup G_{QD}\cup G_{DD}$。

在基于信息网络的学习算法中，构造信息网络的核心问题在于如何为信息网络中节点之间的边的权重赋值。通常，可以基于节点之间距离函数进行构造，最为常见的是使用高斯热核进行定义[11]，如式(1)所示。

$$w_{ij}=\begin{cases}\exp(-d(x_i,x_j)/2\sigma^2), x_i\in knn(x_j)\text{或}x_j\in knn(x_i)\\ 0,\text{其他}\end{cases} \tag{1}$$

其中，$d(x_i,x_j)$ 为距离函数，σ 为调节参数，当 $\sigma\to\infty$ 时，式(1)退化为二值函数，$x_i\in knn(x_j)$ 表示 x_i 是 x_j 的 k 近邻中的节点，当 $k\to\infty$ 时，可以认为式(1)是一种基于全体节点的策略，这种策略通常能保证图的全连通，因此计算开销较大。当 $k<|V|$ 时，式(1)是一种基于局部几何结构的策略，此时需要选择合适的 k 保证图的连通性。本文中 G_Q 和 G_D 都是同构网络，G_Q 节点之间的距离函数 $d(x_i,x_j)$ 可以根据查询数据不同特点选择。例如，文献[12]采用维基词典和维基百科作为数据源计算查询之间的语义相似性，文献[13,14]根据查询序列中的行为构建

Query-Flow 图计算查询之间的关系，文献[15]采用 2 个查询字符串之间的编辑距离度量查询之间的相似性。G_D 上节点之间的距离函数 $d(x_i,x_j)$ 在本文中使用 2 个文档在高维空间的距离表示。

G_{QD} 为异构网络，E_{QD} 上的权重函数 $w_{QD}:Q\times D\to R^+$ 用于描述查询词和文档之间的关系，在此使用点击关系进行描述。

$$w_{QD}(i,j)=\frac{cf(q_i,d_j)}{cf(q_i)} \tag{2}$$

其中，$cf(q_i,d_j)$ 表示通过查询 q_i 点击文档 d_j 的次数，而 $cf(q_i)$ 表示通过查询 q_i 总的点击次数。

在半监督学习中，可以根据预先标记样本的判别信息修改原有的信息网络的结构，这种思想在局部敏感分析(LSDA)和边界 Fisher 分析(MFA)中得到了应用，它们同时采用流形局部结构和判别结构进行降维[16,17]。借鉴这种思想，本文提出了以下面向半监督学习的查询—文档异构信息网络构造算法。

算法 1 结合判别信息的查询—文档异构信息网络构建算法

1) 根据式(1)构造 G_Q 和 G_D，根据式(2)构造 G_{QD}。设 G_Q 上所有节点之间边的平均权重为 w_q，设 G_D 上所有节点之间边的平均权重为 w_d。

2) 设查询标记数据被划分为 c 个类，表示为 $P_q=\{P_q^1,P_q^2,\cdots,P_q^c\}$，其中，$P_q{}^i$ 表示第 i 个类别的标记查询的核心集。令 M_q^i 表示第 i 个查询类别标记的成对约束集，若 $x\in P_q^i$ 且 $y\in P_q^i$，则把 (x,y) 加入 M_q^i。设文档标记数据被划分为 c 个类，表示为 $P_d=\{P_d^1,P_d^2,\cdots,P_d^c\}$，其中，$P_d{}^i$ 表示第 i 个类别的标记文档的核心集。令 M_d^i 表示第 i 个文档类别标记的成对约束集，若 $x\in P_d^i$ 且 $y\in P_d^i$，则把 (x,y) 加入 M_d^i。

3) 若样本对 $(q_l,q_m)\in M_q^i$，q_k 为 q_l 和 q_m 的邻居，$w_{lk}>w_q$ 且 $w_{mk}>w_q$，则把 q_k 加入 $P_q{}^i$，把 (q_l,q_k) 和 (q_m,q_k) 加入 M_q^i。

4) 重复 3)，直到 P_q 不再变化。

5) 若样本对 $(d_l,d_m)\in M_d^i$，d_k 为 d_l 和 d_m 的邻居，$w_{lk}>w_d$ 且 $w_{mk}>w_d$，则把 d_k 加入 $P_d{}^i$，把 (d_l,d_k) 和 (d_m,d_k) 加入 M_d^i。

6) 重复 5)，直到 P_d 不再变化。

7) 若查询样本对 $(q_l,q_m)\in M_q^i$，则修改 G_Q 令 $w_{lm}=1$；若 $q_l\in P_q{}^i$ 且 $q_m\in P_q{}^j(i\neq j)$，则修改 G_Q 令 $w_{lm}=0$。若文档样本对 $(d_l,d_m)\in M_d^i$，则修改 G_D 令 $w_{lm}=1$，若 $d_l\in P_d{}^i$ 且 $d_m\in P_d{}^j(i\neq j)$，则修改 G_D 令 $w_{lm}=0$。

