软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2734−2746 [doi: 10.3724/SP.J.1001.2013.04480]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

专家证据文档识别无向图模型*

毛存礼, 余正涛, 吴则建, 郭剑毅, 线岩团

(昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

通讯作者: 余正涛, E-mail: ztyu@hotmail.com, http://www.liip.cn

摘 要: 专家证据文档识别是专家检索的关键步骤.融合专家候选文档独立页面特征以及页面之间的关联关系,提出了一个专家证据文档识别无向图模型.该方法首先分析各类专家证据文档中的词、URL 链接、专家元数据等独立页面特征以及候选专家证据文档间的链接和内容等关联关系;然后将独立页面特征以及页面之间的关联关系融入到无向图中构建专家证据文档识别无向图模型;最后利用梯度下降方法学习模型中特征的权重,并利用吉布斯采样方法进行专家证据文档识别.通过对比实验验证了该方法的有效性.实验结果表明,该方法有较好的效果.

关键词: 专家证据文档;专家检索;独立页面特征;专家元数据;无向图模型

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 毛存礼,余正涛,吴则建,郭剑毅,线岩团.专家证据文档识别无向图模型.软件学报,2013,24(11):2734−2746. http://www.jos.org.cn/1000-9825/4480.htm

英文引用格式: Mao CL, Yu ZT, Wu ZJ, Guo JY, Xian YT. Undirected graph model for expert evidence document recognition. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2734−2746 (in Chinese). http://www.jos.org.cn/1000-9825/4480.htm

Undirected Graph Model for Expert Evidence Document Recognition

MAO Cun-Li, YU Zheng-Tao, WU Ze-Jian, GUO Jian-Yi, XIAN Yan-Tuan

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Corresponding author: YU Zheng-Tao, E-mail: ztyu@hotmail.com, http://www.liip.cn

Abstract: Expert evidence document recognition is the key step for expert search. Combining specialist candidate document independent page features and correlation among pages, this paper proposes an expert evidence document recognition method based on undirected graph model. First, independent page features such as words, URL links and expert metadata in all kinds of expert evidence document, and correlations such as links and content among candidate expert evidence document are analyzed. Then, independent page features and correlation among pages are integrated into the undirected graph to construct an undirected graph model for expert evidence document recognition. Finally, feature weights are learned in the model by using the gradient descent method and expert evidence document recognition is achieved by utilizing Gibbs Sampling method. The effectiveness of the proposed method is verified by comparison experiment. The experimental results show that the proposed method has a better effect.

Key words: expert evidence document; expert search; independent page feature; expert metadata; undirected graph model

获取高质量的专家证据文档是提高专家检索效果的重要数据资源,研究专家证据文档识别对专家检索具有重要的意义[1,2].TREC 2005 设立了专家检索子任务,将专家检索任务定义为:给定一个查询主题、一个专家列表及专家对应的证据文档集合,要求给出与查询主题相关的专家排序结果[3].中文专家证据文档分为主页、学术资源 CNKI 页、百科页、博客页这 4 种类型.中文专家证据文档识别的目的就是要自动从大量的互联网页面中找出与查询主题紧密相关的专家主页、学术资源页、百科页、博客页构成相关专家的候选证据文档集合,为进

* 基金项目: 国家自然科学基金(61175068); 教育部留学回国人员启动基金; 云南省教育厅科研基金重大专项; 云南省软件工程重点实验室开放基金(2011SE14)

收稿时间: 2013-05-06; 修改时间: 2013-07-12, 2013-08-02; 定稿时间: 2013-08-27

一步开展专家检索任务提供高质量的数据资源.

目前,专家证据文档识别的方法主要集中在专家主页识别方面.如,Xi 等人结合决策树和逻辑回归预测学习方法进行专家主页识别[4];Tang 在 ArnetMiner 中采用了统计学习方法构建分类器识别专家页面[5];Bron 结合语言模型与维基外部链接启发式规则进行主页识别[6];Fang 等人在 TREC 实体检索任务中采用逻辑回归方式选取 URL 链接特征和内容特征通过机器学习方法建立主页识别模型识别专家主页[7];Li 等人采用以上 4 种学习算法对比了链接 URL 特征和页面内容特征对主页识别的效果[8];Fang 等人利用专家证据文档页面之间的相互依赖关系特征,基于逻辑回归模型提出了一个判别式的主页识别图模型[9];Wu 等人融合主页的独立页面特征以及主页之间的关联关系特征,提出了一种基于 Markov 逻辑网的专家主页识别方法[10].

现有的专家证据文档识别方法大都限定为将每一类文档单独进行识别,很少有研究将专家主页、学术资源 CNKI 页、百科页、博客页这 4 类证据文档作为一个整体进行识别.然而,专家证据页面是一组紧密相关的页面,并且页面之间具有较强的相互关联关系.Macdonald 等人融合了证据文档之间紧密关联的关系,提出了基于投票技术的专家证据文档识别方法,从而为专家检索系统获取到高质量的候选证据文档集合[2,11,12].各类专家证据文档都具有自己的特点,非主页的证据文档通常具有更加明显的独立页面特征,专家主页之间的关联关系大都限制在主页之间的链接关系之中.由于专家证据文档之间包含了丰富的关联关系,综合考虑专家紧密相关的各类证据文档之间的特征关联关系能够有效提高专家检索的效果.为此,本文在 Fang[9]和 Wu[10]处理专家主页识别方法的基础上,基于概率图模型[13,14]的思想,综合考虑专家证据文档的独立页面特征以及文档之间的关联关系,提出针对专家证据文档识别的无向图模型,实验结果表明该方法具有较好的效果.

本文第 1 节针对专家证据文档识别的任务提出专家证据文档识别无向图模型的框架及算法流程.第 2 节分析专家证据文档特征分析方法.第 3 节提出专家证据文档识别无向图模型的表示方法,并详细介绍模型推理方法及特征权重学习方法.第 4 节在收集的两个领域的数据集上通过实验验证所提出方法的有效性.第 5 节对本文进行总结并对未来工作提出展望.

