doi: 10.12068/j.issn.1005-3026.2018.02.006

一种局部与全局特征相结合的主题域识别模型

寇 月, 徐宏斌, 申德荣, 聂铁铮
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 传统的主题域识别技术主要局限于单一领域,缺乏领域间的交互式协同,难以保证识别结果的准确性,因此提出一种局部与全局特征相结合的主题域识别模型. 该模型一方面基于实体在领域内的局部特征进行局部识别,另一方面基于领域间协同作用、领域相关度等全局特征对各个局部识别结果进行一致化趋近,从而使识别结果更全面、有效. 另外,针对相似矩阵的更新时机、协同作用的量化以及迭代终止条件的设定三个方面对主题域识别算法进行了优化. 通过实验验证了本文所提出的关键技术的可行性和有效性.

关 键 词: 主题域识别;交互式协同;局部特征;全局特征;领域相关度
中图分类号: TP 311.13 **文献标志码**: A **文章编号**: 1005-3026(2018)02-0176-05

A Topic Domain Identification Model Combining Local and Global Characteristics

KOU Yue, *XU Hong-bin*, *SHEN De-rong*, *NIE Tie-zheng*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: KOU Yue, E-mail: kouyue@cse.neu.edu.cn)

Abstract: Traditional identification techniques focus on a single domain and lack the mutual collaboration among different domains, which often lead to dumb results. So, a topic domain identification model combining local and global characteristics is proposed. Local identification is performed based on entities' local characteristics within one domain. On the other hand, these local identification results tend to be consistent with each other based on the global characteristics such as the collaboration and relevance among domains, which can maintain the accuracy of identification effectively. In addition, some improvements are made for the algorithm of topic domain identification, including similarity matrix updating, collaboration quantifying and iteration terminating. The experiments demonstrate the feasibility and effectiveness of the proposed model.

Key words: topic domain identification; mutual collaboration; local characteristics; global characteristics; domain relevance

随着互联网技术的创新与发展,以实体(如商品、文章、会议、人等)为中心的 Web 搜索系统已经得到广泛的应用. 这些实体连同关联关系构成了一种网络关系结构,即一个信息网络,例如 DBLP,IMDB 和 Flickr 等. 在信息网络中,某些实体具有紧密的关联关系,表达出相同或相似的主题特征,这些实体构成的集合称为主题域. 主题域识别的目标是把信息网络中的实体划分成若干子集,每个子集由彼此主题相关的实体构成. 主题域识别技术可应用于犯罪团伙检测、社区发现、文档聚类等领域,具有重要的应用价值.

尽管当前的主题域识别技术很多,但它们主要局限于单一领域,缺乏领域间的交互式协同,难以保证识别结果的准确性. 例如,两个信息网络隶属于不同的领域(D1 和 D2),分别表示用户之间的项目合作关系和论文合作关系. 假定某企业想寻求一个开发团队,要求团队成员在理论研究和实践开发两个方面都彼此紧密相关. 传统的面向

收稿日期: 2016-08-19
基金项目: 国家重点基础研究发展计划项目(2012CB316201); 国家自然科学基金资助项目(61472070); 中央高校基本科研业务费专项资金资助项目(130404015).
作者简介: 寇 月(1980-),女,辽宁沈阳人,东北大学副教授,博士; 申德荣(1964-),女,辽宁铁岭人,东北大学教授,博士生导师.

单领域的主题域识别将每个领域看作一个独立体，在 D1 和 D2 中的识别结果都无法全面而准确地满足用户的需求. 原因在于：面向单领域的识别方法仅考虑实体在某领域内部的局部特征，而忽略了领域间交互式协同、领域相关度等全局特征，因此无法全面满足用户的需求.

针对现有主题域识别技术的不足，本文提出了一种局部与全局特征相结合的主题域识别模型（topic domain identification model combining local and global characteristics，LG－TDIM），该模型面向多领域且充分考虑了实体的局部特征以及领域间交互式协同、领域相关度等全局特征. 基于 LG－TDIM 模型，提出了主题域识别算法，并从相似矩阵的更新时机、协同作用的量化以及迭代终止条件的设定三个方面对该算法进行了优化，以进一步提高识别的准确性.