8) 若查询 $q_l\in P_q{}^i$ 且文档 $d_m\in P_d{}^i$，则修改 G_{QD} 令 $w_{lm}=1$。

引入判别信息的目的是为了使用样本信息强化现有的网络结构。在基于日志的点击关系中，查询和文档之间非常稀疏，这是导致分类问题性能不佳的主要原因之一，通过算法 1 可以有效地丰富查询和文档之间的关系，同时可以使同类节点之间的联系更为紧凑，不同类别的节点之间的联系更加松散，从而可以更好地进行分类。

在算法 1 中，步骤 3)和步骤 5)的计算复杂度相同。在使用链表存储样本的邻居节点时，寻找样本对的公有邻居即为寻找 2 个链表公有项，在未对链表排序的情况下，时间复杂度为 $O(d_1\times d_2)$，其中，d_1 和 d_2 分别为 2 个样本节点的度，即为链表的长度。在链表已按照边的权重排序的情况下，寻找 2 个链表公有项的时间复杂度为 $O(d_1+d_2)$。在采用式(1)构建基本网络时，节点的度由 k 值决定，其时间复杂可以记为 $O(2k)$。设某一类查询标记中成对约束集的数目使用 $|M_q|$ 表示，类别数目为 c，则步骤 3)的时间复杂度为 $O(2ck|M_q|)$。同理，步骤 5)的时间复杂度为 $O(2ck|M_d|)$，其中，$|M_d|$ 为某一类文档中成对约束集的数目。步骤 7)中修改边权重为 0 的操作是该步骤中最耗时的，其中，寻找 $q_l\in P_q{}^i$ 且 $q_m\in P_q{}^j(i\neq j)$ 的时间复杂度为 $O(|P_q|^2)$，$|P_q|$ 为某一类查询中核心集的数目，由于需要在不同类别的核心集之间交叉搜索，因此，在查询类别上删除修改边权重的时间复杂度为 $O(c^2|P_q|^2)$。由于还要在文档核心集上进行修改，故步骤 7)的时间复杂度为 $O(c^2(|P_q|^2+|P_d|^2))$，其中，$|P_d|$ 为某一类文档类别核心集的数目。同理可知，步骤 8)的时间复杂度为 $O(c^2|P_q||P_d|)$。

2.2 正则化框架

设以上构建的异构信息网络中 $Q=\{q_1,q_2,\cdots,q_l\}$，

$D=\{d_1,d_2,\cdots,d_m\}$，$n=l+m$，聚类的目标是把顶点集划分为 c 个类别。设 $\boldsymbol{F}$ 为 $n\times c$ 矩阵，在此，$\boldsymbol{F}$ 为一个向量函数 $\boldsymbol{F}:V\to R^c$，聚类问题可以描述为对 V 中的顶点 v_i 根据 $y_i=\arg\max_{j\leqslant c}F_{ij}$ 进行类别标记。由于在 $\boldsymbol{F}$ 中包含了查询 Q 和文档 D 的类别标记，因此 $\boldsymbol{F}$ 可以表述为 $\boldsymbol{F}=\left[\boldsymbol{F}_Q,\boldsymbol{F}_D\right]^{\mathrm{T}}$，其中，$\boldsymbol{F}_Q$ 为 $l\times c$ 矩阵，$\boldsymbol{F}_D$ 为 $m\times c$ 矩阵。

根据先验知识，可以由用户对 V 中的若干顶点先进行标记，由此构造 $n\times c$ 的初始化标记矩阵 $\boldsymbol{Y}$，如果 v_i 属于第 j 类，则令 $Y_{ij}=1$，其他情况下都令 $Y_{ij}=0$。同理，$\boldsymbol{Y}$ 可以描述为 $\boldsymbol{Y}=\left[\boldsymbol{Y}_Q,\boldsymbol{Y}_D\right]^{\mathrm{T}}$，其中，$\boldsymbol{Y}_Q$ 为 $l\times c$ 矩阵，$\boldsymbol{Y}_D$ 为 $m\times c$ 矩阵。在图上的半监督聚类问题即为在给定了初始标记矩阵 $\boldsymbol{Y}$ 的情况下求解最优的 $\boldsymbol{F}$。