1 专家证据文档识别无向图模型框架

专家证据文档识别无向图模型框架如图 1 所示,包括以下几个部分:

- 查询模块:利用搜索引擎接口获取相关查询专家的初始页面集.
- 专家消歧模块:采用融合页面关联关系的中文专家谱聚类消岐方法[15],对搜索引擎返回的专家文档集进行人名消歧,得到与查询专家一致的候选文档集.
- 特征分析模块:分析专家独立页面特征信息,如专家姓名、组织机构、联系方式、研究方向、页面 URL 链接信息及不同页面之间的关联关系特征,如页面间存在相互链接关联、内容相似等关联特征.
- 专家证据文档及关联特征无向图构建模块:将专家证据文档标记与文档的独立页面特征关联关系构建为独立页面特征团,将专家各类文档之间的链接特征和内容关联特征构建为文档关联特征团,并将这两类团融合在一起构建出专家证据文档无向图.
- 概率推理模块:专家证据文档识别模型中的概率推理的目标是在已知文档的独立特征以及文档之间的依赖关系特征的前提下,采用动态的蒙特卡洛算法中吉布斯采样算法求解候选证据文档的最大可能性标注.
- 参数学习模块:根据已经标注好的证据文档的语料,采用梯度下降法学习出专家证据文档识别无向图模型中独立页面特征和页面相互依赖关系特征所形成团的权重.
- 查询结果模块:从查询候选文档集中输出与查询紧密相关的专家主页、百科页、CNKI 页、博客页这 4 类证据文档.

专家证据文档识别的具体步骤如下:

Step 1. 调用搜索引擎接口查询相关专家,选取排在前 10 的页面文档集.

Step 2. 融合专家页面特征利用谱聚类方法对同名专家进行消岐,得到各个专家对应的候选文档集.

Step 3. 利用特征选取方法,分析文档独立页面特征及文档之间关联特征.

Step 4. 将专家证据文档独立页面特征及关联关系特征融合到无向图中,构建出专家证据文档识别无向图.

Step 5. 利用吉布斯采样方法进行概率推理求解页面标记类型,并利用利用梯度下降方法学习模型中特征权重.

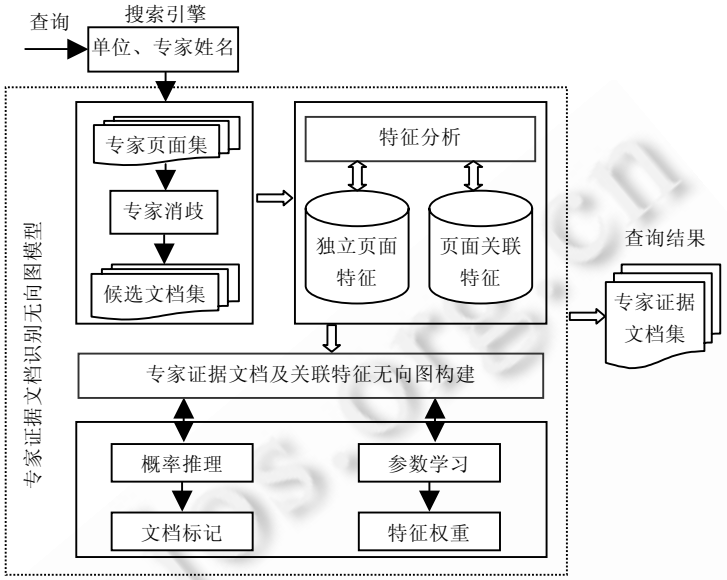

Fig.1 Flow diagram for expert evidence document recognition

图 1 专家证据文档识别流程框图

2 特征分析

2.1 专家独立页面特征分析

专家主页、专家百科页、专家学术资源页及专家博客页这 4 类证据文档有各自明显的特点:

- 专家主页信息简洁并带有明显的标签,一般包含专家的姓名、地址、联系电话及科研成果,如获得自然科学基金、发表的论文、参与的项目等,同时还包含一些非常重要的词,如教授、博士、项目、科研等信息.
- 专家百科页的内容大都是百科网站运营商直接摘自专家主页或者对主页的概括,就页面内容而言很难与专家主页进行区分.但是百科页面也有自己的特点,比如,国内主要的几个百科网站在其链接当中往往都有类似“baike”的字样,并且百科页面大都有导航栏等.
- 专家学术资源页与专家的百科页类似,包含专家的基础信息和专家论文发表的情况,具有较为固定的格式,比如,链接中包含“cnki”字样.
- 专家博客页页面 URL 地址中通常包含博客页面标志词“blog”.

针对以上专家证据文档特点分析,专家证据文档独立页面特征主要包括以下 3 类:

(1) 页面 URL 地址链接特征:页面 URL 地址链接中包含文档标记类型的特征词,如 baike,wiki,cnki,blog,bokee 等.

(2) 元数据词特征:描述专家信息的元数据,如,姓名、组织机构、职称、邮箱地址、联系电话、研究领域、项目、论文、专利、获奖等.

(3) 页面链接特征:证据文档页面中包含指向同一域名下的二层或者三层链接,如指向发表论文对应的

pdf 链接、指向研究成果详细信息页面链接等.

提取专家证据文档独立页面特征的基本思想是:首先对页面 URL 地址进行分词,提取页面 URL 地址链接特征;然后对网页去标签、分词,利用模板匹配的方法提取出网页中对应的专家元数据词特征;最后,利用正则表达式提取出文档中页面链接特征.专家证据文档的部分独立页面特征见表 1.

Table 1 Examples for part of the independent characteristics of the page in expert evidence documents

表 1 专家证据文档部分独立页面特征示例

特征类型	特征取值
标题是否包括专家姓名及组织机构	{0,1}
页面中 pdf 链接数量是否大于 10(包括参考文献)	{0,1}
页面中是否包含专家姓名、组织结构等基本信息	{0,1}
页面地址是否包含某个链接特征词,如 baike,wiki,cnki,dblp,blog,bokee 等)	{0,1}
页面中是否包含导航栏	{0,1}
正文是否包含邮箱地址	{0,1}
正文是否包含电话号码	{0,1}
页面中是否包含某个特征词	{0,1}
…	…

构建的专家独立页面特征与页面标记之间关系的无向图结构如图 2 所示.