1　相关工作

1.1　基于局部特征的主题域识别

目前，大多数主题域识别技术面向单领域信息网络，考虑的因素包括实体的模式特征、实例特征和图的拓扑结构等局部特征.

实体的模式特征主要包括实体类别、元路径等. 例如，文献[1]通过计算属性名相似性来进行模式聚类，文献[2－3]在关联实体匹配时考虑了元路径信息. 基于实例特征的主题域识别技术的基本思想是依据实例间的相似度来计算实体间主题相关度. 例如，文献[4]提出一种混合子空间聚类算法，考虑了实体共现程度及属性值相似度等实例特征；文献[5]将实体的模式特征与属性特征相结合，提出一种增量式验证算法. 另外，还有一些主题域识别技术侧重于利用图的拓扑结构进行识别. 例如，文献[6]提出一种星型网络模式下的聚类算法，文献[7]提出一种边密度约束的识别方法.

1.2　基于全局特征的主题域识别

目前基于全局特征的主题域识别技术还不多见，但与之相关的一些聚类算法已被提出，包括多视角聚类和多领域聚类.

多视角聚类是从不同视角对实体聚类，最终达到全局层面各个视角下聚类结果的一致化. 例如，文献[8]和文献[9]分别提出了 CCA 算法和 CSC 算法，能够利用其他视角下的聚类结果来动态地更新某视角下的拉普拉斯矩阵；但该类算法一般需要满足一定约束条件，如要求各个信息网络中节点完全相同，并且要求最后各个局部聚类结果也是相同的. 多领域聚类算法可适用于各个信息网络中节点与边都不同的情况. 例如，文献[10]提出了 CGC 算法，基于非负矩阵因式分解实现全局层面的聚类.

本文提出的主题域识别技术与上述工作的不同之处在于：

1）传统的主题域识别方法主要面向单一领域，仅考虑实体在某领域内部的局部特征. 本文提出的主题域识别模型面向多领域，力求将局部特征与全局特征充分结合来提升识别的准确性.

2）虽然一些方法支持多领域的主题域识别，但它们大多要求各领域满足一定的约束限制，且易导致领域间的过度影响. 本文提出的主题域识别模型可适用于各领域节点与边均不同的情况，并可避免出现领域间过度影响的情况.

2　LG－TDIM 模型

定义 1　信息网络. 信息网络用来表示实体之间的关联关系，某领域 i 的信息网络用 $D_i(V, E, \Psi)$ 表示. 其中，V 表示实体节点集合，E 表示节点间关联边集合，Ψ 用来量化实体之间的关联强度.

定义 2　主题域识别. 给定一组信息网络 $\{D_0,\cdots,D_d\}$，主题域识别的目标是生成主题聚类集 $\{C_0,\cdots,C_k\}$，要求隶属于同一主题的实体被划分在同一聚类 C_i 中，每个聚类只包含一个主题.

针对主题域识别，本文提出了 LG－TDIM 模型（如图 1 所示），该模型将主题域识别过程分为两个阶段.

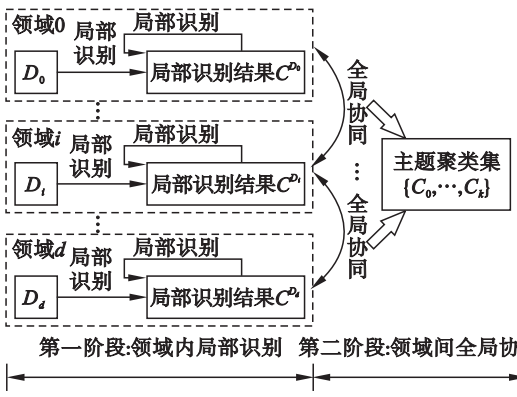

图 1　局部与全局特征相结合的主题域识别模型

Fig. 1　Topic domain identification model combining local and global characteristics

第一阶段：领域内局部识别. 给定 $\{D_0,\cdots,D_d\}$，领域内局部识别目标是对 $D_i(i=0,\cdots,d)$ 生

成一组聚类 $C^{D_i}=\{C_0^{D_i},\cdots,C_m^{D_i}\}$. 该阶段考虑了实体在 D_i 的局部特征(如 D_i 的拓扑结构、领域内实体间关联强度等).

