互联网科技专家搜索系统①

莫 倩, 张传想

(北京工商大学 计算机与信息工程学院, 北京 100037)

摘 要: 提出和设计了一个为用户自动收集、分析和整理科技专家信息的科技专家搜索系统. 描述了互联网科技专家搜索系统的体系结构、主要特征和关键技术. 系统采用基于特征向量的分类算法, 设计了一种适合于互联网大规模科技专家信息抽取的方法, 利用互联网上的信息资源高效的抽取专家信息、挖掘专家学术关系, 为科技专家发现和搜索研究提供了一种新的思路.

关键词: 专家搜索; 信息抽取; 社会网络搜索; 学术搜索; 专家发现

Internet Technology Expert Search System

MO Qian, ZHANG Chuan-Xiang

(School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract: This paper proposes and designs a technology experts searching system which can automatically collect, analysis and access information of technology expert. It describes the system's structure, main function and key technology. Based on the classification algorithm of feature vector, this paper addresses to design a kind of method suitable for large-scale information extraction, it can effectively extract and mining academic social networks from the information resources on the Internet. It provides a new train of thought for science and technology expert information extraction research.

Key words: expert search; information extraction; social network search; academic search; expert finding

在互联网应用走向多元化的今天, 伴随着各类领域搜索引擎的出现, 搜索引擎的搜索功能日益完善, 学术搜索也逐渐成为新的研究焦点. 人们想要提升业务技能, 进行科研合作, 就需要搜索某一领域科技专家的相关信息、掌握前沿的科技成果, 而传统的搜索引擎无法方便快捷地提供全面详尽的科技专家信息.

本文通过对互联网科技专家信息特征的研究与抽取, 实现互联网上大规模分散的专家信息的采集, 并分析处理这些采集到的专家信息, 构建完整的专家知识库. 最终实现用户只需输入一次关键词, 即搜索科技专家的名字或某一专业领域, 就可获得该专家或领域的多方位信息, 能够快捷有效地实现专家信息的全面搜索, 大幅度地减少获取信息的时间.

1 相关工作

基于领域的搜索引擎的研究是未来搜索引擎的发展趋势, 而学术搜索的研究与应用也蓬勃兴起, 例如专家发现和专家关系搜索. 早在 2005 年的文本挖掘会议(TREC)上提出了一种用于专家发现的方法, 并提供了企业级专家搜索追踪平台[1], K. Balog 等和 Y. Ma, S. Shi 等曾提出了扩展语言模型用于专家发现的研究[2,3]. 这些研究提出了一些用于专家发现的方法, 但是对专家网络关系的提取却未能进一步的描述, 而目前社会网络关系搜索也是热门的内容, 其目标是致力于发现人们之间的社会网络关系, 例如早期研究和开发的 Web 社会网络关系的 ReferralWeb 系统[4], 以及微软亚洲研究院开发设计人立方关系搜索是一款新型社会化搜索引擎, 帮助人们搜索主体的社会网络关系. 此外, 还有一些相关的学术搜索引擎, 例如 Google 学术搜索、微软学术搜索、Rexa.info 等, 尽管学术搜索方面已做了大量的工作, 但是还没有提出一个比较完整全面的系统, 没能够进一步挖掘出专家发现及学术搜索

① 基金项目:国家自然科学基金(61170112);北京市教委科技创新平台专项(2011101)
收稿时间:2012-10-22;收到修改稿时间:2012-11-28

中存在的隐性知识，于是我们研究和设计了针对科技专家的领域搜索系统，全面综合整理分析专家基本信息、专家学术信息、以及专家学术关系.

2 科技专家搜索系统概述

本系统总体分成三个模块：专家信息采集模块、专家数据处理模块、专家信息查询与检索模块. 科技专家搜索系统的总体框图如图 1 所示.