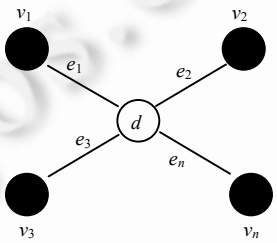

Fig.2 Expert independent page features and page labeled graph structure

图 2 专家独立页面特征与页面标注图结构

图 2 中,白色节点 d 表示候选专家证据文档的标记,黑色节点 $v_i(1\leqslant i\leqslant n)$表示专家证据文档独立页面特征,边 e_i 表示独立页面特征与候选文档标记间的依赖关系.候选文档 d 的每个独立页面特征 v_i 与对应的候选文档标记形成图结构中的一个团.

2.2 专家证据文档页面关联关系分析

为了分析专家证据文档之间的关联关系,需要从互联网中抓取要检索的专家相关的大量页面,并对抓取的页面进行分类整理,将每个专家相关的页面整理成一簇.专家页面抓取的过程是:首先,调用 Google API 检索指定姓名和组织机构的专家页面;然后,通过收集搜索引擎返回的前若干个页面获得高质量的候选专家证据文档集合;最后,将每一位专家相关的文档整理成一个候选文档集合.本文将每位专家对应的候选文档集合定义为一簇,为此,专家证据页面间的关联关系分为簇内关联关系和簇间关联关系.分析同一专家候选文档集中页面间的关联关系称为簇内关联关系分析,分析不同专家候选文档集合间的页面关联关系称为簇间关联关系分析.

(1) 专家消歧

由于存在中文专家同名现象,为了获取高质量的候选专家证据文档,需要对同名专家进行人名消歧.而专家页面之间在内容上会存在很多关联,这些关联关系对专家唯一性的确定有极大的支撑作用.为此,针对中文人名和专家属性特点,采用融合页面关联关系的中文专家谱聚类消岐方法[15].在课题组前期相关工作中,已通过实验证明了该方法具有较好的效果.中文专家人名消岐的思想是:首先,采用 TF-IDF 算法计算基于词的特征权重,利用余弦相似度算法计算页面之间的相似度,得到专家页面初始相似度矩阵;然后,以专家关联关系特征作为半监督约束信息对初始相似度矩阵进行校正,利用基于谱聚类[16,17]的方法构建专家消歧模型.

(2) 关联关系分析

专家之间的关联关系体现在证据文档之间的相互关联,包括页面链接关系和内容关联关系.本文将专家证据文档之间的关联关系分为以下 3 种情况,其中,A,B 为簇间关联关系,C 为簇内关联关系.

- 关联 A:相互链接的页面属于同一种类型的专家证据文档.例如,专家主页往往有链接指向另一位专家的主页,专家学术资源页大都有链接指向另一个专家的学术资源页等,但很少有专家主页指向另一个专家的博客页或学术资源页,如图 3(a)所示.
- 关联 B:一个专家大都只有 1 个主页,而且不会与其他专家共用一个主页,如果一个页面同时被判定为两个人的主页,那么这个页面不是主页,如图 3(b)所示.
- 关联 C:同一专家证据文档之间存在链接关联,例如,专家主页中大都有链接指向专家自己的博客、微博等页面的链接;同一个专家的不同类型的证据文档之间具有内容相似的关联关系,例如,专家百科页的内容往往是直接摘自专家的主页或来源于专家的主页;每位专家大都只有 1 个同一类型的证据文档,如博客页、主页等.因此,同一专家的相关页面中如果两个页面有链接关系,则这两个页面属于不同的证据文档类别.专家证据文档簇内关联关系如图 3(c)所示,不同类型的页面之间都有边相连.

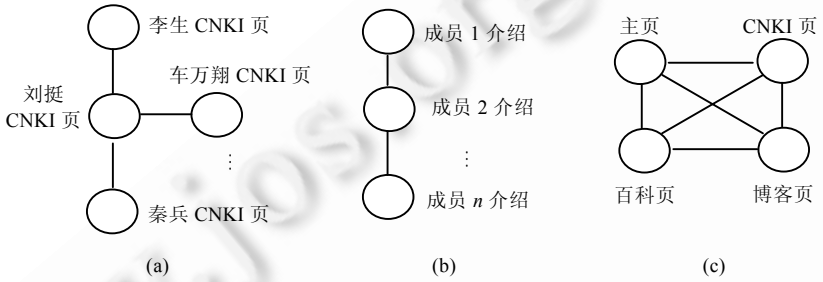

Fig.3 Examples for expert evidence document link relationship

图 3 专家证据文档链接关联关系示例

在图 3(a)中,专家刘挺 CNKI**页中包含了许多指向合作者的 CNKI 页面链接;在图 3(b)中,同一个页面中同时出现实验室所有成员的介绍,这个页面不能同时作为所有成员的共同主页.

(3) 专家证据文档相似度计算

文本相似度的计算方法往往是基于词特征,而专家证据文档的内容相似性是指两篇文档表述的专家信息是否相似.因此,可通过计算表征专家页面信息的命名实体之间的相似度来表示两篇证据文档内容的相似度.专家证据文档相似度计算的思想是:首先对候选专家证据文档进行去标签、分词、词性标注等语料预处理;然后,基于融合长距离依赖特征的专家证据文档中文命名实体识别方法[18]识别专家证据文档中的命名实体;最后,建立文档中候选实体的向量空间模型并利用向量之间的余弦夹角进行候选文档的相似度计算[19,20].例如,两篇候选文档 D_i,D_j,经过命名实体识别后得到的向量为 $D_i=(C_{1i},C_{2i},C_{3i},\ldots,C_{mi})^T$ 与 $D_j=(C_{1j},C_{2j},C_{3j},\ldots,C_{mj})^T$,则候选文档 D_i,D_j 的相似度系数为

$$Cos(D_i,D_j)=\frac{\sum_{k=1}^{m}C_{ki}C_{kj}}{\sqrt{\sum_{k=1}^{m}C_{ki}^2\sum_{k=1}^{m}C_{kj}^2}} \tag{1}$$

在相似度计算方法中,C_{ki} 和 C_{kj} 为命名实体在候选文档中的权重,其值可以利用 TF/IDF 方法[21]取得.同时,需要结合专家证据文档识别的具体任务,给余弦系数设定一个阈值来判断两篇候选文档是否相似.