第二阶段:领域间全局协同. 给定各个领域的局部识别结果($C^{D_0},\cdots,C^{D_d}$),领域间全局协同是指利用其他领域 $D_j(i\neq j)$ 对 D_i 的影响,对 D_i 的局部结果进行修正,使其趋于一致化,最终形成全局层面的聚类结果,即 $\{C_0,\cdots,C_k\}$. 该阶段考虑了领域间的交互式协同作用(如 D_i 的识别结果对 D_j 的相似矩阵的影响等).

3 主题域识别算法

3.1 算法描述

LG-TDIM 模型中的第一阶段用来将关联紧密的实体聚在一起而形成局部识别结果. 基于实体间关联强度,可以为各个领域构建一个相似矩阵,它能够反映出实体的局部特征. 在第二阶段,将相似矩阵之间的相互影响作为全局特征,迭代地对其修正,从而使局部识别结果趋于一致化. 针对上述过程,本文提出一种主题域识别算法,具体步骤如下.

步骤 1 针对每个领域 z 的相似矩阵(记为 $\boldsymbol{S}^z=(S_{ij}^z)$)进行初始化. 本文将实体间关联强度 $\boldsymbol{\Psi}=(\psi_{ij})$ 作为相似矩阵中各元素的初始值.

步骤 2 基于谱聚类的思想对每个 D_z 中的节点进行局部聚类. 首先,基于 $\boldsymbol{S}^z$ 构建对角矩阵 $\boldsymbol{T}$ 和拉普拉斯矩阵 $\boldsymbol{L}_{n\times n}=\boldsymbol{T}-\boldsymbol{S}^z$,并构建特征向量空间 $\boldsymbol{Y}$. 然后,基于 k-means 思想对 $\boldsymbol{Y}$ 中的行进行聚类,得到领域 z 的局部识别结果 $C^{D_z}=\{C_0^{D_z},\cdots,C_m^{D_z}\}$.

步骤 3 将各个 C^{D_z} 以局部识别矩阵 $\boldsymbol{M}^z=(M_{ij}^z)$ 表示. 其中,M_{ij}^z 表示在领域 z 中节点 i 与 j 是否在同一个聚类中,如果隶属于同一聚类则为 1,否则为 -1.

步骤 4 若迭代次数已达上限,则返回($C^{D_0},\cdots,C^{D_d}$),算法终止. 否则,根据其他领域 $z'(z'\neq z)$ 的 $\boldsymbol{M}^{z'}$ 来调整 $\boldsymbol{S}^z$,基于调整后的 $\boldsymbol{S}^z$ 对 D_z 重新构建局部识别矩阵 $\boldsymbol{M}^z$,再次执行步骤 4. 注意,当各个领域内实体数量不同时,相似矩阵的大小也不相同. 为简单起见,这里假定同一实体在不同领域中行列序号是一致的. 如果 $M_{ij}^z=M_{ij}^{z'}=1$(或 $M_{ij}^z=M_{ij}^{z'}=-1$),则说明在领域 z 与 z' 中对于节点 i 与 j 的识别结果是趋于一致的,z 与 z' 相互协同,应一同提升(或降低)S_{ij}^z 和 $S_{ij}^{z'}$ 的取值. 如果 $M_{ij}^z\neq M_{ij}^{z'}$,则说明领域 z 与 z' 是相互抵消的,需要对 M_{ij}^z 和 $M_{ij}^{z'}$ 进行综合考虑. 本文将其他领域对 S_{ij}^z 取值的协同作用量化为 $E(S_{ij}^z)$(见式(1)),基于领域间交互式协同作用来调整各个领域的相似矩阵(见式(2)).

$$E(S_{ij}^z)=\begin{cases}\exp\left(\sum_{z'}M_{ij}^{z'}\right), & \text{领域 } z' \text{ 包含节点 } i \text{ 与 } j;\\ 1, & \text{其他.}\end{cases} \tag{1}$$

$$S_{ij}^z=S_{ij}^z\times E(S_{ij}^z)\ . \tag{2}$$

3.2 优化策略

虽然 3.1 节中的算法综合考虑了实体的局部特征和全局特征,但还存在以下问题:

1) 某领域可能会受到其他领域的过度影响,即全局特征被过度强化.

