第 36 卷 第 7 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.7 Computer Engineering April 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)07—0255—02 文献标识码：A 中图分类号：TP391

一种专家地图构建方法

王树锋，胡智喜
(常州工学院计算机信息工程学院常州市软件技术与应用重点实验室，常州 213002)

摘 要：专家地图是专家个人知识和技能信息及专家协作网络信息的记录。基于图论的方法给出专家发现社会协作网的算法。定义一种从企业局域网的异构数据文档中发现专家地图的任务，给出实现该任务的多种模型，并将模型应用于 TRCKMS 系统的开发中。在 W3C 语料库上进行的实验表明，该模型能够提高专家发现的效率。

关键词：专家地图；社会协作网；知识管理系统；专家发现

Expert Profile Construction Method

WANG Shu-feng, HU Zhi-xi
(Changzhou Key Laboratory of Software Technology and Applications, School of Computer Information Engineering, Changzhou Institute of Technology, Changzhou 213002)

【Abstract】The expert profile of an individual is a record of the types and areas of skills of that individual and a description of his collaboration network. This paper gives a social profile algorithm of finding experts. It defines the task of automatically determining an expert profile from a heterogeneous corpus made up of a large organization's intranet, proposes multiple models for addressing the profiling task, and applies them to TRCKMS. An experiment based on the W3C-corpus shows a significant improvement on the efficiency of finding experts.

【Key words】expert profile; social collaboration network; knowledge management system; expert find

1 概述

在开发盾构施工风险控制知识管理系统(Tunnel Risk Control Knowledge Management System, TRCKMS)中，实现一个发现专家的功能，目的是根据用户输入的信息，在后台执行一个搜索算法快速查询某专家的研究领域及与该专家有协作关系的其他专家，选择其中最合适的专家，咨询解决现场技术人员无法处理的风险控制问题。

本文将上述问题形式化为从企业局域网的异构数据文档中发现专家地图，提出多种模型实现这个任务，并把这些模型应用到 TRCKMS 系统开发中。

2 专家发现系统

最初的专家发现系统是检索企业员工知识和技能数据库的过程，系统的重点是将分散的、异构的员工知识和技能数据库整合成一个数据仓库进行数据检索[1-2]。后来的专家发现系统主要在企业局域网的异构数据文档中自动抽取专家信息[3]最近几年，专家发现系统仍然主要用于工业领域，解决特定组织提出的要求。文献[4]于 2005 年引入专家发现的任务，为研究者提供了通用的平台用于评价专家发现的方法和技术。

3 专家发现方法

在开发专家发现功能时，本文提出的任务是：根据用户输入的专家信息，在后台执行一个搜索算法快速准确地查找到专家的研究领域及与该专家有协作关系的其他专家。

3.1 专家地图

专家地图是一条记录，描述专家拥有的特定的领域知识或特有的专业技能及能证明具有这些知识和技能的支撑材料。可以用向量表示专家地图，向量中的每一个元素表示某领域知识或专业技能，元素值的大小表示相应能力的高低。

3.2 构建专家地图的算法

发现专家地图有 2 个步骤：发现并确定专家可能具有的领域知识或专业技能；计算得到与此对应的表示能力大小的度量值。

一般某专家具有的知识领域或专业技能是确定的，表示为 $KA=\{ka_i \mid i=1,2,\cdots,n\}$，主要实现发现专家的第 2 个步骤。某专家 ca 的专家地图是一个具有 n 个元素的向量，向量中第 i 个元素表示专家的知识领域或专业技能 ka_i，该元素值的大小表示专家在该领域具有的能力大小 $score(ca,ka_i)$。

$$profile(ca)=\langle score(ca,ka_1),score(ca,ka_2),\cdots,score(ca,ka_n)\rangle$$