首先，定义同构图 G_Q 和 G_D 上的代价函数定义如下

$$R_{Q,D}=\frac{\alpha}{2}\sum_{i,j\in Q}\boldsymbol{W}_{QQ,ij}\left\|\frac{\boldsymbol{F}_{Qi}}{\sqrt{\boldsymbol{D}_{QQ,ii}}}-\frac{\boldsymbol{F}_{Qj}}{\sqrt{\boldsymbol{D}_{QQ,jj}}}\right\|^2+\frac{\alpha}{2}\sum_{i,j\in D}\boldsymbol{W}_{PP,ij}\left\|\frac{\boldsymbol{F}_{Pi}}{\sqrt{\boldsymbol{D}_{DD,ii}}}-\frac{\boldsymbol{F}_{Pj}}{\sqrt{\boldsymbol{D}_{DD,jj}}}\right\|^2+\mu\sum_{i\in Q}\left\|\boldsymbol{F}_{Pi}-\boldsymbol{Y}_{Qi}\right\|^2+\mu\sum_{i\in D}\left\|\boldsymbol{F}_{Di}-\boldsymbol{Y}_{Di}\right\|^2 \tag{3}$$

其中，$\boldsymbol{W}_{QQ}$ 表示 G_Q 上的邻接矩阵，$\boldsymbol{D}_{QQ}$ 为对角矩阵，$\boldsymbol{D}_{QQ,ii}=\sum_j\boldsymbol{W}_{QQ,ij}$。$\boldsymbol{W}_{DD}$ 表示 G_D 上的邻接矩阵，$\boldsymbol{D}_{DD}$ 为对角矩阵，$\boldsymbol{D}_{DD,ii}=\sum_j\boldsymbol{W}_{DD,ij}$。式（3）中的第一项和第二项可以认为是进行平滑约束，即认为一个好的代价函数要求相似顶点上的输出值尽量一致，可以把它理解为全局约束。第三项和第四项为拟合约束，即一个好的代价函数要求输出值和初始化的标记值不能偏移太远，可以把它理解为局部约束。全局约束和局部约束由参数 α 和 μ 进行平衡。

对式(3)进行代数变换

$$\begin{aligned}R_{Q,D}&=\alpha\boldsymbol{F}_Q^{\mathrm{T}}\boldsymbol{D}_{QQ}^{-\frac{1}{2}}(\boldsymbol{D}_{QQ}-\boldsymbol{W}_{QQ})\boldsymbol{D}_{QQ}^{-\frac{1}{2}}\boldsymbol{F}_Q+\alpha\boldsymbol{F}_D^{\mathrm{T}}\boldsymbol{D}_{DD}^{-\frac{1}{2}}(\boldsymbol{D}_{DD}-\boldsymbol{W}_{DD})\boldsymbol{D}_{DD}^{-\frac{1}{2}}\boldsymbol{F}_D+\mu\sum_{i\in Q,D}\left\|\boldsymbol{F}_i-\boldsymbol{Y}_i\right\|^2\\&=\alpha\boldsymbol{F}_Q^{\mathrm{T}}(\boldsymbol{I}_Q-\boldsymbol{S}_{QQ})\boldsymbol{F}_Q+\alpha\boldsymbol{F}_D^{\mathrm{T}}(I_D-\boldsymbol{S}_{DD})\boldsymbol{F}_D+\mu\sum_{i\in Q,D}\left\|\boldsymbol{F}_i-\boldsymbol{Y}_i\right\|^2\\&=\alpha\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}\boldsymbol{I}_Q&0\\0&\boldsymbol{I}_D\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\\boldsymbol{F}_D\end{bmatrix}-\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\cdot\begin{bmatrix}\alpha\boldsymbol{S}_{QQ}&0\\0&\alpha\boldsymbol{S}_{DD}\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\\boldsymbol{F}_D\end{bmatrix}+\mu\sum_{i\in Q,D}\left\|\boldsymbol{F}_i-\boldsymbol{Y}_i\right\|^2\end{aligned} \tag{4}$$

其中，$\boldsymbol{S}_{QQ}=\boldsymbol{D}_{QQ}^{-\frac{1}{2}}\boldsymbol{W}_{QQ}\boldsymbol{D}_{QQ}^{-\frac{1}{2}}$，$\boldsymbol{S}_{DD}=\boldsymbol{D}_{DD}^{-\frac{1}{2}}\boldsymbol{W}_{DD}\boldsymbol{D}_{DD}^{-\frac{1}{2}}$，$\boldsymbol{I}_Q$ 为 $l\times l$ 的单位矩阵，$\boldsymbol{I}_D$ 为 $m\times m$ 的单位矩阵。

查询和查询之间的关系以及文档和文档之间的关系可以通过 G_{QD} 互相传播并相互增强，为此定义 G_{QD} 上的代价函数如下

$$R_{QD}=\frac{\beta}{2}\sum_{i\in Q,j\in D}\boldsymbol{W}_{QD,ij}\left\|\frac{\boldsymbol{F}_{Qj}}{\sqrt{\boldsymbol{D}_{QD,ii}}}-\frac{\boldsymbol{F}_{Di}}{\sqrt{\boldsymbol{D}_{DQ,jj}}}\right\|^2+\frac{\beta}{2}\sum_{i\in Q,j\in D}\boldsymbol{W}_{DQ,ij}\left\|\frac{\boldsymbol{F}_{Di}}{\sqrt{\boldsymbol{D}_{DQ,jj}}}-\frac{\boldsymbol{F}_{Qj}}{\sqrt{\boldsymbol{D}_{QD,ii}}}\right\|^2 \tag{5}$$

