整合搜索引擎结果的专家检索

毕文静 沈华伟 刘悦 许洪波 程学旗
（中国科学院 计算技术研究所, 北京 100190）
E-mail: biwenjing@software.ict.ac.cn

摘要： 目前，专家检索的研究多基于企业内部语料，较少引入外部资源。万维网中包含丰富的信息。本文以现有的两个典型专家检索模型为基础，根据模型的特点，整合搜索引擎检索结果与企业内部语料，实现专家检索。实验证明搜索引擎检索结果的引入能够很好的改善专家检索的效果。

关键字： 专家检索、专家查找、搜索引擎结果

Combing Intranet Evidence and Search Engine Results for Expertise Retrieval

Wenjing Bi, Huawei Shen, Yue Liu, HongboXu, Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences

ABSTRACT

Expertise retrieval in enterprise has been largely explored on the collections crawled from the intranets of enterprises or organizations. However, only limited external information has been used to help improve expertise retrieval. As the World Wide Web (WWW) has abundance information, we attempt to combine such information with intranet evidence for expertise retrieval. In this paper, different strategies are adopted to extract search engine results for two different widely-used models based on language model. Experimental results demonstrate that the introduction of search engine results can significantly improve expertise retrieval.

Key words: expertise retrieval, expert finding, search engine results

1. 引言

企业内的专家检索（Expertise Retrieval，ER）是当今信息检索领域的一个热点。它作用于海量数据之上，针对用户的输入，返回企业内的专家。这些专家具有一定的专业知识或技能，能够帮助用户解决他们所遇到的问题。从 2005 年开始，为了鼓励专家检索的发展，文本检索会议(Text REtrieval Conference, TREC) 引入企业搜索，并将专家检索作为其中的一个重要子任务[1]，其间，出现了许多可用的模型和方法。

目前，企业背景下专家检索的研究策略大体包含以下几个方面：1） 通过文档为专家建立全面的模型，在此基础上确定专家与查询之间的关系。K.Balog 博士在 2006 年提出的专家模型[2]是其典型代表。它建立在语言模型的框架之上，根据文档与专家的关系，为每个专家建立语言模型 θ_{ca}，并在模型基础上进行检索。2007 年，清华大学综合专家的各方面特点，提出了 CDD 模型[3]，其核心与专家模型相似。同时， Fang 等人在概率模型的基础上将本策略定位成基于专家描述的策略[4]。2）将文档作为联接专家和查询的纽带，即首先确定文档和查询的相关性，根据专家和文档的连接关系，得到专家和查询的相关性。其代表是与专家模型同时提出的文档模型[2]。此外 Craig Macdonald 等人提出的投票模型也基于这种策略[5]，该模型引入数据融合技术，将专家的排序问题转化为投票问题，票的权重由专家在文档中出现的位置以及专家出现和查询出现的距离来确定。3）充分利用链接关系。如利用专家和文档

本课题受国家重点基础研究计划（973）课题“大规模文本内容计算（2004CB318109）”和国家高技术研究发展计划（863）项目“网络文本的倾向性分析（2007AA01Z441）”资助

的 Pagerank 值等，或模拟知识在文档和专家中传播的过程。Serdyukov.P 在[6]中提出的相关性传递模型正是基于这种策略。它将用户查找专家的过程看作是用户在专家文档关系图上游走的过程，充分利用文档与文档、专家与专家、文档与专家、专家与文档之间的关系来对用户的行为进行模拟，并提出了 4 种传递模型[7]。以上对专家检索的研究都是基于企业内部语料，近期部分学者开始引入外部信息改善专家检索的效果。

Troy 等人使用 WordNet 对查询词进行扩展，识别查询词的同义词，虽然 WordNet 中包含的专家信息较少但是仍然有效[8]。Chu-Carroll 等人采用了 GoogleScholar 中的信息来优化专家检索[9]。GoogleScholar 是垂直搜索引擎的典型代表，它可以检索海量网页和部分学术数据库，且这些资源都是免费的。Chu-Carroll 等人首先通过查询 GoogleScholar 得到作者（专家）列表，并记录文献和引用信息。通过出版文献的质量、文献被引用的情况以及专家在作者列表中的位置确定作者（专家）对知识的掌握程度。Jiang 等人在[10]中，在外部搜索引擎的检索结果上进行专家检索。他们为每个候选专家建立查询，在商业搜索引擎中进行检索，并将检索结果作为原始语料进行专家检索，且发现，该检索结果与企业内专家检索的结果相差不大。在 Jiang 的工作中，只是在搜索引擎的结果上进行了专家检索，对结果并没有进一步利用。本文的工作重点将集中在引入搜索引擎结果，与企业语料结合，改善专家检索的效果。