2) 该算法视各个领域是平等的,而实际上它们与领域 z 具有不同的相关性,应区别对待.

3) 该算法的迭代终止条件单纯由迭代次数的上限所决定.

为此,本文从相似矩阵的更新时机、协同作用的量化和迭代终止条件的设定三个方面对 3.1 中的算法进行了优化.

首先,重新设置相似矩阵的更新时机:对于相似矩阵中的 S_{ij}^z,当领域 z' 对 z 有正向(或反向)促进作用时,如果 $S_{ij}^z=\max\{S_{i0}^z,\cdots,S_{i(n-1)}^z\}$(或 $S_{ij}^z=\min\{S_{i0}^z,\cdots,S_{i(n-1)}^z\}$),则放弃本次对 S_{ij}^z 的更新. 这样,节点之间的相似度不会被过度调整.

其次,在量化领域之间协同作用时,如果 z' 与 z 的相关度较高,则对于 z 来说,z' 中的识别结果要比其他领域更具有说服力,因此应加强 z' 对 z 的促进作用,反之亦然. 直观来看,如果两个领域所包含的公共节点越多,它们就越相关. 如果某公共节点在其他领域中比较少见,那么它就更能反映该领域的特性. 结合公共节点的数量及其重要性,定义领域间相关度(式(3)),并重新量化领域间的协同作用(式(4)).

$$P(z,z')=\begin{cases}\dfrac{|D_z.V\cap D_{z'}.V|}{|D_z.V\cup D_{z'}.V|}\times\displaystyle\sum_{V_i\in D_z.V\cap D_{z'}.V}\ln\frac{d+1}{|\{D_j\mid V_i\in D_j.V\}|}, & z\neq z';\\ 1, & \text{其他.}\end{cases} \tag{3}$$

$$E(S_{ij}^z)=\begin{cases}\exp\left(\sum_{z'}P(z,z')\times M_{ij}^{z'}\right), & z' \text{ 包含节点 } i \text{ 与 } j;\\ 1, & \text{其他.}\end{cases} \tag{4}$$

式中：D_z. V 表示领域 D_z 中的节点集合.

第三，针对迭代终止条件的设定，通过比较前后两次迭代所产生的聚类结果的模块度（式(5)）来量化两次迭代之间的差异（式(6)）. 若 ε 趋于1，则结果趋于稳定，迭代即可终止.

$$Q(C^{D_z}) = \sum_{C_i^{D_z} \in C^{D_z}} \frac{\psi_{ii} - \psi_{ij}}{| C_i^{D_z} V |}, \tag{5}$$

$$\varepsilon = \sum_{z=0}^{d} \frac{Q'(C^{D_z})}{Q(C^{D_z})}. \tag{6}$$

式中 ψ_{ii}，ψ_{ij} 分别表示在聚类 i 内部的实体间总关联强度之和与聚类之间的总关联强度之和.

4　实验测试

本文主要针对 3 个数据集（Iris，Wine 和 DBLP）进行测试，其中 Iris 和 Wine 来自 UCI 数据库（http://archive. ics. uci. edu/ml/datasets. html）. Iris 和 Wine 各包含 150 个鸢尾花样本和 178 个葡萄酒样本，每个样本由若干属性描述，每种属性构成一个信息网络. 对于 DBLP，选用了 SIGMOD，VLDB，ICDE，SIGKDD 的 9 628 名作者及 10 175 篇论文，每个会议形成一个信息网络. 本实验环境设置为 Intel Core－i7（3. 4 GHz）处理器及 8 GB 内存，JDK1. 8.