其中，$\boldsymbol{W}_{QD}$ 为 $\boldsymbol{G}_{QD}$ 上的邻接矩阵，$\boldsymbol{W}_{DQ}$ 为 $\boldsymbol{W}_{QD}$ 的转置矩阵。$\boldsymbol{D}_{QD}$ 和 $\boldsymbol{D}_{DQ}$ 都为对角矩阵，$\boldsymbol{D}_{QD,ii}=\sum_j\boldsymbol{W}_{QD,ij}$，$\boldsymbol{D}_{DQ,ii}=\sum_j\boldsymbol{W}_{DQ,ij}$。参数 β 和式(3)中的参数 α 满足 $\alpha+\beta=1$，且都介于 0 和 1 之间。对式(5)进行代数变换可得

$$\begin{aligned}R_{QD}&=\frac{\beta}{2}\left(\boldsymbol{F}_Q^{\mathrm{T}}\boldsymbol{F}_Q-2\boldsymbol{F}_Q^{\mathrm{T}}\boldsymbol{S}_{QD}\boldsymbol{F}_D+\boldsymbol{F}_D^{\mathrm{T}}\boldsymbol{F}_D\right)+\frac{\beta}{2}\left(\boldsymbol{F}_D^{\mathrm{T}}\boldsymbol{F}_D-2\boldsymbol{F}_D^{\mathrm{T}}\boldsymbol{S}_{DQ}\boldsymbol{F}_Q+\boldsymbol{F}_Q^{\mathrm{T}}\boldsymbol{F}_Q\right)\\&=\beta\left(\boldsymbol{F}_Q^{\mathrm{T}}\boldsymbol{F}_Q-\boldsymbol{F}_Q^{\mathrm{T}}\boldsymbol{S}_{QD}\boldsymbol{F}_D-\boldsymbol{F}_D^{\mathrm{T}}S_{DQ}\boldsymbol{F}_Q+\boldsymbol{F}_D^{\mathrm{T}}\boldsymbol{F}_D\right)\\&=\beta\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}\boldsymbol{I}_Q&0\\0&\boldsymbol{I}_D\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\\boldsymbol{F}_D\end{bmatrix}-\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}0&\beta\boldsymbol{S}_{QD}\\\beta\boldsymbol{S}_{DQ}&0\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\\boldsymbol{F}_D\end{bmatrix}\end{aligned} \tag{6}$$

其中，$\boldsymbol{S}_{QD}=\boldsymbol{D}_{QD}^{-\frac{1}{2}}\boldsymbol{W}_{QD}\boldsymbol{D}_{DQ}^{-\frac{1}{2}}$，$\boldsymbol{S}_{DQ}=\boldsymbol{D}_{DQ}^{-\frac{1}{2}}\boldsymbol{W}_{DQ}\boldsymbol{D}_{QD}^{-\frac{1}{2}}$。

综合式(4)和式(6)，在图 G 上的代价函数可以表示为 $R=R_{Q,D}+R_{QD}$，有

$$R=\alpha\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}\boldsymbol{I}_Q&0\\0&\boldsymbol{I}_D\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\\boldsymbol{F}_D\end{bmatrix}-\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\cdot\begin{bmatrix}\alpha\boldsymbol{S}_{QQ}&0\\0&\alpha\boldsymbol{S}_{DD}\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\\boldsymbol{F}_P\end{bmatrix}+\mu\sum_{i\in Q,D}\left\|\boldsymbol{F}_i-\boldsymbol{Y}_i\right\|^2+$$

$$\beta\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}\boldsymbol{I}_Q & 0\\ 0 & \boldsymbol{I}_D\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\ \boldsymbol{F}_D\end{bmatrix}-$$

$$\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}0 & \beta\boldsymbol{S}_{QD}\\ \beta\boldsymbol{S}_{DQ} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\ \boldsymbol{F}_D\end{bmatrix}$$

$$=(\alpha+\beta)\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}\boldsymbol{I}_Q & 0\\ 0 & \boldsymbol{I}_D\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\ \boldsymbol{F}_D\end{bmatrix}-$$

$$\left[\boldsymbol{F}_Q^{\mathrm{T}},\boldsymbol{F}_D^{\mathrm{T}}\right]\begin{bmatrix}\alpha\boldsymbol{S}_{QQ} & \beta\boldsymbol{S}_{QD}\\ \beta\boldsymbol{S}_{DQ} & \alpha\boldsymbol{S}_{DD}\end{bmatrix}\begin{bmatrix}\boldsymbol{F}_Q\\ \boldsymbol{F}_D\end{bmatrix}+\mu\sum_{i\in Q,D}\left\|\boldsymbol{F}_i-\boldsymbol{Y}_i\right\|^2$$

$$=\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\ -\boldsymbol{F}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{F}+\mu\sum_{i\in Q,D}\left\|\boldsymbol{F}_i-\boldsymbol{Y}_i\right\|^2 \tag{7}$$