** http://dbpub.cnki.net/grid2008/DetailAuthor/-DetailView.aspx?authorId=-06994824

3 专家证据文档识别无向图模型

3.1 马尔可夫网理论

Markov 网,也称为 Markov 随机场(Markov random field,简称 MRF)或者无向图模型(undirected graphic model),是一个随机变量集合 $X=(X_1,X_2,\ldots,X_n)\in\chi$的联合概率分布模型[13,14].马尔可夫网由一个无向图 G 和定义于 G 上的一组势函数ϕ_k组成.无向图上的每个节点都代表一个随机变量,所有节点都有边相连的最大子图,称为团.马尔可夫网中的每一个团都对应着一个非负实函数,称为势函数,团中节点的不同状态对应势函数不同的函数值.Markov 网所代表的随机变量集合的联合概率分布为

$$P(X=x)=\frac{1}{Z}\prod_k\phi_k(x_{\{k\}}) \tag{2}$$

其中,$x_{\{k\}}$表示 Markov 网中第 k 个团的状态,即对应第 k 个团中所有变量的取值状态;$\phi_k(x_{\{k\}})$表示第 k 个团对应的势函数;Z 是归一化因子,且 $Z=\sum_{x\in\chi}\phi_k(x_k)$.

将势函数表示为对数线性模型的形式:

$$\phi_k(x_{\{k\}})=\exp\left\{\sum_k\theta_k f_k(x_{\{k\}})\right\} \tag{3}$$

Markov 网的联合概率分布转化为

$$\begin{aligned}P(X=x)&=\frac{1}{Z}\prod_k\phi_k(x_{\{k\}})\\&=\frac{1}{Z}\prod_k\exp\{\theta_k f_k(x_{\{k\}})\}\\&=\frac{1}{Z}\exp\left(\sum_k\theta_k f_k(x_{\{k\}})\right)\end{aligned} \tag{4}$$

其中,$f_k(x_{\{k\}})$表示第 k 个团中随机变量的特征函数,θ_k表示第 k 个团对应的权重.其中,为计算方便,可以将特征函数定义为

$$f(x_{\{k\}})=\begin{cases}1, & \text{特征与标记一致}\\0, & \text{否则}\end{cases} \tag{5}$$

3.2 模型表示

通过融合专家证据文档的独立页面特征以及页面之间的关联关系特征,构建专家证据文档识别无向图模型 $G=\langle V,E\rangle$,其中,V 表示顶点,E 表示顶点之间的边,并且 $V=T\cup X$,$E=E_{T,T}\cup E_{X,T}$,其中,$x_i\in X$ 表示的独立页面特征节点,$t_i\in T$表示候选专家证据文档节点,$E_{T,T}$表示候选专家证据文档节点之间的关联关系形成的边,$E_{X,T}$表示独立页面特征节点和对应的专家证据文档节点之间形成的边.任意两个节点之间都有边相连的最大子图称为团.专家证据文档识别无向图模型图结构示例如图 4 所示.

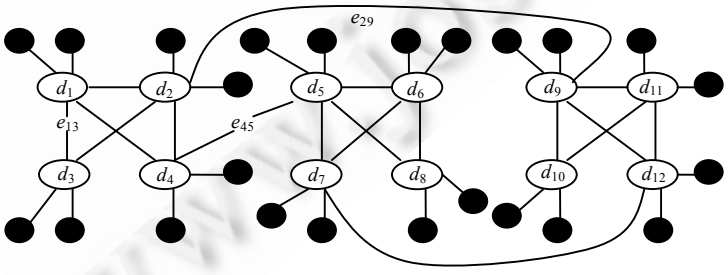

Fig.4 Schematic diagram for undirected graph model of the expert evidence document recognition

图 4 专家证据文档识别无向图模型图结构示意

在图 4 中,黑色节点表示独立页面特征,白色节点表示要标注的候选专家证据文档节点,黑色节点与白色节点之间的边表示独立页面特征与候选专家证据文档标记之间的依赖关系,白色节点之间的边表示候选专家证据文档标记之间的依赖关系.根据本文第 2.2 节定义的关联关系,如果 t_k 和 t_j 为不同页面并且它们之间有边相连,则 $e_{kj}\in A$,比如,图 4 中 e_{45} 表示文档 d_4 和文档 d_5 有链接关系;如果 t_k 和 t_j 是相同页面,则 $e_{kj}\in B$,比如,e_{29} 表示文档 d_2 与文档 d_9 为同一篇文档;如果 t_k 和 t_j 是对相同的查询条件返回的页面,则 $e_{kj}\in C$,例如,e_{13} 表示文档 d_1 与文档 d_3 为同一个专家的证据文档.

专家证据文档识别无向图中包含两种类型的团,即专家证据文档独立页面特征与候选专家证据文档标记形成的团以及专家证据文档之间的关联关系形成的团.通过对专家证据文档识别无向图中的不同类型的团进行分析,根据公式(4)马尔可夫网的联合概率分布,专家证据文档识别无向图模型的联合概率分布可以表示为

$$P(X=x)=\frac{1}{Z}\exp\left[\left(\sum\nolimits_i \theta_i \chi_i\{x_i\}\right)+\left(\sum\nolimits_j \theta_j \chi_j\{x_j\}\right)\right] \tag{6}$$

其中,i 表示页面的独立特征与页面标注之间形成的团,j 表示相互依赖的页面标注节点之间形成的团.

针对专家证据文档识别任务的特点,在专家证据文档识别无向图模型中,同一类型特征在图结构中形成的不同的团应当具有同样相同的权重,为此进行如下规定:

(1) 同一独立页面特征对同一类别的不同页面标注的贡献一致;

(2) 同一种类型页面之间的依赖关系对不同的页面标注的贡献一致.