文献[5]提出的专家发现方法，是通过计算某个知识领域或专业技能 q 中，候选专家 ca 是专家的概率 $p_{ES}(ca \mid q)$ 来选择专家的。假设在知识领域或专业技能 q 中，候选专家 ca 的排列顺序表示为：$rank_{ES}(ca,q)=1,2,\cdots,m$，则候选专家的排列顺序与 $p_{ES}(ca \mid q)$ 有如下关系：

$$p_{ES}(ca_i \mid q) \geqslant p_{ES}(ca_j \mid q) \Rightarrow rank_{ES}(ca_i,q) \leqslant rank_{ES}(ca_j,q)$$

因此，可以用概率值的大小或排列顺序值的大小描述某候选专家具有的知识能力，即：

$$score(ca,ka)=p_{ES}(ca \mid ka) \text{ 或 } score(ca,ka)=1/rank_{ES}(ca)$$

用这种方法确定的专家能力是相对能力，即这种能力不是某个候选专家在某个知识领域的绝对能力，因此，在给定的知识领域，如果候选专家排序靠前，则意味着该候选专家

基金项目：国家自然科学基金资助项目(50778109)；常州工学院校级基金资助项目(YN0818)
作者简介：王树锋(1968－)，男，副教授、博士，主研方向：智能计算，知识科学；胡智喜，讲师
收稿日期：2009-12-25 **E-mail：**wangsf@czu.cn

具有较高的知识能力。用上述方法计算得到的结果作为基准值来评价本文提出的专家地图算法的性能。

方法 1

一个候选专家的知识能力可以用该候选专家与特定知识领域中公开出版的文章的相关度来描述。对于知识领域 ka，以 ka 作为关键字进行搜索，从检索结果中选取排在前面的 n 篇文章 D_{ka}。在这些文章中统计出与该候选专家有关的文章数，通过计算能够得到特定知识领域 ka 中候选专家 ca 具有的知识能力，表示为

$$score(ca,ka)=\sum_{d\in D_{ka}} relevance(d,ka)A(d,ca)$$

当候选专家 ca 的名字出现在文章 d 中时，$A(d,ca)$ 为 1，否则为 0。为了简化计算，这里不考虑该候选专家对这篇文章的贡献，即不考虑在文章中的排名。

用标准的语言模型(language modeling)[6]计算特定知识领域中文章的相关度。

$relevance(d,ka)=p(ka\mid\theta_d)$ 表示特定知识领域 ca 中文章 d 出现的可能性。

$p(q\mid\theta_d)=\prod_{t\in q}\{(1-\lambda)p(t\mid d)+\lambda p(t)\}$ 考虑了背景模型 $p(t)$ 和每一个术语 $t\in q$ 的估计值 $p(t\mid d)$，通过平滑处理，线性组合得到该语言模型。

方法 2

候选专家的知识能力也可以通过比较候选专家具有的与特定知识领域有关的关键字的相似程度来确定。

对于每一篇文章，首先选择分词，去掉停用词，利用 $TF\cdot IDF$ [6]计算，选出前 20 个词作为该文章的关键词，从文章 d 中抽取的一组关键字表示为 $KW(d)$。

对于知识领域 ka，获取其关键字的方法是：以 ka 作为关键字搜索 W3C 语料库，从检索结果中选取排在前面的 n 篇文章 D_{ka}。组合这些文章中每一篇的关键字得到作为知识领域 ka 的关键字，表示为

$$KW_{ka}=\bigcup_{d\in D_{ka}} KW(d)$$

用同样的方法，可以得到与候选专家有关的关键字集合，表示为

$$KW_{ca}=\bigcup_{d\in D,A(d,ca)=1} KW(d)$$

特定知识领域和特定候选专家的关键字共现率的大小就是该候选专家在特定领域中的知识能力，表示为

$$score(ca,ka)=\mid KW_{ka}\cap KW_{ca}\mid/\mid KW_{ka}\mid$$

利用上述 2 种方法，能够发现众多的专家。为了保证质量，采用一种优化的方法，选出最合适的专家。该优化方法的直观想法是提供一个门槛值，将发现的数量众多的专家按某种原则选出真正的专家。该原则描述为：一个知识领域是专家地图的一部分，当且仅当该专家在这个知识领域中的排名位于前列，即所谓专家就是具备某个领域的知识，同时比其他同行有更多知识的候选专家。实现这个原则，实际上就是修正候选专家知识能力的计算方法，表示为

$$score'(ca,ka)=\begin{cases} score(ca,ka) & \text{if } \mid\{ca'\mid score(ca',ka)<score(ca,ka)\}\mid<f \\ 0 & \text{other}\end{cases}$$