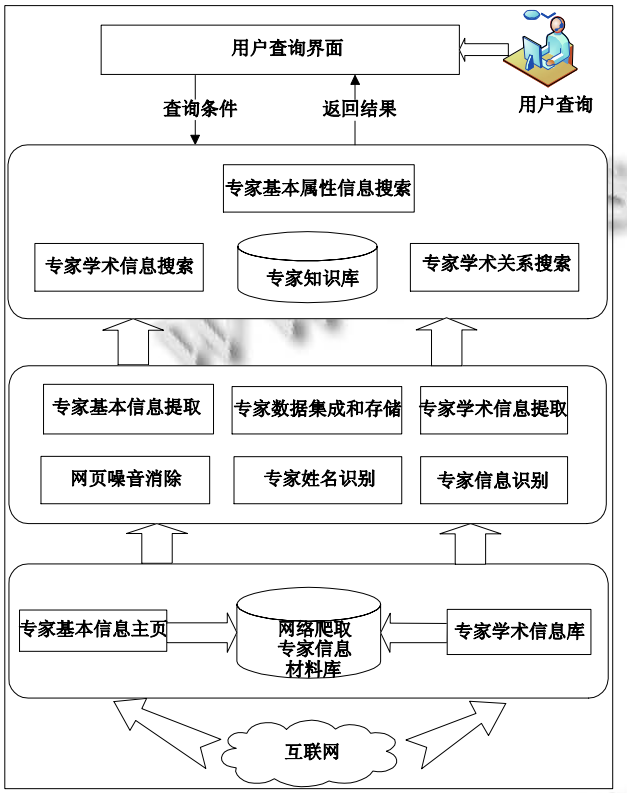

图 1 科技专家搜索系统总体框图

专家信息采集模块：信息采集模块采用网络爬虫根据一定的条件采集含有专家信息页的链接或文本内容，收集含有专家信息的专家个人网页，分析专家个人网页的内容，构建专家材料库. 采集的科技专家个人页面或者专家相关介绍页面，通常包含专家特定的基本属性，如姓名、职称、教育背景、联系方式、所在机构、E-mail、研究方向等，专家主页中通常还包含专家相关的学术论文，而学术论文信息中含有专家相关的学术信息，如论文作者、合著作者、作者所在机构、摘要、关键词等. 根据上述特定字段，确定筛选专家个人页面的条件，筛选出专家主页的链接. 这一模块是专家属性信息抽取的准备阶段.

专家数据处理模块：专家数据处理模块针对采集的专家个人页面的文本信息数据进行处理，主要包括网页噪音消除、结构化信息分离，专家主页文本内容的解析即专家信息识别、专家基本信息提取、专家学术信息提取，该模块还包括对专家数据的集成和存储等，根据专家姓名集成提取的专家信息和学术成果，并存储到一个专家知识库. 本系统采用数据库存储和索引的方法提供了存储和索引的接口以方便向专家知识库中存储和集成数据. 从自然表达的文本内容中抽取出结构化的专家信息是实现专家搜索系统的关键部分.

专家信息查询与检索模块：专家信息查询与检索服务模块提供专家搜索、专家学术信息搜索、专家学术网络关系搜索等个性化服务. 为了具有更友好的服务特性，科技专家搜索系统除了提供用户科技专家信息集成查询结果外，还形成可视化专家学术关系图，为用户提供更清晰更加明了的专家信息检索和展现.

实现上述功能模块在很多方面都具有很大的挑战性. 诸如需要解决采用什么方法抽取专家的基本属性信息，如何根据专家发表的学术论文和成果构成专家的学术信息，以及如何从学术信息中获取专家的学术关系. 基于以上考虑，本文提出了一系列方法，以专家为对象进行信息组织，构成专家信息模型，建立结构化的专家数据信息，来实现专家的多维信息模块建立.

3 各模块的关键技术与实现

3.1 专家属性信息

研究过程中我们发现，85%以上的科技专家主页在高等院校的网页上，部分是在某些研究中心的专家介绍页面上，我们从中随机抽取了 1000 名科技专家主页或者专家介绍性页面并经过网页噪音过滤，去除没用的链接等噪音，通过研究专家网页的内容从中发现 70%以上的专家信息是以自然语言文本的形式给出，部分是以列表的形式给出信息. 因此，这就要求我们设计一个通用的信息抽取方法来完成专家信息属性值的抽取. 基于互联网的语料规模大，系统效率要求高，专家姓名和专家属性信息表达方式多样化，本系统采用监督式机器学习的方法来抽取专家基本信息，下面重点介绍本系统专家基本信息抽取的方法.