商业搜索引擎作用于万维网中的数据，这些数据不仅包含普通信息，而且包含丰富的个人信息，且可以免费获取。如果能够充分利用这些个人信息，将有利于专家知识和技能的确定。本文整合了搜索引擎的结果，并将问题着眼于以下几个方面：1）引入搜索引擎结果能否改善专家检索的结果；2） 针对某些特殊查询，引入搜索引擎结果的检索效果如何；3）搜索结果的引入对不同模型的影响是否相同。

本文结构安排如下：第二章简单介绍专家检索中的两个典型模型，专家模型和文档模型，两者也是本文的工作基础；第三章详细介绍本文的方法；第四章在标准数据集上进行实验并对实验结果进行分析；最后总结全文进行展望。

2. 专家检索模型

目前，针对专家检索问题，众多学者已经提出了很多有效的模型，其中专家模型和文档模型是提出最早也是应用最广泛的两个模型，它们由 K.Balog 博士于 2006 年同时提出[2]。两者分别从专家为中心和文档为中心两个角度解决专家检索问题，且都可在语言模型上得到理论支持。本文将以这两种模型为基础，引入搜索引擎结果改进专家检索。本节将对这两个模型进行简单介绍。

在语言模型的基础上，专家检索问题可以描述为：给定一个查询话题 q，判断某个候选专家 ca 与该话题 q 相关的可能性 $p(ca|q)$，并进行排名。根据 Bayes 法则，在查询已给定的情况下，专家检索问题可以转化为如何确定 $p(q|ca)$ 的问题。在确定 $p(q|ca)$ 的过程中，专家模型和文档模型采用了不同的策略。

专家模型为每个候选专家 ca 建立模型 θ_{ca}，由此，$p(q|ca)$ 的大小可以用 θ_{ca} 生成查询 q 的概率 $p(q|\theta_{ca})$ 来表示，其中 $p(q|\theta_{ca})$ 的值可通过公式（1）得到：

$$p(q|\theta_{ca})=\prod_{t\in q}p(t|\theta_{ca})^{n(t,q)} \quad (1)$$

t 为查询 q 所包含的单个词项，$n(t,q)$ 表示词项 t 在查询中出现的次数。为了获取 $p(t|\theta_{ca})$，专家模型使用 $p(t|ca)$ 对 $p(t|\theta_{ca})$ 进行模拟，并引入词项在所有文档集中出现的概率 $p(t)$ 对其进行平滑：

$$p(t|\theta_{ca})=(1-\lambda_{ca})p(t|ca)+\lambda_{ca}p(t) \quad (2)$$

其中，λ_{ca} 为平滑系数。由于专家的知识散布在文档之中，$p(t|ca)$ 可表示为：

$$p(t|ca)=\sum_{d\in D}p(t|d,ca)p(d|ca) \tag{3}$$

其中，$p(d|ca)$ 为给定某专家，生成某文档的概率，$p(t|d,ca)$ 为给定专家和文档后，生成某查询词项的概率。

文档模型将文档看成连接查询和专家的纽带：

$$p(q|ca)=\sum_{d}p(q|d)p(d|ca) \tag{4}$$

其中，$p(q|d)$ 为某文档 d 产生查询词 q 的概率。为了计算 $p(q|d)$，文档模型为每个文档构建模型 θ_d（文档模型的命名也来源于此）。假定查询中的词项相互独立，并引入 Jelinek-Mercer 平滑后，得到：

$$p(q|\theta_d)=\prod_{t\in q}((1-\lambda_d)p(t|d)+\lambda_d p(t))^{n(t,q)} \tag{5}$$