其中，$\boldsymbol{S}=\begin{bmatrix}\alpha\boldsymbol{S}_{QQ} & \beta\boldsymbol{S}_{QD}\\ \beta\boldsymbol{S}_{DQ} & \alpha\boldsymbol{S}_{DD}\end{bmatrix}$，参数$\alpha$和$\beta$分别用户控制不同的子图在整个正则化框架中的作用，参数值越大则最终的结果越依赖于该子图。

在给定了初始标记矩阵$\boldsymbol{Y}$的情况下求解最优的$\boldsymbol{F}$可以表达为

$$\boldsymbol{F}^*=\arg\min R(\boldsymbol{F}) \tag{8}$$

根据式(6)对R进行微分并令其等于0[4,18]，有

$$\frac{\partial R}{\partial \boldsymbol{F}}\Big|_{\boldsymbol{F}=\boldsymbol{F}^*}=\boldsymbol{F}^*-\boldsymbol{S}\boldsymbol{F}^*+\mu(\boldsymbol{F}^*-\boldsymbol{Y})=0 \tag{9}$$

式(9)可以进一步转换为

$$\boldsymbol{F}^*-\frac{1}{1+\mu}\boldsymbol{S}\boldsymbol{F}^*-\frac{\mu}{1+\mu}\boldsymbol{Y}=0 \tag{10}$$

引入新的变量$\mu_\alpha=\dfrac{1}{1+\mu}$，$\mu_\beta=\dfrac{\mu}{1+\mu}$，变换后可得

$$\boldsymbol{F}^*=\mu_\beta(\boldsymbol{I}-\mu_\alpha\boldsymbol{S})^{-1}\boldsymbol{Y} \tag{11}$$

其中，$\boldsymbol{I}$为$n\times n$单位矩阵。式(11)即为异构信息网络上正则化框架的封闭解，对节点i可以根据$\arg\max_{j\leqslant c}\boldsymbol{F}_{ij}^*$进行类别标记。

2.3 迭代算法

根据式(11)进行求解需要计算$\boldsymbol{I}-\mu_\alpha\boldsymbol{S}$的逆矩阵，在矩阵的规模很大的情况下其时间复杂度太高，因此，给出以下的迭代算法。

算法 2 查询—文档异构信息网络上的半监督学习迭代算法

1) 对于类别$\forall j\in\{1,\cdots,c\}$，对于图中的节点$\forall i\in\{1,\cdots,n\}$，构造$n\times c$的初始化标记矩阵$\boldsymbol{Y}$。

2) 根据同构节点之间的相似性度量分别构造同构网络上的$\boldsymbol{W}_{QQ}$，$\boldsymbol{W}_{DD}$，根据异构节点之间的关系构造$\boldsymbol{W}_{QD}$及相应的转置矩阵$\boldsymbol{W}_{DQ}$。

3) 构造$\boldsymbol{W}_{QQ}$、$\boldsymbol{W}_{DD}$、$\boldsymbol{W}_{QD}$、$\boldsymbol{W}_{DQ}$的对角度矩阵$\boldsymbol{D}_{QQ,ii}=\sum_j\boldsymbol{W}_{QQ,ij}$，$\boldsymbol{D}_{DD,ii}=\sum_j\boldsymbol{W}_{DD,ij}$，$\boldsymbol{D}_{QD,ii}=\sum_j\boldsymbol{W}_{QD,ij}$，$\boldsymbol{D}_{DQ,ii}=\sum_j\boldsymbol{W}_{DQ,ij}$。

4) 构造矩阵$\boldsymbol{S}=\begin{bmatrix}\alpha\boldsymbol{S}_{QQ} & \beta\boldsymbol{S}_{QD}\\ \beta\boldsymbol{S}_{DQ} & \alpha\boldsymbol{S}_{DD}\end{bmatrix}$。

5) 迭代计算$\boldsymbol{F}(t+1)=\mu_\alpha\boldsymbol{S}\boldsymbol{F}(t)+(1-\mu_\alpha)\boldsymbol{Y}$，其中$\mu_\alpha$为介于(0,1)之间的参数。

6) 设$\boldsymbol{F}^*$为$\{\boldsymbol{F}(t)\}$序列的极限，则图G中的节点v_i根据$y_i=\arg\max_{j\leqslant c}F_{ij}$进行类别标记。

和文献[4]的分析类似，可以证明迭代过程中$\boldsymbol{F}(t)$的极限即为式(11)中得到的封闭解。在迭代过程中，异构信息网络上的每个节点都不断地把标记信息传播给自己的邻居节点直到它们达到一个稳定的状态。