其中，f 是该领域中所有候选专家中排在最前面的数值。

3.3 实验结果

用 W3C 语料库进行实验。该语料库大小约为 5.7 GB，有 330 037 篇文档和 1 092 个候选专家。每个专家有唯一的 id 号、姓名、一个或多个 E-mail。本实验采用 2005 版 TREC 中指定的 50 个知识领域发现专家。

在 TREC2005 中，标题实际上是 W3C 工作组的名称，专家是在对应标题下工作组的成员。一个人是多个工作组的成员，意味着他是多个标题下的专家。用 W3C 工作组名作为知识领域，如果一个人是工作组的成员，则这个知识领域就是专家地图的一部分。实验结果见表 1。从表中可以看出，在 *MAP* 和 *MRR* 2 个指标上，方法 1 和方法 2 的性能好于测试基准值，其中，方法 1 性能更好些。

表 1　发现专家方法比较

方法	*MAP*	*MRR*
概率方法基准值	0.370	0.387
排名方法基准值	0.253	0.234
方法 1 结果	0.457	0.493
方法 2 结果	0.447	0.476

优化方法的实质是在候选专家中遴选排在最前面的 f 个专家，f 越小，意味着包含的专家越少，这些专家也越符合要求。f 选多大，选取这样大小的 f 会滤掉多少可能的专家，表 2、表 3 给出了相应的实验结果。其中，“优化”表示从原始搜索结果中过滤掉的比例；“错误率”表示不正确过滤的比例。从表中可以看出，选择不同的 f，方法 1 和方法 2 在性能上都得到改进。

表 2　方法 1 采用过滤技术后发现专家方法比较

方法	执行专家发现方法的结果		采用优化技术后性能的改变	
	MAP	*MRR*	优化	错误率
方法 1	0.457	0.493	-	-
f=100	0.445	0.501	0.518	0.101
f=50	0.445	0.525	0.674	0.156
f=30	0.442	0.548	0.836	0.221
f=20	0.438	0.667	0.902	0.256
f=10	0.458	0.654	0.945	0.319
f=5	0.402	0.649	0.976	0.387

表 3　方法 2 采用过滤技术后发现专家方法比较

方法	执行专家发现方法的结果		采用优化技术后性能的改变	
	MAP	*MRR*	优化	错误率
方法 2	0.447	0.476	-	-
f=100	0.355	0.465	0.625	0.058
f=50	0.396	0.502	0.818	0.067
f=30	0.267	0.476	0.907	0.085
f=20	0.234	0.461	0.965	0.108
f=10	0.229	0.432	0.987	0.132
f=5	0.256	0.453	0.991	0.186

4　专家协作发现算法

专家协作，也称为社会协作网。协作网络有助于进一步确定该候选人在一个“圈子”内的学术地位；他仅是一个能力很强的专家，还是一个“图书馆式”的人才，不仅具有丰富的专业知识，而且还可以指导人们在更宽广的领域中找到所需要的专家。

协作网络 CN 是有向图 (V,E)，其中，节点表示个人；边 $(x,y)\in E$ 表示个人之间的关系，边上的权重表示个人 x 与个人 y 之间的协作强度。专家协作网络 $CN(t)$ 是一种特殊的 CN，用来描述特定知识领域 t 中专家之间的协作关系。

构建专家协作网 $CN(q)$ 的算法描述如下：

初始条件：文章集 D_q

第 1 步 计算每个候选专家相关度。

$$\forall x\in V: R(x,q)=\sum_{d\in D_q\wedge A(x,d)} relevance(d,q)$$

第 2 步 计算候选专家之间的协作权重。

$$w(x,y)=\sum_{d\in D_{xy}} relevance(d,q)$$

第 3 步 连接专家协作网中的有向边，规范化边上的权重，允许存在环。

（下转第 259 页）