其中，t 表示单个查询词项，λ_d 为平滑系数，$n(t,q)$ 为查询词 t 在查询 q 中出现的次数。

两模型的实现流程见下图：

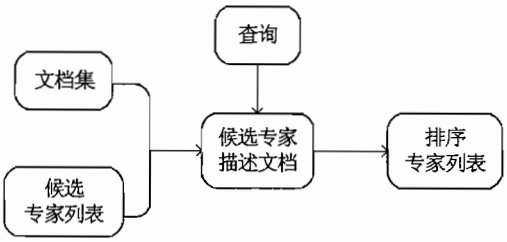

图 1 专家模型流程图

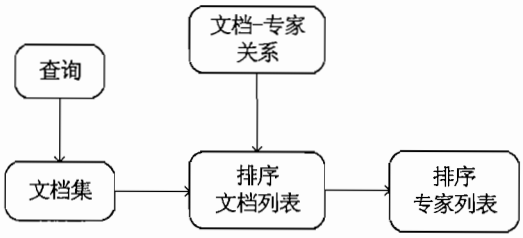

图 2 文档模型流程图

3. 模型和方法

针对上一章提到的两种经典的专家检索模型，本节分别介绍如何利用商业搜索引擎的检索结果提高每个模型的检索效果。大体框架可以分为三个部分：1）从给定的企业数据集中获取候选专家列表；2）根据候选专家的姓名、Email 等信息，构建查询，使用商业搜索引擎获取候选专家的相关文档集，针对两种不同模型，采用相应的策略整合返回的相关文档集以及企业数据集；3）对于每个查询输入，使用两种经典的专家检索模型，返回与查询相关的、按相关度排序的候选专家列表。

3.1 获取候选专家

给定语料（本文中指企业数据集）后，语料中出现的所有人物均视作候选专家。候选专家通常以人名全称、人名简称、Email 地址等形式出现。如何从语料中识别出候选专家是专家检索的重要组成部分，专家列表是否完整和准确直接影响专家检索的效果。由于不同的候选专家可以具有相同的人名（特别是人名简称），获取候选专家也是专家检索的一个难点。

不同的语料中，专家出现的形式不同，因此获取专家列表的方法也不同。有些语料自身提供候选专家列表，如 2005 年和 2006 年 TREC 企业检索任务所使用的 W3C（World Wide Web Consortium）语料。有些语料提供候选专家出现的标准形式，如 2007 年和 2008 年 TREC 企业检索任务所使用的 CERC（CSIRO Enterprise Research Corpus）语料中，候选专家出现的

标准形式是格式为 firstname.lastname@csiro.au 的邮件地址。考虑到 W3C 语料主要以 Email、代码等为主，不能很好地代表企业数据，本文采用 CERC 语料。

针对 CERC 语料，考虑到候选专家的标准形式，我们首先使用正则表达式得到语料中出现的所有可能的 Email 地址，然后，使用一些启发式规则，过滤掉不适合作为候选专家的 Email 地址，例如邮件服务器不是以 csiro.au 为后缀的 Email 地址。过滤后所得到的所有 Email 地址用于识别候选专家。有些候选专家拥有多个 Email 地址，例如 Don Michel 拥有 don.michel@csiro.au 和 don.michel@marine.csiro.au。考虑到这一因素，我们对相似度较大的 Email 地址进行了合并。最后，根据候选专家的 Email 地址，得到候选专家的人名全称、人名简称等出现形式。通过以上策略，我们得到了候选专家列表和相应的不同形式的标识信息。我们最终得到的候选专家个数为 3624 个。

3.2 获取商业搜索引擎检索结果

针对每个候选专家，使用专家的 Email 等标识信息构建查询。查询作用于商业搜索引擎，由于商业搜索引擎的检索对象是整个万维网中的网页，因此，过于宽泛的查询会返回很多不相关页面。在文献[10]中，Jiang 给出了三种可能的构建形式：1)全称加组织名；2）E-mail 地址；3）以上两者的或操作。本文采用第三种方式，即包含专家 E-mail 地址或包含专家全称和组织名的网页都将被返回。这种方式获得的文档最多，信息也最丰富。

输入查询后，搜索引擎返回查询直接结果，每个结果中，包括一个指向结果网页的超链接，一个结果摘要以及 URL 地址等。网页的全文可以在搜索引擎的缓冲区中获得。本文将结果分为两类：引擎摘要（超链接+结果摘要）和网页全文。针对不同的基础模型，对搜索结果的选择也不相同。具体方法我们将在下一节中阐述。

3.3 检索模型

本文采用两种基本模型，分别是专家模型和文档模型，模型的详细介绍见第 2 节。两者分别从专家为中心和文档为中心的思想解决专家检索问题。其中，专家模型将所有与某专家相关的文档按照某种方式组织起来形成该专家的描述文档，并对描述文档组成的描述文档集进行检索，返回的相关文档所表示的专家即为相似专家。文档模型则将文档看成连接查询和专家的纽带，先检索与查询相关的文档，并假定文档中出现的所有人都是候选专家，文档则是相应专家的支持文档，通过整合支持文档与查询的相关度获得专家与查询关联的程度[2]。