在本算法中，步骤5)是整个算法中最为耗费时间的步骤。由于$\boldsymbol{S}$为稀疏矩阵，因此采用邻接链表作为存储结构。当$\boldsymbol{S}$中的某一行与$\boldsymbol{F}(t)$中的某一列相乘时，只需要使用该行中的非0项参与计算，而$\boldsymbol{S}$中所有行中非0项之和为异构网络中边的数目$|E|$，$\boldsymbol{F}(t)$中列的数目为类别数c，因此计算$\boldsymbol{S}\boldsymbol{F}(t)$的时间复杂度为$O(c|E|)$。与初始标记矩阵进行合并过程即为2个$|V|$行$c$列的矩阵相加，其时间复杂度为$O(c|V|)$，$|V|$为异构网络中的节点数。设迭代次数为$N$，则整个算法的时间复杂度为$O(N(c|E|+|V|))$。在封闭解中需要对矩阵求逆，在不同的算法下其时间复杂度不同，文献[19]证明了其下界为$O(|V|^2\log|V|)$。在大规模稀疏异构网络中$|E|<<|V|^2$，因此迭代算法更有优势。

3 实验与分析

3.1 数据集

为评价本文模型，数据集中不仅要包含文本和类别数据，同时需要在这些文本及类别上的点击数据，然而，当前并没有这样的标准测试集，因此，必须手工构建。采用的原始数据集是来自于搜狗搜索引擎2008年6月的点击记录，在每条点击记录中包含了查询词、用户ID、点击的URL、该URL在返回结果中的排名等信息。在该数据中共包含了

51 043 933 条日志记录，5 736 696 个不同的查询，15 951 082 个不同的文档。为了清除记录中的噪声数据，获得不同类别的点击信息，利用搜狗提供的基于 URL 的分类目录进行处理，得到 5 个类别的数据如表 1 所示。

表 1　　实验数据的类别统计信息

类别	旅游	体育	军事	财经	IT
文档	12 004	28 772	12 707	33 381	15 171
查询	3 823	7 719	4 237	9 066	4 487
平均点击数	11.13	12.73	9.96	14.68	12.38

3.2　性能评价

由于当前经典的半监督学习算法都只能应用于同构网络，因此，把经典的半监督学习算法用于异构网络时主要采用 2 种方案：一是从异构信息网络中抽取同构子网进行分析，例如，在本文中只抽取查询子图 G_Q 或只抽取文档子图 G_D 进行分析；二是忽略异构数据的类型信息，只考虑异构节点之间的关系，即只抽取查询和文档之间的点击关系 G_{QD} 进行分析。在本文中选择经典的基于图的半监督聚类算法(LLGC)[4]采用以上方法与本文算法进行比较。本文算法称为 QDGC。在 LLGC 中只有参数 α 需要设置，设为 0.5。在 QDGC 中，平等地对待各个子网，$\alpha=\beta=1$，$\mu_\alpha=\mu_\beta=0.5$。

使用准确率(*AC*)和归一化的互信息(*NMI*)评价聚类的性能，*AC* 被定义为 $AC=\dfrac{\sum_{i=1}^{c}a_i}{n}$，$c$ 为类别数目，a_i 表示正确分类到第 i 类的样本数，n 为样本总数。*NMI* 被定义为[20]

$$NMI=\frac{\sum_{l=1}^{c}\sum_{h=1}^{c}n_{l,h}\log\left(\frac{nn_{l,h}}{n_l\hat{n}_h}\right)}{\sqrt{\left(\sum_{l=1}^{c}n_l\log\frac{n_l}{n}\right)\left(\sum_{h=1}^{c}\hat{n}_h\log\frac{\hat{n}_h}{n}\right)}} \tag{12}$$

其中，n_l 是第 l 个类别中的样本数，$\hat{n}_h$ 是聚类结果中第 h 个类别中的样本数，$n_{l,h}$ 表示被同时被包含在标准结果类别 C_l 中以及聚类结果类别 $\hat{C}_h$ 中的样本数。当聚类结果与标准结果越符合，*NMI* 值越大，当两者完全重合时 *NMI*=1。

在实验中，首先只对一种类型的数据进行标记。首先，随机地从每种类型的查询中抽取 1%、2%、3%、4%、5%的查询数据，然后使用 LLGC 分别在 G_Q 和 G_{QD} 上计算查询的类别，使用 QDGC 在异构信息网络上计算查询的类别，在每个样本比例上进行了 5 次实验取平均值，其结果如表 2 所示。在表 3 中采用了类似的方法，只对文档进行标记，然后计算文档分类的准确率。

