基于专长词表的图情领域专家检索与评价*

陆 伟 刘 杰 秦喜艳

摘 要 以武汉大学专家检索系统 WHU-ES 为平台,借鉴基于相关文档集的归并排序法和基于词典进行查询扩展的方法,对图情领域专家检索进行实验与评价,包括专家排序和专长识别两个方面。利用基于词典进行查询扩展的方法对基于相关文档集的归并排序法进行改进,实验结果表明,利用专长词表可以有效地提高专家检索的查准率和专长识别的效果。未来研究中需进一步解决词表的规范性生成问题。图 3。表 3。参考文献 10。

关键词 专家检索 专长词表 专家排序

分类号 G353.4

ABSTRACT The document-combing-based method is one of the common ways in expert search area. However there are still many deficiencies about this approach. In this paper, the way to expand query based on dictionary is used to improve the expert search effectiveness. Through semi-automatic construction of expertise vocabulary and the establishment of algorithm model of experts sorting and expertise suggestion, this paper takes the Expert Search System (WHU-ES) as a platform to carry out expert search experiment in the field of library and information science. The results show that the use of expertise vocabulary can significantly improve the precision and the effectiveness of expertise recognition. 3 figs. 3 tabs. 10 refs.

KEY WORDS Expert search. Expertise vocabulary. Experts sort.

CLASS NUMBER G353.4

1 引言

专家检索的目的是为了快速准确地找到某领域或组织内的专家,以获取帮助,共享其掌握的隐性专长知识等[1],主要利用组织和领域内外部能够表征专家专长的各种文档和资源,识别专家在某给定查询主题(领域)的专长(相关性)程度,并按程度高低排序显示专家结果列表的过程[2]。广义上的专家检索还包括对专家专长识别研究,即给定查询主题(专家名)返回能表征专家专长的专长词的过程。

专家检索的一个核心过程是专家排序。专家排序是利用信息检索模型,对与查询主题相关的专家列表进行排序的过程[3]。目前典型的专家排序方法有两种[4]:基于相关文档集的归并排序法和基于专家档案的方法。前者是先利用传统信息检索方法检索出相关的文档集合,从而对每一个查询得到一组已排序的带有得分的文档,然后进一步利用专家在文档中的特征信息对文档集合与得分采用相关模型进行归并,生成专家得分,最后以专家得分递减顺序排序返回专家列表;后者是通过抽取每位专家高度相关的特征信息生成专家档案,然后计算检索主题词与档案文件的相似度,返回相关排序专家列表。

基于相关文档集的归并排序法是目前研究专家排序最常用的方法。然而,在利用这种方法进行专家检索研究时往往会出现这种问题:在检索词与专家专长的匹配上,只是简单的字词匹配。如检索“信息检索”领域的专家,只会把出现

* 本文系国家自科基金重点项目“基于生命周期理论的数字信息资源深度开发与管理机制研究”(项目编号:70833005)和教育部人文社科规划项目“专家专长智能识别与检索系统实现研究”(项目编号:09YJA870021)研究成果之一。

"信息检索"的文档和文档对应的专家检索出来,即使标识专家专长的文档中出现了"XML 检索"或者"多媒体检索"这样的专长词,也不会对返回的专家结果产生任何影响。出现这种情况是因为,单纯的利用查询主题对文档进行检索往往会忽视词的语义,使得那些与查询内容相同但是表述不同的文档排名靠后或者根本无法被检索出来。而在实际的自然语言里,同一个概念经常会有多种不同的表达方式,并且词与词之间存在着等同、属分以及相关等多种关系。

这个问题的解决方法之一就是基于词典进行查询扩展。其基本思想是根据语义信息、外部语料库统计或是对全部文档中的词或者词组进行分析,计算每对词或词组间的关联程度,生成词表。当需要查询扩展时,将与查询用词关联程度最高的词或词组加入原始查询中以生成新的查询[5]。

这种词表控制机制很早就已经被引入传统的信息检索中,但却很少有人把词表控制技术进一步引入专家检索中,利用词表对查询词进行重构,计算专家得分。本文将以图书情报领域作为研究范畴,尝试把专家专长词表引入专家检索,以改进检索效果。同时,专长词表的作用不仅仅限定于专家检索的查询扩展上,本文将利用专长词表进行专家专长提示的研究。

2 研究思路与方法

2.1 总体思路

借鉴基于相关文档集的归并排序法和基于词典进行查询扩展的方法,本文以武汉大学专家检索系统(WHU-ES)[2]为平台进行图情领域专家检索的实验,包括专家排序和专长识别两个部分。首先,半自动的构建具有三级结构的专家专长词表。建立一级词和二级词,然后利用基于向量的空间模型进行文本聚类构建三级词。其次,通过蜘蛛程序采集已定义的可以表征专家专长信息的搜索引擎和学术数据资源,并生成专家数据集索引文件。然后,提取专家姓名列表归并形成专家列表文件,利用专家数据集索引文件生成专家—文档映射文件(Expert-Docs,该文档记录了数据集中每个专家对应出现的文档列表),利用专家专长词表生成专长—文档映射文件(Word-Docs,该文件记录了数据集中每个专长词对应出现的文档列表)。最后,利用 WHU-ES 的 Web 检索接口,实现可视化呈现专家排序和专家专长提示结果的功能。整个检索过程的体系结构如图 1 所示。