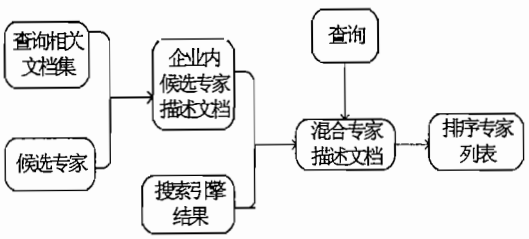

图 3 应用于新语料上的专家模型（专家模型 F)流程图

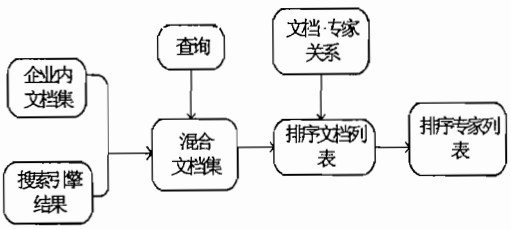

图 4 应用于新语料上的文档模型（文档模型 F)流程图

在实现专家模型时，获得专家描述文档之后，需要为描述文档集建立索引。而实现文档模型时，需要对所有的文档集建立索引。根据两者索引建立的方式不同，我们选择不同的搜索引擎结果。针对专家模型，我们将引擎摘要与专家描述文档集结合作为新的语料集。而针对文档模型，我们将网页的全文加入到文档集中，并在此基础上建立索引。改进后的模型流程见上图 3 与图 4。

4. 实验及结果分析

4.1 实验数据集及评价标准

实验数据集采用由澳大利亚科学与事业研究中心（CSIRO）提供的 CERC 语料，该语料采集于 2007 年 5 月，所有文档都以 HTML 网页的形式给出，大小为 4.2GB，包含 370715 个文档，与 CSIRO 内部搜索引擎使用的文档集相似。我们首先对语料进行了预处理：删除所有的 HTML 标签，并剔除所有不规范的文档，。

专家检索返回的是排序后的人物列表，从检索评价的角度上看，它和文本检索返回的结果相似，因此文本检索的评价指标可以直接应用在专家检索上。本文采用的评价指标主要有 MAP、MRR、RR、P@N 等[11]。

4.2 实验过程及结果分析

首先，在 CERC 的 370715 个文档上，采用第 3.1 节的专家识别方法，得到专家列表。然后对列表进行清洗，删除部分非专家的地址，如 read.read@csiro.au 等，得到最后的 3624 个专家，它们包含了 TREC07 年的所有结果。

第 2 节介绍的专家模型中，通常假定如果某专家在某文档中出现，那么该文档的全部内容都是对该专家技能和知识的描述。在实验中，我们对这个过强的假设进行了弱化，引入窗口策略[12]。即，在一个文档中，文档中的词语与专家的相关性和两者在文档中出现的距离成正比，即，如果一个词出现的位置与专家的位置距离越近，那么它和专家的相关性也越大，反之，如果距离越远，相关性则越小。本文采用的窗口大小为 100，即只有专家前后的 50 个单词可以被加入到描述文档中。

此外，在文档模型中，本文引入多域检索，即在检索过程中，对出现在不同文档域的词项给予不同的权重。本文将文档分为 5 个域，分别是 URL、标题、关键字、锚文本以及文档正文。并为每个域赋予不同的权重，实验结果最理想的权重分配是 3、1、1、3、1。

本文采用 TREC07 的查询，表 1 给出了文档模型、专家模型以及两个扩展模型在 TREC07 下检索的结果。从实验结果中，我们可以发现，文档模型的 MAP 值比专家模型的 MAP 值高出两个百分点。笔者认为这可能是由于专家模型倾向于获得专家的总体描述，对具体查询的敏感度不足造成的。此外，虽然两者的 MAP 值相差较大，但 MRR 值和 P@20 不相上下，这说明，对专家总体的关注不影响专家模型找到排名比较靠前的专家。

表 1 各模型实现结果对比表

模型	MAP	MRR	P@20
专家模型	0.3158	0.4592	0.0590
文档模型	0.3249	0.4579	0.0620
专家模型 F	**0. 3520**	**0. .4995**	0.0660
文档模型 F	0.3445	0. 4729	**0. 0840**

表 2 Topic38 检索结果对比表

专家	Num	MAP	RR	P@20
专家模型	2	0.0198	0.0385	0.0000
文档模型	3	0.1844	0.0769	0.1000
专家模型 F	3	0.3070	**1.0000**	0.1000
文档模型 F	**4**	**0.4583**	1.0000	**0.1500**