根据以上规定,专家证据文档识别无向图模型的联合概率分布可以进一步简化表示为

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_m \theta_m\left(\sum_n \chi_n\{x_n\}\right)\right) \tag{7}$$

其中,m 表示特征的种类,n 表示第 i 个特征对应图中团的数量,Z 是归一化因子.由此可见,改进后的专家证据文档识别无向图模型中的参数得到了有效的改进,简化了模型的参数学习及推理难度.

3.3 模型表示

(1) 概率推理

概率图模型的基本推理问题就是在已知证据变量 $E=e$ 的前提下,求解模型的最大可能性解释问题(MPE 问题)[13,14].为此,专家证据文档识别模型中的推理问题就是在已知页面的独立特征以及页面之间的依赖关系特征的前提下,求解候选证据文档的最大可能性标注问题.概率公式表示如公式(8)所示:

$$\begin{aligned}\arg\max_y P(y\mid e)&=\arg\max_y \frac{1}{Z_x}\exp\left(\sum_{i=1}^{m}\theta_i\sum_{j=1}^{n}\chi_j\{x_j\}\right)\\&=\arg\max_y \sum_{i=1}^{m}\theta_i\sum_{j=1}^{n}\chi_j\{x_j\}\end{aligned} \tag{8}$$

在图的结构为链状或者树状的情况下,精确推理算法可以高效地进行,如变量消元算法、团树传播算法等.由于识别专家证据文档的图结构较为复杂,精确推理算法往往行不通,因此采用近似推理算法对公式(8)进行求解.由于静态蒙特卡洛算法要产生高维随机变量的随机数非常困难,为此,本文采用动态的蒙特卡洛算法中的吉布斯采样算法对上式进行求解.该算法求解的基本思想是:在所有变量的联合状态空间中与 $E=e$ 一致的子空间里进行随机漫步,先任选一个起点,以后的每一步都只依赖于前一步的状态.下面给出求解算法的伪代码:

Gibbs Sampling(G,m,E,e,Q,q,ρ).

输入:G——用于证据文档识别的无向图模型;m——样本量;

E——证据变量(独立页面的特征节点);e——证据变量的取值;

Q——查询变量(候选页面节点);q——查询变量的一组取值;

ρ——非证据变量的抽样顺序.

输出: $\arg\max_y(y\mid e)$.

Step 1. for (q in Q)
Step 2. m_q=0; j=0;
Step 3. 随机生成一个与 E=e 一致的样本 D_1;
Step 4. if (D_1 与 Q=q 一致)
Step 5. m_q=m_q+1;
Step 6. end if
Step 7. for (i=2 to m)
Step 8. D_i=D_{i-1};
Step 9. for (ρ中的每一个变量 Z)
Step 10. 设 Y=mb(Z),y_i 是 Y 在 D_i 中的当前取值,从 $P(Z|Y=y_i)$;
Step 11. 用抽样结果代替 D_i 中 Z 的取值;
Step 12. end for
Step 13. if (D_i 与 Q=q 一致)
Step 14. m_q=m_q+1;
Step 15. end if
Step 16. end for
Step 17. P_j=m_q/m;
Step 18. j++;
Step 19. end for
Step 20. return max(P_j).

(2) 参数学习

一般概率图模型的学习分为结构学习和参数学习两个方面.结构学习既要确定网络的结构又要确定网络中的参数[13,14];参数学习是已知网络结构的前提下,确定网络参数的问题.由于本文提出的专家证据文档识别的无向图模型中图结构是确定的,为此,学习问题就只涉及参数学习.针对专家证据文档识别的具体任务,参数学习就是根据已经标注好的证据文档的语料,学习出联合概率公式中独立页面特征和页面相互依赖关系特征所形成团的权重.采用最大似然估计方法对特征的权重进行估计,求解过程如下:

Step 1. 对公式(9)中联合概率分布取对数后,得到对应的似然函数,如公式(10)所示:

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_{i=1}^{m}\theta_i\sum_{j=1}^{n}\chi_j\{x_j\}\right) \tag{9}$$

$$\log P(X=x)=\sum_{i=1}^{m}\theta_i\sum_{j=1}^{n}\chi_j\{x_j\}-\log Z \tag{10}$$

Setp 2. 对公式(10)中的一个权重进行求导:

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\log P(X=x)&=\frac{\partial}{\partial\theta_i}\sum_{i=1}^{m}\theta_i\sum_{j=1}^{n}\chi_j\{x_j\}-\frac{\partial}{\partial\theta_i}\log Z\\&=\sum_{j=1}^{n}\chi_j\{x_j\}-\frac{1}{Z}\frac{\partial}{\partial\theta_i}Z\\&=\sum_{j=1}^{n}\chi_j\{x_j\}-\frac{1}{Z}\sum_{x'}\frac{\partial}{\partial\theta_i}\exp\left(\sum_{i=1}^{m}\theta_i\sum_{j=1}^{n}\chi_j\{x_j\}\right)\\&=\sum_{j=1}^{n}\chi_j\{x_j\}-\sum_{x'}P_{\theta_i}(X=x')\sum_{j=1}^{n}\chi_j\{x'_j\}\end{aligned} \tag{11}$$

由上式可知:在给定独立页面特征节点真值的情况下,只需要计算当前赋值情况下特征之和以及在所有可能赋值情况下带概率的和,就可以求得当前的权重;然后,利用梯度下降方法可以求得所有的未知权重.

4 实验与结果分析

4.1 数 据

为了验证模型的有效性,从计算机科学技术和生物学这两个学科领域随机收集各 600 个专家,对于每个专家,利用 Google API 通过检索专家的姓名和组织机构,收集搜索引擎返回排在最前面的 10 个页面.另外,由于一些专家在其单位的网站上有他们的页面,并介绍了详细信息,例如教育经历、学术论文、参与项目、联系方式等,将这些页面也作为专家的主页.百科页面大都将同一姓名的专家在同一个页面中显示.在收集语料过程中,采用融合页面关联关系的中文专家谱聚类消岐方法[15]过滤同一百科页面中其他专家的信息,只保留与查询主题相关的专家信息,并且过滤掉检索结果中的非网页文档,比如 pdf 文件、word 文件等.对于一些专家存在自己某一类型的证据文档但是搜索引擎返回结果中没有的情况,通过人工方式将专家证据文档加入语料库.表 2 给出了收集语料的详细统计.