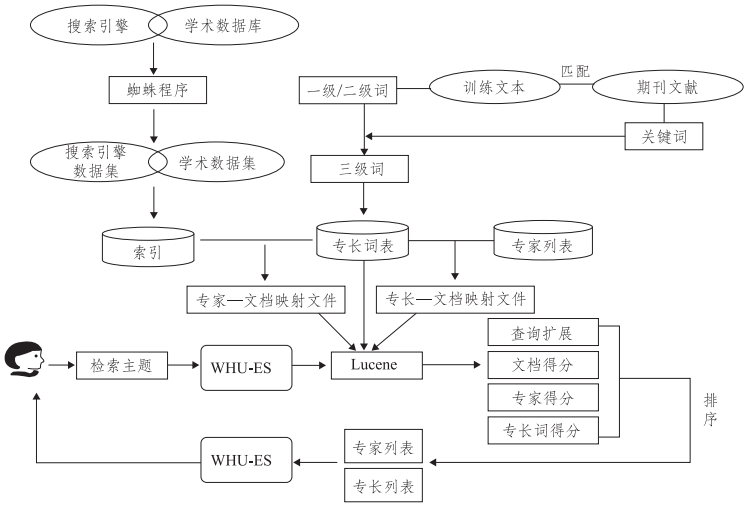

图 1 检索过程体系结构

2.2 词表构建的方法

2.2.1 一、二级词的建立

本文将图书情报领域的研究范围限定于图书馆学、情报学和档案学,改造了管理科学主题词表[6]中的部分主题词,并与中国图书馆分类法[7]中图书馆学、情报学及档案学的分类方法结合起来,形成专长词表的构建基础。这一步主要是构建词表的一、二级类目。表1展现了前两级词汇建立的部分结果。

表1 一、二级词汇部分结果

级别	领域词汇		
一级词	图书馆学	情报学	档案学
二级词	文献学	信息资源管理	档案管理
	主题法	信息组织	文书学
	分类法	信息检索	档案编研
	目录学	信息计量学	档案鉴定
	文献信息编目	竞争情报	档案修复
	数字图书馆	信息分析与预测	电子政务
	图书馆工作	信息咨询与服务	档案自动化
	……	……	……

由于本文的词汇收集过程主要是对三级词,其聚类分析过程也基本是针对三级词进行的,以下将重点介绍三级词的构建。

2.2.2 三级词类目的建立

三级词词汇类目的建立要广泛地收集图情领域的词汇,保证专长词的数量和全面性。一般来说,图书情报领域的学术期刊基本可以代表这个领域最全面和权威的研究领域。因此,本文选取了万方数据库中覆盖图书情报领域的期刊文献库(2000—2008年间30种核心期刊上的所有文献)作为三级词的数据源。然后利用向量空间模型进行文本聚类来确定三级词,步骤如下:

(1)在已经确定的二级词类目的基础上,为每个二级词建立训练文本。在Google Scholar和期刊数据库、图情领域经典著作等中检索和获取与这个二级词相关的、能确切代表这个二级词的文献(被引次数高的权威文献、综述性的文献等)。将这些文献规整并进行文本预处理后作为此二级词的训练文本。

(2)对于已经采集的期刊文献中的每一篇文章,它的关键词比较准确地反映了这篇文章的主题和该作者的研究方向,可以作为候选的三级专长词;摘要经综合并进行文本预处理后作为被匹配的目标文档。

(3)用每个词的训练文本与每一篇期刊文章的目标文档进行相关度计算,获取得分,设定阈值,如果大于设定的阈值,则认为这两者是相关的,进而期刊文档所对应的关键词和训练文本所对应的二级词是相关的,将关键词划入该二级词的类下。

(4)对获取的三级词列表进行去重等处理。在文本相似度计算方面用到的模型是向量空间模型[8],其基本思想是以索引词构成的向量来表示文本:$(w_1, w_2, w_3 \cdots\cdots w_n)$。其中 w_i 为第 i 个索引词的权重。采用以下的TF-IDF公式来计算权值:

$$w_i(d) = \frac{tf_i(d)\log(\frac{N}{n_i}+0.1)}{\sqrt{\sum_{i=1}^{n}(tf_i(d))^2 \times \log^2(\frac{N}{n_i}+0.1)}} \tag{1}$$

$w_i(d)$ 是一个索引词 t_i 在 d 中的权值,$tf_i(d)$ 为索引词 t_i 在文本 d 中的词频,N 为文档总数,n_i 为文档集中出现 t_i 的文本数。