表 2　　查询分类性能比较

样本百分比/%	LLGC(G_Q)		LLGC(G_{QD})		QDGC	
	AC/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%
1	62.2	41.7	52.7	31.5	73.6	59.8
2	65.4	42.7	53.4	32.3	75.3	61.7
3	64.7	413.8	53.5	32.7	74.1	60.2
4	67.3	45.0	58.0	37.4	76.6	63.4
5	68.9	47.5	62.3	44.0	77.8	65.2

表 3　　文档分类性能比较

样本百分比%	LLGC(G_Q)		LLGC(G_{QD})		QDGC	
	AC/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%
1	77.5	62.1	56.3	35.5	78.6	65.2
2	78.2	64.7	59.2	40.5	80.7	67.9
3	81.1	69.0	59.9	40.2	81.0	69.1
4	81.9	70.2	62.1	45.8	83.2	73.6
5	82.2	72.1	63.4	46.0	83.1	74.0

从表 2 可知，QDGC 在查询分类上的准确率提高了很多，但在文档分类的准确率上与 LLGC(G_Q) 相比提高不明显，分析可能是由于文档之间的内容信息已经提供了较为丰富的分类特征，因此性能提升不明显。在只考虑点击关系 G_{QD} 的情况下，无论是查询还是文档的分类效果都远逊于其他方法，其原因就在于点击关系较为稀疏，标记的数据有可能位于不连通的子图，此时会大为影响聚类的性能。

实际应用中，用户可能同时对查询和文档具有先验知识，可以同时对查询和文档进行标记。为研究在标记非常稀疏的情况下算法的性能，同时随机选取一定比例的查询和文档，使用 $(q\%,d\%)$ 表示标记查询和文档的比例，分别选取 $(q\%,d\%)=[(0.1\%,0.1\%),(0.2\%,0.2\%),\cdots,(0.5\%,0.5\%)]$，在每个样本比例上进行了 5 次实验然后取平均值，其结果分别如表 4 和表 5 所示。在此使用了 SS-NMF 作为对比算法，SS-NMF 是一种基于异构数据及成对约束进行非负矩阵分解的半监督聚类的迭代框架，在已知的文献中，它是性能最优的一种算法[7]。QDGC($G1$)表示在基于算法 1 构造的异构信息网络上运行算法 2，

表 4 标记稀疏情况下查询分类性能比较

样本百分比(q%, d%)	LLGC(G_D)		LLGC(G_{QD})		SS-NMF		QDGC		QDGC($G1$)	
	AC/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%
(0.1%，0.1%)	35.7	22.5	42.4	27.1	78.3	63.2	64.9	47.4	78.6	63.8
(0.2%，0.2%)	45.0	28.3	47.8	29.4	80.1	65.3	69.2	55.8	81.2	66.2
(0.3%，0.3%)	49.7	31.8	50.1	30.8	84.2	71.1	73.0	61.3	83.4	70.2
(0.4%，0.4%)	54.5	33.2	53.6	32.9	83.0	72.1	74.5	62.6	84.5	72.9
(0.5%，0.5%)	56.2	36.7	58.7	35.6	83.5	73.8	78.2	66.1	84.7	74.5

表 5 标记稀疏情况下文档分类性能比较

样本百分比(q%, d%)	LLGC(G_D)		LLGC(G_{QD})		SS-NMF		QDGC		QDGC($G1$)	
	AC/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%	*AC*/%	*NMI*/%
(0.1%，0.1%)	63.3	42.7	49.0	28.3	79.3	67.5	72.2	60.8	81.1	68.8
(0.2%，0.2%)	69.0	53.5	53.1	32.6	84.4	74.0	75.5	63.6	85.7	74.2
(0.3%，0.3%)	72.7	58.3	54.8	33.0	85.2	74.1	78.4	66.3	86.0	76.0
(0.4%，0.4%)	72.0	58.0	57.2	35.4	85.3	75.7	80.2	69.7	84.6	75.2
(0.5%，0.5%)	73.2	60.6	59.8	41.2	85.9	76.2	83.8	72.2	87.1	77.5

而 QDGC 表示在没有使用算法 1 的网络上运行算法 2。

由于 LLGC(G_Q)无法使用预先给出的文档标记，而 LLGC(G_D)无法使用预先给出的查询标记，因此准确率不高。LLGC(G_{QD})虽然能同时利用查询和文档标记，但是由于它只利用了查询和文档之间的点击关系，而在实际中该数据往往也较为稀疏，有可能标记的数据位于一个独立的子网中，因此效果不佳。在 QDGC($G1$)中，在标记稀疏的情况下显然优势更为明显，其原因就在于它能更充分地利用查询和文档本身的内容信息，并借助于相互之间的关系互相传播。因此，它的效果远胜于 LLGC。