Table 2 Statistics for the experimental data

表 2 实验数据统计

学科	专家总数	页面类型	页面数	平均含有词语数	关联 A 的边的总数	关联 B 的边的总数
计算机科学技术	600	主页	451	451	72	9
		百科页	445	443	14	18
		CNKI 页	523	377	218	27
		博客页	108	489	29	0
		其他	4 473	427	31	14
		合计	6 000	–	364	68
		超过一个候选页面专家数		没有候选页面专家数	少于 4 个候选页面专家数	最大团节点数
		487		92	132	4
生物学	600	页面类型	页面数	平均含有词语数	关联 A 的边的总数	关联 B 的边的总数
		主页	396	422	46	13
		百科页	503	394	9	25
		CNKI 页	539	364	156	40
		博客页	97	452	18	0
		其他	4 465	436	28	15
		合计	6 000	–	257	93
		超过一个候选页面专家数		没有候选页面专家数	少于 4 个候选页面专家数	最大团节点数
		325		103	76	4

4.2 实验设置及结果分析

为了验证所提出的方法的有效性,我们将本文提出的无向图模型 undirected graph(UG)model 与逻辑回归 logistic regression(LR)模型、支持向量机(SVM)模型、马尔可夫逻辑网 markov logic networks(MLN)模型进行对比,并根据本文第 2 节定义的 3 种特征类型对比在使用相同特征的情况下各种方法的效果.实验中统计了各种方法的准确率 Precision(P)和召回率 Recall(R)两个指标.在这两个指标的基础上,利用 F 值作为衡量所提出的方法的最终评测指标.准确率、召回率以及 F 值的公式表示如下:

$$准确率(P)=\frac{系统正确识别的专家证据文档数}{系统识别的专家证据文档数}\times 100\% \tag{12}$$

$$召回率(R)=\frac{系统正确识别的专家证据文档数}{专家证据文档总数}\times 100\% \tag{13}$$

$$F值=\frac{2\times P\times R}{P+R}\times 100\% \tag{14}$$

以下通过 4 个实验验证了本文提出无向图模型对专家主页、百科页、CNKI 页、博客页这 4 种类型的证据文档识别的效果.其中,实验 1 将不同类型的专家证据文档认为是相互独立的,分别对每种类型的文档进行识别实验;实验 2、实验 3 及实验 4 则是将专家主页、百科页、CNKI 页、博客页这 4 类文档作为一个整体,识别

效果是指对这 4 种文档的整体识别效果.实验 1 是指具体的某一种类型的证据文档识别效果.实验 2 与实验 3 不同的是:实验 2 的训练数据和测试数据只是第 4.1 节中某一个学科领域的数据;实验 3 则是整个数据集,并且训练和测试的数据不是同一个学科领域的数据.实验 4 验证了不同规模的训练数据下,无向图模型对专家证据文档识别效果的影响.

实验 1. 分别对每种类型的证据文档进行识别.

实验 1 将专家主页、百科页、CNKI 页、博客页这 4 种类型的证据文档认为是相互独立的页面,在第 4.1 节收集的整个数据集上分别对每种类型的证据文档进行识别.表 3 给出了 LR,SVM,MLN 及 UG 这 4 种方法识别的准确率(*P*)、召回率(*R*).由于 LR 和 SVM 这两种模型要求特征是独立的,实验中 LR 和 SVM 方法只考虑专家证据文档的独立页面特征;MLN 和 UG 模型中使用的特征之间具有相关性,为此,可以同时使用专家证据文档的多种特征.

Table 3 Comparison of recognizing each evidence document respectively using different methods on the computer and biology data sets

表 3 在计算机与生物学混合数据集上不同方法分别对每种证据文档识别的比较

方法	文档类型							
	主页		百科页		CNKI 页		博客页	
	P (%)	*R* (%)	*P* (%)	*R* (%)	*P* (%)	*R* (%)	*P* (%)	*R* (%)
LR_a	84.9	79.8	86.7	87.1	89.4	84.9	86.1	72.3
SVM_a	85.2	80.3	88.5	84.6	90.3	85.2	87.7	70.5
MLN_{ab}	86.1	83.7	89.3	87.9	92.7	86.5	89.4	72.1
UG_{ab}	87.3	84.5	92.2	86.9	94.4	89.3	91.2	73.0
UG_{abc}	88.1	85.3	95.3	87.4	95.1	90.0	92.6	72.8

在表 3 中,LR_a 指融合专家独立页面特征的逻辑回归方法,SVM_a 指融合专家独立页面特征的 SVM 方法,MLN_{ab} 指融合专家独立页面特征及专家证据文档簇间关联特征的马尔可夫逻辑网方法,UG_{ab} 指融合专家独立页面特征、专家证据文档簇间关联特征的无向图模型,UG_{abc} 指融合专家独立页面特征、专家证据文档簇间关联及簇内关联特征的无向图模型.从表 3 可以看出:以上各种方法对专家百科页、CNKI 页具有较好的识别效果,对博客页的识别效果最低,对主页的识别效果接近中间效果.以上方法对各种证据文档识别的 *F* 值效果比较如图 5 所示.对百科页及 CNKI 页的识别效果最好,是由于收集的实验数据中,大多数专家都有百科页和 CNKI 页,并且这两种类型的页面专家之间的簇间关联关系更明显,而且这两类页面具有明显的标记特征,如,百科页的 URL 中带有“baike”,CNKI 页的页面 URL 中带有“CNKI”标记,因此,独立页面特征和簇间关联特征有助于提高识别效果.由于拥有博客页的专家较少,并且不同专家之间博客页的簇间关联不明显,因此效果较差.在主页识别方面,由于专家主页没有类似于百科页那样明显的标记,必须综合独立页面特征(如姓名、专业、职称、研究方向、发表论文等专家基本信息)和专家之间的关联特征进行分析.