实验 1　比较本文提出的主题域识别方法 LG－TDIM与传统的基于局部特征的主题域识别方法L－TDIM（数据重复率为 0）的识别准确性（NMI 值），如图 2 所示. 当数据重复率为 0 时，表明领域间无交集，此时仅考虑局部特征. 随着数据重复率的增加，全局特征愈发被考虑，识别准确性逐渐提高.

实验 2　将未经优化的主题域识别算法 LG－TDIM与优化后的算法 LG－TDIM′进行比较. 如图 3 所示，在领域间数据重复率较低的情况下，LG－TDIM′算法略优于 LG－TDIM 算法，此时领域间的协同作用不会产生过度影响. 但随着领域间数据重复率的增加，LG－TDIM′算法的优势越发明显.

实验 3　针对不同算法，评估了迭代次数对识别准确性的影响，如图 4 所示. 其中，LG－TDIM1，LG－TDIM2 和 LG－TDIM3 分别代表对相似矩阵的更新时机、协同作用的量化和迭代终止条件的设定进行优化后的模型. 实验结果表明，三项优化措施均是有效的. 随着迭代次数的增加，LG－TDIM′算法的优势越发明显.

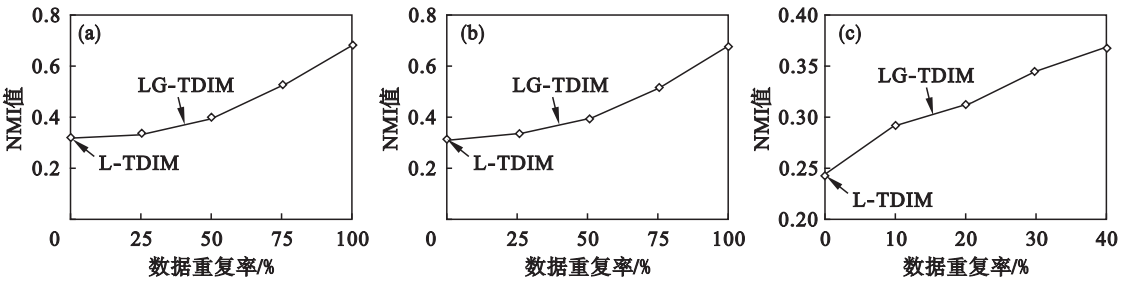

图 2　LG－TDIM 算法与 L－TDIM 算法的识别准确性比较

Fig. 2　Accuracy comparison between LG-TDIM and L-TDIM

（a）—Iris 数据集；（b）—Wine 数据集；（c）—DBLP 数据集.

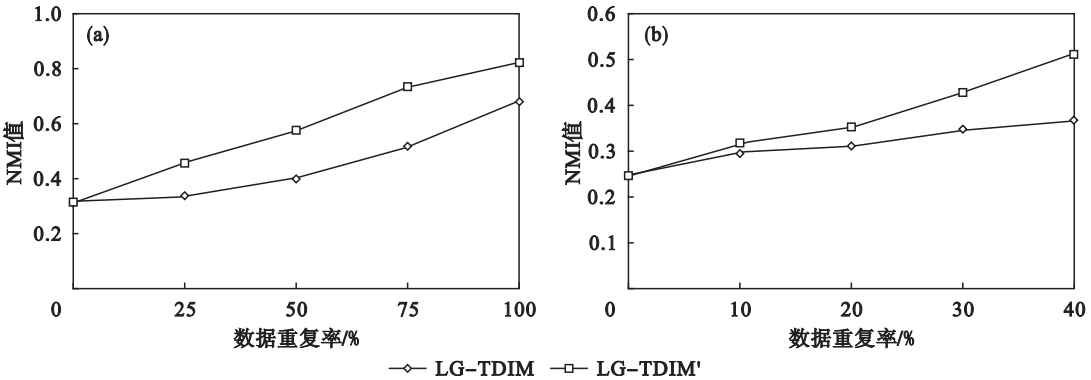

图 3　LG－TDIM′算法与 LG－TDIM 算法的识别准确性比较

Fig. 3　Accuracy comparison between LG-TDIM′ and LG-TDIM

（a）—Wine 数据集；（b）—DBLP 数据集.