在众多搜索引擎中，我们采用 Google 作为本文使用的搜索引擎，因为它的检索效率比较高且检索结果得到了人们的普遍认可。我们对 3624 个专家都进行了检索，并通过一个抓取工具获得检索的结果。对于专家模型，我们只将前 20 个检索结果加入到专家描述文档集中，因为如果加入的结果过多，会使模型对专家的描述产生偏移，而如果加入的结果较少，则达不到扩充专家描述文档的效果。对于文档模型，我们将检索的前 40 个全文加入到 CERC

文档集中，因为即使引入了偏移，第一步检索文档时也会将偏移剔除，而如果引入文档过少，可能会漏掉部分有用信息。根据不同模型的特点，使用相应的方法，整合搜索引擎结果与 CERC 语料，并对新语料建立索引。索引工具使用 Firtex，词根还原使用 Snowball。

从表 1 中，我们可以发现引入搜索引擎结果可以同时改善专家模型和文档模型的检索结果，并且对前者的改善优于后者。引入搜索引擎结果后，专家模型的 MAP 值提高了近 4 个百分点，文档模型的 MAP 值提高了 2 个百分点。笔者认为这可能是由于只依靠内部资源，专家模型很难为每个专家建立完整的模型所造成的。此外，两者的 MRR 值和 P@20 也都有一定的提高，由此，我们可以认为，引入搜索引擎结果，不仅可以提高检索的平均准确率，而且可以有效优化专家检索的排名。

下面对单个检索进行分析，以 Topic38 为例，它的查询词为" managing salinity"，即水盐控制，所有研究水土盐度的专家都应当返回，其检索结果见表 2，其中 Num 表示模型返回的专家数。从表 2 中，我们可以发现，虽然两基本模型都返回了一定数量的专家，但它们的 MAP 值、RR 值以及 P@20 都比较低。笔者认为，这主要是由于返回专家的位置不靠前造成的。观察改进后的两个模型，我们发现，与基本模型相比，MAP 值、RR 值以及 P@20 都有显著提高，尤其是 RR，它的值均为 1，即返回的第一个专家即为相关专家。笔者为分析其原因，将所有的 4 个答案专家输入 Google 进行查询。发现有为数不少的包含有 salinity 关键字的文档被返回。以专家 jeffrey.turner@csiro.au 为例，它的前 5 个文档摘要中就有 2 个包含有 salinity 关键字。

5. 总结和展望

本文基于专家检索的两个典型模型——专家模型和文档模型，引入搜索引擎检索结果，改善专家检索效果。实验表明：引入搜索引擎结果可以有效改善两基本模型的检索效果；引入搜索引擎结果对专家模型的改善多于文档模型。针对单个话题，本文以 Topic38 为例进行简单分析，发现搜索引擎结果的引入对查询的 MAP 值、RR 值以及 P@20 都有显著提高。本文的下一步工作是，详细分析以上模型对每个查询的检索效果，并进行进一步改进。

参考文献

[1] Craswell N, de Vries A, Soboroff I. Overview of the TREC-2005 Enterprise Track[A]. 2005.

[2] Balog K, Azzopardi L, de Rijke M. Formal models for expert finding in enterprise corpora[A]. ACM New York, NY, USA, 2006:43-50.

[3] Fu Y, Xiang R, Liu Y, et al. A CDD-based formal model for expert finding[A]. ACM New York, NY, USA, 2007:881-884.

[4] Fang H, Zhai C. Probabilistic Models for Expert Finding[J]. LECTURE NOTES IN COMPUTER SCIENCE, 2007,4425: 418

[5] Macdonald C, Ounis I. Voting for candidates: adapting data fusion techniques for an expert search task[A]. ACM Press New York, NY, USA, 2006:387-396.

[6] Serdyukov P, Rode H, Hiemstra D. University of Twente at the TREC 2007 Enterprise Track: Modeling relevance propagation for the expert search task[A]. 2007.

[7] Serdyukov P, Rode H, Hiemstra D. Modeling Multi-step Relevance Propagation for Expert Finding[A]. CIKM, 2008.

[8] A.Iroy GZ. Case Western Reserve University at the TREC 2006 enterprise track.[J]. In Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), 2006:

[9] J.Chu-Carroll GA, P.Duboue,D,Gondek,J.Murdock and J.Prager. IBM in TREC 2006 enterprise track.[J]. In Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), 2006:

[10] Jiepu Jiang SH, Wei lu Expertise Retrieval Using Search Engine Results[J]. 2008:

[11] Buckley C, Voorhees E. Evaluating evaluation measure stability[A]. ACM New York, NY, USA, 2000:33-40.

[12] Balog K. People Search in the Enterprise[J]. 2008: 214