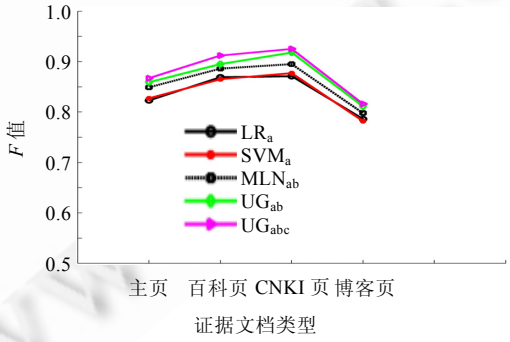

Fig.5 *F* value on each kind of evidence document recognition with different methods

图 5 不同方法对每一种证据文档识别的 *F* 值

从实验结果可以看出,独立地对各类证据文档进行识别,马尔可夫逻辑网模型和无向图模型的效果与其他方法相比都具有较好的效果,其中,无向图模型效果最好.

实验 2. 同一学科领域专家证据文档的识别.

表 4 分别给出了在计算机科学及生物学两个学科领域数据集上进行证据文档识别的准确率(*P*)、召回率(*R*)及 *F* 值的比较.从表 4 可以看出,LR_a 和 SVM_a 这两种方法很接近,但还不能说明哪一个方法更好.我们可以看出:在这两个数据集上,MLN_{ab} 及 UG_{ab} 的效果明显超过 LR_a 和 SVM_a 这两种方法,UG_{ab} 的效果略优于 MLN_{ab},UG_{abc} 效果最好.这是由于实验收集的数据集大多专家拥有多个证据文档,因此,同一个专家的证据文档之间表现出较强的簇内关联特征.同时,考虑 3 种类型的特征以后,UG_{abc} 具有更好的效果,与 MLN_{ab} 的实验效果相比,UG_{abc} 的识别效果更明显.

Table 4 Comparison of expert evidence document recognition on the same subject area

表 4 同一学科领域专家证据文档识别的比较

方法	计算机科学			生物学		
	P (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)
LR_a	88.9	88.3	88.6	89.5	85.8	87.6
SVM_a	90.4	86.1	88.2	90.2	85.7	87.9
MLN_{ab}	91.3	88.1	89.7	91.6	85.9	88.7
UG_{ab}	92.4	88.9	90.6	92.3	86.3	89.2
UG_{abc}	93.6	89.5	91.5†	93.1	87.5	90.2†

实验 3. 不同学科领域混合数据集上专家证据文档的识别.

实验 2 分别验证了在计算机科学与生物学这两个学科领域收集的数据集上,专家证据文档识别取得了较好的效果,如果再增加不同类别学科领域的专家数据作为实验数据,实验效果是否会随着学科类别数的增加而改变?为此,实验 3 将计算机科学及生物学这两个学科领域的专家页面数据混合在一起作为整个实验的数据集.表 5 给出了在混合数据集上进行证据文档识别的准确率(*P*)、召回率(*R*)及 *F* 值的比较结果.

Table 5 Comparison of expert evidence document recognition on the computer and biology data sets

表 5 计算机与生物学混合数据集上专家证据文档识别比较

方法	*P* (%)	*R* (%)	*F* (%)
LR_a	88.5	81.9	85.1
SVM_a	89.6	81.6	85.4
MLN_{ab}	91.7	82.2	86.7
UG_{ab}	92.1	83.9	87.8
UG_{abc}	93.3	84.5	88.7†

通过对比表 4、表 5 给出的实验结果可以看出,实验 3 在混合数据集上的效果与实验 2 分别在同一学科领域进行证据文档识别的效果相比没有明显的差别.由此可见,本文方法对专家证据文档的识别效果与实验数据所选择的专家学科领域的类别数关系不大.另外,对比表 3~表 5 的结果我们可以看出,表 3 中分别针对专家单个证据文档识别的效果的平均值明显低于表 4、表 5 中把专家各种证据文档作为一个整体进行识别的效果.这是由于表 3 中将专家的各类证据文档认为是相互独立的,不能很好地利用证据文档的簇内关联特征.表 4、表 5 则是充分利用了专家证据文档的簇间关联特征和簇内关联特征,因此能够取得较好的效果.

实验 4. 无向图模型在不同训练数据规模的识别效果.

为了验证我们提出的无向图模型对专家证据文档识别的效果是否与训练数据集的规模大小有明显的关系,我们分别针对计算机、生物学及这两个学科领域的混合数据集进行了不同数据规模下的专家证据文档识别实验,实验结果如图 6 所示.从图 6 中我们可以看出:当训练数据的规模分别为 100~200 个专家时,计算机科学技术和生物学这两个领域进行证据文档识别的 *F* 值与混合数据集上识别的 *F* 值基本接近;随着训练数据规模的逐渐增大,专家数量达到 300 人时,相应的 *F* 值也在逐渐增大;当专家人数达到 400~500 人之间时,不同数据集上专家证据文档识别的 *F* 值变化越来越小;数据规模再增大,相应的 *F* 值基本没变化.由此可见,本文提出的无向图模型对专家证据文档识别的效果是比较稳定的.

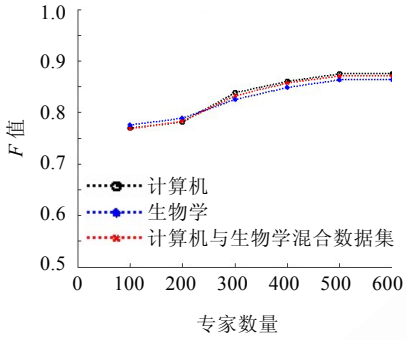

Fig.6 F value of the undirected graph model on different size of training data

图 6 无向图模型在不同训练数据规模的 F 值

5 总结与展望

专家证据文档识别是专家检索的关键步骤.针对专家证据文档具有页面独立特征及页面间链接关系和内容关联关系的特点,本文融合了专家证据文档的独立页面特征以及文档之间的关联关系特征,提出了一种专家证据文档识别无向图模型,并通过模拟实验验证了方法的有效性.实验结果表明,本文提出的方法具有较好的效果.在未来的研究中,本文提出的方法还可以进一步扩展,用于其他实体任务检索.另外,还可以通过证据文档中的独立页面属性特征、页面标记特征及一些潜在的关联特征进一步改进本文提出的无向图模型,在确保召回率不降低的情况下,提高专家证据文档识别的准确率.