专家名字识别方面，我们直接使用汉语词法分析系统 ICTCLAS，该系统是中国科学院计算技术研究所研制的一个集分词、词性标注、未登录词识别、词

性标注于一体的汉语词法分析系统, 其中基于角色标注的中国人名自动识别方法, 人名识别的正确率和召回率分别达到 95.57%和 95.23%[8].

专家信息抽取是提取专家属性信息的过程, 首先我们定义了要抽取的专家属性信息, 如图 2 展示了科技专家属性信息.

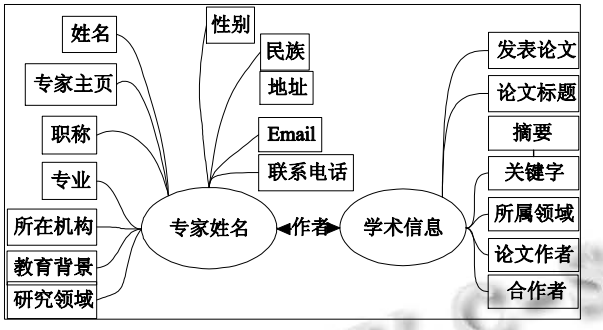

图 2 科技专家属性信息

3.2 专家基本信息抽取的方法和实现

系统专家信息抽取方法和流程如图 3 所示.

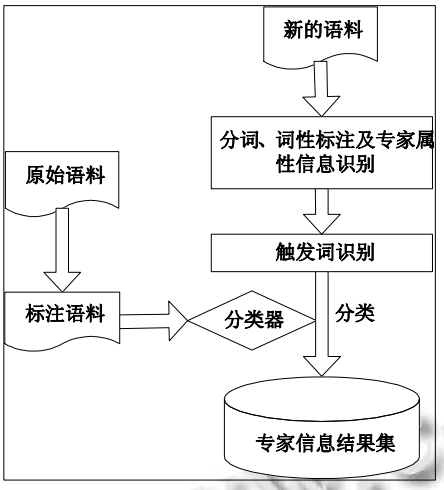

图 3 专家属性信息抽取方法和流程

3.2.1 分类器语料集的标注

由图 3 所示, 我们使用 ICTCLAS[ICTCLAS2011]分词系统进行词法分析, ICTCLAS2011 版可以自己添加语料集. 首先我们构建专家属性信息抽取的标注语料集, 然后从标注的语料中训练出分类器, 提取出对应的专家基本信息和学术信息, 分类器采用基于语义相似度计算的特征向量形式. 系统首先采用 ICTCLAS 词法分析软件对专家基本属性信息进行标记, 建立专家属性信息词典. 在构建标注语料集过程中, 我们对专家主页上的文本内容进行标注, 即手工标注出一定规模的专家基本属性信息语料, 每一个标注包含所有可能的属性标记, 然后根据训练有素的标记语料模型进行分类, 确定标记所对应的图 2 中专家信息属性, 如“中国工程院院士/jobtitle”, 对应专家属性标记的“职称”.

3.2.2 分类器的模型表示

在构建分类器模型的过程中借鉴文献[9]中使用的分类器模型, 即分类器采用特征向量的表示方法, 特征向量表示法主要是把对象实例数值化成向量形式, 然后对于给定的一组训练数据 $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$, 其中 $y_i \in \{-1,+1\}$,学习一个分类函数 f, 使得对于给定的新的特征向量 x, f 能够正确地分类, 即 $f(x)= y$. 因此, 我们把专家姓名和属性看作是一个二元关系<x, y>, 专家姓名和属性周边的词语和符号构成了特征向量的内容.

在分类器训练的形成阶段, 我们对特征向量词场中的词引入语义相似度计算方法, 借鉴文献[5]提供的“基于《知网》的词汇语义相似度计算”方法, 计算词与词之间的语义相似度. 该方法采用了“整体的相似度等于部分相似度加权平均”的做法. 首先将一个整体分解成部分, 再将两个整体的各个部分进行组合配对, 通过计算每个组合对的相似度的加权平均得到整体的相似度. 对于两个词语的相似度, 该方法采用根据上下位关系得到语义距离并进行转换的方法. 对于两个汉语词语 w_1 和 w_2, 如果 w_1 有 n 个概念表述: $S_{11}, S_{12}, \cdots, S_{1n}$, $w2$ 有 m 个概念表述: $S_{2}1, S_{22}, \cdots, S_{2n}$, 则 w_1 和 w_2 的相似度是各个概念的相似度之最大值, 用公式表示如下:

$$Sim(W_1,W_2) = \max_{i=1...n,\, j=1...m} Sim(S_{1i}, S_{2j})$$

3.2.3 分类器的训练与测试

本系统的分类器采用简单高效、训练与分类速度快的中心分类法[6]. 这样, 我们就从训练语料中训练出专家性别、籍贯、住址、教育背景、职称、研究领域、所在机构等中心向量和各自最佳的阈值以及数据集中专家姓名和基本信息所间隔的最大窗口. 我们在系统中采用“和中心”来表示中心向量, 即 $\overline{C_i^s} = \sum_{de C_i} \overline{d}$, 决策规则为:

$$C = \arg\max_{C_i} \left(sim\left(\overline{d}, \overline{C_i}\right)\right) = \arg\max_{C_i} \left(\overline{d}, \overline{C_i}\right)$$

选取训练样本中离中心最远的距离作为评判测试

样本的阈值; 选取训练样本中专家姓名与专家属性信息所间隔的最大词数作为该属性信息的窗口. 表 1 是专家基本信息训练语料中间隔的最大窗口长度.

表 1 专家基本信息属性项对应的窗口长度

专家信息属性项	窗口长度
性别	4
籍贯	8
地址	8
教育背景	16
职称	40
研究领域	45
所在机构	45

在测试阶段首先使用现有的词法分析软件 ICTCLAS 进行词法分析, 然后建立一部触发词库从未经标注的语料中发现专家的基本属性信息. 对于一个新的句子, 先对其进行分词、词性标注和专家属性信息识别, 然后解析该句子, 如果发现满足某触发条件的词语, 则开始抽取专家基本属性信息并归类.

3.2.4 分类器产生的结果及分析

当寻找出专家姓名, 并在词语满足触发条件的时候, 识别出专家属性最大的短语及处于并列地位的属性短语. 对于任一文本信息短语, 如果文本信息存在着专家属性信息, 则可以挖掘出对应的特征向量. 然后与训练好的分类器中的语义向量进行相似度运算. 这样计算出该属性短语对应的向量相似度, 其中较大的, 且超过阈值, 则该属性信息就属于该属性.

测试结果如表 2 所示, 表中 P 表示专家基本属性信息提取的准确率, R 表示召回率, F 测度综合以上两个标准, 结果表明平均准确率达到了 85%以上. 实验证明, 分类器中引入语义相似度计算, 能有效提高专家属性信息抽取的效果.

表 2 专家基本信息抽取效果

信息属性	测试集大小	P(%)	R(%)	F(%)
地址	95	0.925573	0.836467	0.878767
性别	96	0.915367	0.879326	0.896984
籍贯	94	0.883129	0.855768	0.869233
教育背景	96	0.909283	0.860691	0.884319
职称	93	0.887351	0.856253	0.871525
研究领域	95	0.876532	0.833216	0.854325
所在单位	98	0.897358	0.827851	0.861204

3.3 专家网络关系提取

本文按照专家学术论文信息及相互之间合作关系特点, 将专家关系分为: $x\in${导师, 学生, 合作}, x 表示某类特定关系, 下面为本文根据学术论文建立专家关系的抽取方法的具体步骤:

步骤 1: 关系特征词提取. 对于定义好的专家关系类别, 提取各个类别的关系特征词, 例如具有合作关系的关系特征词同事、校友、合伙人、搭档、合著、合作等, 建立关系类别的关系特征词库并记录下来.

步骤 2: 提取专家姓名. 在专家论文信息中, 会有若干专家姓名与某一确定专家姓名同时出现在同一论文中, 因此, 我们需要将与该确定专家共现的专家提取出来, 作为相关专家列表, 再进一步确定专家列表中的特定关系专家的姓名. 然后将确定专家姓名和关系类别特征词组合作为关键词利用搜索引擎查询, 在返回的信息中提取所要的专家的姓名, 以此构造相关专家姓名列表: $EL(x, y)$: x 表示某类特定关系, $x\in${导师, 学生, 合作}, $y\in\{1, 2, 3, \cdots\}$, y 值对应不同的相关人名.