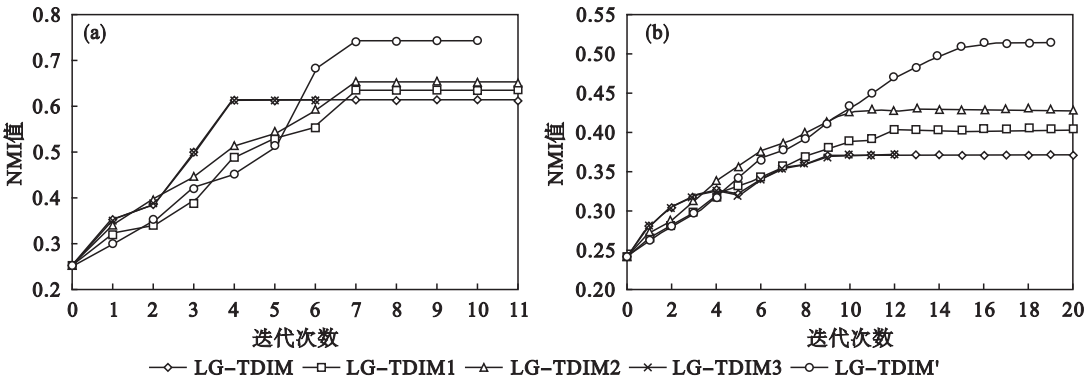

图 4 不同算法的迭代次数对识别准确性的影响

Fig. 4 Effect of iterations on accuracy in different algorithms

(a)—Iris 数据集; (b)—DBLP 数据集.

5 结 语

针对现有主题域识别技术的不足,本文提出了一种局部与全局特征相结合的主题域识别模型,一方面基于实体在领域内的局部特征进行局部识别,另一方面基于领域间协同作用、领域相关度等全局特征对各个局部识别结果进行一致化趋近,从而使得识别结果更为全面、有效. 最后通过实验验证了本文提出的关键技术的可行性和有效性.

下一步工作将对算法的计算复杂度进行分析,提出优化策略来降低其执行代价.

参考文献:

[1] Mahmoud H A, Aboulnaga A. Schema clustering and retrieval for multi-domain pay-as-you-go data integration systems [C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. Indianapolis, 2010: 411 – 422.

[2] Sun Y Z, Han J W, Yan X F, et al. PathSim: meta path-based top-*k* similarity search in heterogeneous information networks [J]. *Proceedings of the VLDB Endowment*, 2011, 4(11): 992 – 1003.

[3] Sun Y Z, Norick B, Han J W, et al. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks [C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, 2012: 1348 – 1356.

[4] Lee J, Hwang S, Nie Z, et al. Query result clustering for object-level search [C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, 2009: 1205 – 1214.

[5] 寇月,申德荣,刘恒,等. 异构网络中关联实体识别模型及增量式验证算法研究[J]. 计算机学报, 2013, 36(10): 2096 – 2108.
(Kou Yue, Shen De-rong, Liu Heng, et al. Research on related entity identification model and incremental verification algorithm for heterogeneous networks[J]. *Chinese Journal of Computers*, 2013, 36(10): 2096 – 2108.)

[6] Gu Y, Gao C, Cong G, et al. Effective and efficient clustering methods for correlated probabilistic graphs [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(5): 1117 – 1130.

[7] Pattillo J, Veremyev A, Butenko S, et al. On the maximum quasi-clique problem [J]. *Discrete Applied Mathematics*, 2013, 161(1/2): 244 – 257.

[8] Chaudhuri K, Kakade S M, Livescu K, et al. Multi-view clustering via canonical correlation analysis [C]// Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, 2009: 129 – 136.

[9] Kumar A, Daumé H. A co-training approach for multi-view spectral clustering[C]//Proceedings of the 28th International Conference on Machine Learning. Bellevue, 2011: 393 – 400.

[10] Cheng W, Zhang X, Guo Z, et al. Flexible and robust co-regularized multi-domain graph clustering[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, 2013: 320 – 328.