致谢 在此,我们感谢对本文的工作给予支持和建议的昆明理工大学智能信息处理重点实验室专家检索资源获取与学习排序课题组全体成员.

References:

[1] Macdonald C, Ounis I. Voting for candidates: Adapting data fusion techniques for an expert search task. In: Proc. of the 15th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2006. 387−396. [doi: 10.1145/1183614.1183671]

[2] Macdonald C, Hannah D, Ounis I. High quality expertise evidence for expert search. Lecture Notes in Computer Science, 2008, 4956:283−295. [doi: 10.1007/978-3-540-78646-7_27]

[3] Craswell N, de Vries AP, Soboroff I. Overview of the trec-2005 enterprise track. In: Proc. of the TREC 2005 Conf. New York: IEEE Press, 2005. 199−205.

[4] Xi WS, Fox EA, Tan RP, Shu J. Machine learning approach for homepage finding task. In: Proc. of the 9th Int'l Symp. on String Processing and Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2002. 145−159. [doi: 10.1007/3-540-45735-6_14]

[5] Tang J, Zhang D, Yao LM. Social network extraction of academic researchers. In: Proc. of the 17th IEEE Int'l Conf. on Data Mining (ICDM 2007). Washington: IEEE Press, 2007. 292−301. [doi: 10.1109/ICDM.2007.30]

[6] Bron M, Balog K, de Rijke M. Ranking related entities: Components and analyses. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2010. 1079−1088. [doi: 10.1145/ 1871437.1871574]

[7] Li LN, Yu ZT, Zou JJ, Su L, Xian YT, Mao CL. Research on entity homepage recognition method. Journal of Computational Information System, 2009,5(6):1617−1624.

[8] Fang Y, Si L, Yu ZT, Xian YT, Xu YB. Entity retrieval with hierarchical relevance model. In: Proc. of the 18th Text REtrieval Conf. (TREC 2009). New York: IEEE Press, 2009.

[9] Fang Y, Si L, Mathur AP. Discriminative graphical models for faculty homepage discovery. Journal of Information Retrieval, 2010, 13(6):618−635. [doi: 10.1007/s10791-010-9127-7]

[10] Wu ZJ, Yu ZT, Su L, Liu L, Xian YT. Research on the method of expert homepage recognition based on Markov logic networks. Journal of Computational Information System, 2012,8(3):1089−1096.

[11] Macdonald C, Ounis I. Voting for candidates: Adapting data fusion techniques for an expert search task. In: Proc. of the CIKM 2006. New York: ACM Press, 2006. 387−396. [doi: 10.1145/1183614.1183671]

[12] Balog K, Azzopardi L, de Rijke M. Formal models for expert finding in enterprise corpora. In: Proc. of the SIGIR 2006. New York: ACM Press, 2006. 43−50. [doi: 10.1145/1148170.1148181]

[13] Jordan MI. Graphical models. Statistical Science, 2004,19(1):140−155. [doi: 10.1214/088342304000000026]

[14] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. Cambridge: Massachusetts Institute of Technology Press, 2009. [doi: 10.1007/978-3-642-38466-0_28]

[15] Tian W, Shen T, Yu ZT, Guo JY, Xian YT. A Chinese expert name disambiguation approach based on spectral clustering with the expert page-associated relationships. Lecture Notes in Electrical Engineering, 2013,256:2013. 245−253.

[16] Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Dietterich TG, Becker S, Ghahramani Z, eds. Advances in Neural Information Processing Systems (NIPS) 14. Cambridge: MIT Press, 2002. 894−856.

[17] Wang L, Bo LF, Jiao LC. Density-Sensitive semi-supervised spectral clustering. Ruan Jian Xue Bao/Journal of Software, 2007, 18(10):2412−2422 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2412.htm [doi: 10.1360/jos182412]

[18] Wu ZJ, Yu ZT, Guo JY, Mao CL, Zhang YM. Fusion of long distance dependency features for Chinese named entity recognition based on Markov logic networks. In: Proc. of the Natural Language Processing and Chinese Computing. Natural Language Processing and Chinese Computing Communications in Computer and Information Science, 2012,333:132−142. [doi: 10.1007/978-3-642-34456-5_13]

[19] Luenberger DG. Optimization by Vector Space Methods. Hoboken: Wiley-Interscience, 1997.

[20] Zhang D, Lee WS. Question classification using support vector machines. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval. New York: ACM Press, 2003. 26−32. [doi: 10.1145/860435.860443]

[21] Aizawa A. An information-theoretic perspective of TF-IDF measures. Information Processing & Management, 2003,39(1):45−65. [doi: 10.1016/S0306-4573(02)00021-3]

附中文参考文献:

[17] 王玲,薄列峰,焦李成.密度敏感的半监督谱聚类.软件学报,2007,18(10):2412−2422. http://www.jos.org.cn/1000-9825/18/2412.htm [doi: 10.1360/jos182412]

毛存礼(1977—),男,云南曲靖人,博士生,讲师,CCF 学生会员,主要研究领域为自然语言处理,信息检索.
E-mail: maocunli@163.com

郭剑毅(1964－),女,教授,CCF 会员,主要研究领域为自然语言处理,信息检索,机器翻译.
E-mail: gjade86@hotmail.com

余正涛(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索,中文问答系统.
E-mail: ztyu@hotmail.com

线岩团(1981－),男,博士生,讲师,CCF 会员,主要研究领域为自然语言处理,机器翻译.
E-mail: 195426286@qq.com

吴则建(1986－),男,硕士,主要研究领域为自然语言处理,信息检索.
E-mail: zejian.km@gmail.com