步骤 3: 提取专家网络关系. 提取专家姓

名列表后, 针对每一类专家关系, 将专家姓名列表, 相关专家姓名列表, 关系特征词及专家基本信息结合起来, 利用其在搜索引擎返回文档中的结果来确定专家网络关系. 例如根据关系特征词提取专家姓名及相关专家姓名列表后, 然后再以专家基本信息作为相关条件进行判别, 例如合著者之间, 根据两人的职称、学历、职位关系等来进一步判别是学生还是合作关系.

结果分析: 针对本文研究的专家关系, 基于《同义词词林》构造关系特征词集, 然后用测试数据构建专家姓名列表, 提取相关专家姓名列表, 最后提取专家关系. 实验结果如表 3 所示, 表中 P 表示专家的准确率, R 表示召回率, F 测度综合以上两个标准, 反应了专家

关系提取方法的总体效果，本文的方法在准确率达到 86%以上 .

表 3 专家关系抽取的总体效果

信息属性	测试集大小	P(%)	R(%)	F(%)
导师	96	89.3	87.2	88.2
学生	94	87.2	85.6	86.8
合作	95	87.5	86.3	86.9

4 结语

本文主要描述了科技专家搜索系统的体系结构和主要特征，提出了专家信息抽取和专家关系提取的方法，实验结果表明，本文专家信息抽取和专家关系抽取的方法的平均准确率都达到了 85%以上，具有实际应用价值. 利用此方法建立的科技专家系统已经进入应用阶段，并且为用户提供个性化的专家信息检索服务，用户使用过程中缩小了查寻范围，快速、有效地在浩瀚的网页中搜索到科技专家相关信息. 本文研究和设计的科技专家搜索系统，具有非常深远的理论意义和巨大的实际应用价值.

参考文献

1 Craswell N, de Vries AP, Soboroff I. Overview of the trec-2005 enterprise track. TREC'05. 2005: 199–205.

2 Balog K, Azzopardi L, de Rijke M. Formal models for expert finding in enterprise corpora. Proc. of SIGIR'06. 2006: 43–55.

3 Nie Z, Ma Y, Shi S, Wen JR, Ma WY. Web object retrieval. Proc. of WWW'07, 2007: 81–90.

4 Kautz H, Selman B, Shah M. Referral web: Combining social networks and collaborative ltering. Communications of the ACM, 1997,40(3):63–65.

5 Liu Q, Li SJ. Lexical semantic similarity computation based on HowNet. Computational Linguistics and Chinese Language Processing, August 2002,7(2):59–76.

6 Han E, Karypis G. Centroid-Based Document Classification Analysis & Experimental Result. PKDD 2000.

7 Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge University Press, Cambirdge University, 2000.

8 张华平,刘群.基于角色标注的中国人名自动识别研究.计算机学报,2004,27(1).

9 于满泉.面向人物追踪的知识挖掘研究.北京:中国科学院计算技术研究所,2006.

10 Miller S, Fox H, Ramshaw LA, Weischedel RM. A Novel Use of Statistical Parsing to Extract Information from Text. Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, WA: Association for Computational Linguistics, 2000: 226–233.

11 Che WX, Liu T, Li S. Automatic Entity Relation Extraction. Journal of Chinese Information Processing, 2005,19(2):1–6.

12 Matsuo Y, Mori J, Hamasaki M. POLYPHONET: an advanced social network extraction system from the web. Proc. of the 15th International Conference on World Wide Web. ACM Press, New York, NY, 2006: 397–406.

13 Zhang SX, Wen J, Qin Y, et al. Study about automatic entity relation extraction. Journal of Harbin Engineering University, 2006,27(S):370–374.

14 Zhao P, Geng HT, Cai QS. An approach of Chinese text representation based on semantic and statistic feature. Journal of Chinese Computer System, 2007, 28(7): 1311–1313.

15 Grishman R. Information extraction: Techniques and challenges. MT Pazienza, editor, Information Extraction. Springer-Verlag, LNAI. 1997.

16 Xun E, Huang C, Zhou M. A unified statistical model for the identification of english basenp. Proc. of ACL'00. 2000.

17 庞剑锋,卜东波,白硕.基于向量空间模型的文本自动分类系统的研究与实现.计算机应用研究,2001,9(9):36–30.