从表中可知，本文方法和 SS-NMF 的效果较为接近。但是，由于 SS-NMF 算法中不仅要求用户指定同一个类别的样本标记(must-link)，同时还要求用户显示地指明不能在同一个类别中的样本标记(cannot-link)，这种方法对用户提出了更高的要求。因此，从这一角度而言，本文算法具有自己的优势。另一方面，本文算法 1 中的思想和 SS-NMF 也有类似之处，算法 1 会根据预标记样本的判别信息修改图结构，使同类样本之间的联系更为紧凑，不同类别的样本之间的联系更加松散，从 QDGC($G1$)和 QDGC 的对比结果可以发现在标记样本越少的情况下，算法 1 的作用越明显。

4 结束语

本文根据查询和文档自身内容特征和点击关系构建查询—文档异构信息网络，并引入样本标记的判别信息强化网络结构，并提出了查询—文档异构信息网络上半监督聚类的正则化框架和迭代算法。在大规模商业搜索引擎查询日志上的实验表明本文方法优于传统的半监督学习方法。尤其是在给出少量标签的情况下，本文方法能更充分的利用查询和文档本身的内容信息，并借助于相互之间的关系互相传播，从而大大提升聚类的性能。

本文的半监督聚类框架是建立在随机选择标记样本的基础上，然而，初始标记样本的选择对分类的性能影响较大，因此，下一步的研究工作将针对如何自动选择更有代表能力的样本进行标记的问题。

参考文献：

[1] SUN Y, YU Y, HAN J. Ranking-based clustering of heterogeneous information networks with star network schema[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining[C]. Paris, France, 2009. 797-806.

[2] SUN Y, HAN J. Mining heterogeneous information networks: a structural analysis approach[J]. SIGKDD Explorations, 2012, 14(2):20-28.

[3] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. The Journal of Machine Learning Research, 2006, 7: 2399-

2434.

[4] ZHOU D, BOUSQUET O, LAL T N, *et al*. Learning with local and global consistency[J]. Advances in Neural Information Processing Systems, 2004, 16:321-328.

[5] LI X, WANG Y Y, ACERO A. Learning query intent from regularized click graphs[A]. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Singapore, Singapore, 2008. 339-346.

[6] WU W, LI H, XU J. Learning query and document similarities from click-through bipartite graph with metadata[A]. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining[C]. Roman, Italy, 2013.687-696.

[7] CHEN Y, WANG L, DONG M. Non-negative matrix factorization for semisupervised heterogeneous data coclustering[J]. Knowledge and Data Engineering, 2010, 22(10): 1459-1474.

[8] DENG H, HAN J, ZHAO B, *et al*. Probabilistic topic models with biased propagation on heterogeneous information networks[A]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. San Diego, CA, 2011. 1271-1279.

[9] DENG H, HAN J, LYU M R, *et al*. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking[A]. Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries[C]. New York, USA, 2012. 71-80.

[10] ZHOU Z H, LI M. Semi-supervised learning by disagreement[J]. Knowledge and Information Systems, 2010, 24(3): 415-439.

[11] JOHNSON R, ZHANG T. Graph-based semi-supervised learning and spectral kernel design[J]. IEEE Transactions on Information Theory, 2008, 54(1): 275-288.

[12] LUCCHESE C, ORLANDO S, PEREGO R, *et al*. Identifying task-based sessions in search engine query logs[A]. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining[C]. Hong Kong ,China, 2011. 277-286.

[13] BOLDI P, BONCHI F, CASTILLO C, *et al*. The query-flow graph: model and applications[A]. Proceedings of the 17th ACM Conference on Information and Knowledge Management[C]. Napa Valley ,USA, 2008. 609-618.

[14] BOLDI P, BONCHI F, CASTILLO C, *et al*. Query suggestions using query-flow graphs[A]. Proceedings of the 2009 Workshop on Web Search Click Data[C]. Barcelona, Spain, 2009. 56-63.

[15] JONES R, KLINKNER K L. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs[A]. Proceedings of the 17th ACM Conference on Information and Knowledge Management[C]. Napa Valley ,USA ,2008. 699-708.

[16] PAN J, KONG F S, WANG R Q. Locality sensitive discriminant transductive learning[J]. Journal of Zhejiang University, Engineering Science, 2012, 46(6): 987-994.

[17] CHEN X, CHEN S, XUE H, *et al*. A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data[J]. Pattern Recognition, 2012, 45(5): 2005-2018.

[18] DENG H, LYU M R, KING I. A generalized Co-HITS algorithm and its application to bipartite graphs[A] Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Paris,France,2009. 239-248.

[19] RAZ R. On the complexity of matrix product[J]. SIAM Journal on Computing, 2003, 32(5): 1356-1369.

[20] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions[J]. The Journal of Machine Learning Research, 2003, 3: 583-617.

作者简介：

刘钰峰（1974-），男，湖南邵阳人，湖南大学博士生，主要研究方向为智能信息检索与机器学习。

李仁发（1956-），男，湖南郴州人，湖南大学教授、博士生导师，主要研究方向为嵌入式计算、网络计算和无线网络。