两文档之间的相似度可以用其对应的向量之间的夹角余弦来表示,则文档 d_i 与 d_j 的相似度可以表示为:

$$sim(d_i, d_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_k(d_i) \times w_k(d_j)}{\sqrt{(\sum_{k=1}^{n} w_k^2(d_i))(\sum_{k=1}^{n} w_k^2(d_j))}} \tag{2}$$

2.3 专家排序模型

对于专家排序,用户输入专长名返回的是这个专长领域内排序后的专家名列表。计算专家排序得分的基本思路是:利用专家—文档映射文件记录的数据集中每个专家对应出现的文档列表,将用户检索专家的过程转化为提交主题词返回相关文档,以及针对相关文档,结合 Expert-Docs 生成专家得分并排序显示的过程。

首先,获取查询主题后,结合建好的专长词表进行查询扩展(对于用户输入的查询主题词,使用其下位词来扩展),然后分别在索引中检索。利用 Lucene[9] 的评分机制计算返回的每篇文档的相关性得分并与每次查询该文档的权值相结合,得到加权后的文档得分。再结合专家—文档映射文件,对每个索引文档对应的专家加上刚才得到的文档得分。最后按专家得分递减的顺序排序返回专家列表。

Lucene 的文档评分公式如下:

$$w_i = \sum_{t \text{ in } q} tf(t \text{ in } d) * idf(t) * boost(t.filed \text{ in } d) * lengthNormal(t.filed \text{ in } d) \quad (3)$$

查询扩展和专家得分计算过程如图 2 所示。

该检索过程的基本伪码如下文:

```
if (queryString. equals(上位词)) //对查询词 queryString 进行判断
{
    String[ ] tempWords = topWords. getSubWords( );
    //获取此上位词下对应的下位词集(其中包含了此上位词)
    for (int w = 0; w < tempWords. length; w + + ) //对扩展后的查询词集检索
    {
      Query query = tempWords[ w ];
      //检索查询词得到返回文档 hits 结构,该结构包含文档号,得分等详细信息
      Hits hits = searcher. search(query);
      for (int i = 0; i < hits. length; i + + ) //遍历检索得到的所有返回文档
      {
        for (int j = 0; j < ExpertNames. length; j + + )
        {
          if (ExpertDocs[ j ]. contains(hits. id(i)))
            //遍历所有专家信息,若文档中含有该专家信息,则增加专家得分。如果循环到的查询词为上位词,对文档赋以较高权重;如果为下位词,赋以较低权重。
            ExpertScore[ j ] + = hits. score(i) * W;
        }
      }
    }
}
```

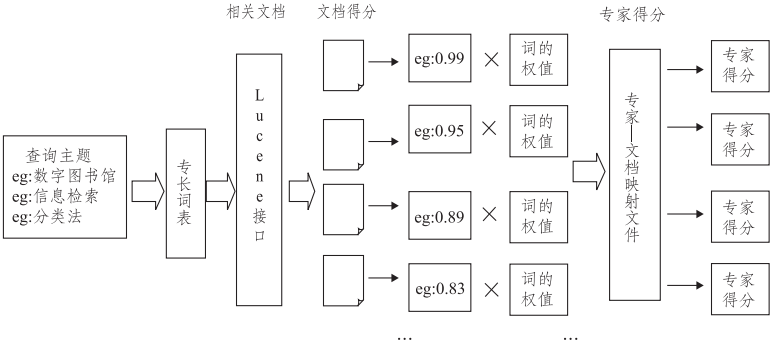

图 2 查询扩展和专家得分计算过程

在一次检索中,Lucene 所使用的向量空间检索模型会自动地对文档内查询词多次出现的情况作平滑处理。而在本文的实验中,由于查询扩展会对某个查询词的所有下位词进行多次检索,如果一篇文档中多次出现某个查询词的下位词,会在多次检索后显著地影响该文档的得分,进而影响这个文档对应的专家得分。本文采取的平滑策略是,对于获取查询主题进行

扩展的多次检索,使用哈希表来保存每次检索的每个返回文档与其已出现次数的映射。当一个文档在第一次检索结果中出现时,将其得分乘以较高的权值,之后随着该文档出现次数的增多而递减,这样就可以通过动态改变文档权值来达到检索的平滑效果。

2.4 专家专长提示模型

专家专长提示模型的构建需要有专长词表的辅助,专家相关专长得分的计算类似于专家得分的计算。不同的是,该模型要求输入的查询主题是专家名,而最终得出的是一系列该专家相关专长排序。

本实验中表示专家专长的词汇是本文所建立的专长词表中的二级词,即对“图书馆学”、“情报学”、“档案学”这几个一级词细分后的二级类目词(见表1)。专家专长检索最终返回的将是以专长词得分降序排列的专家专长词,以反映专家在各领域的贡献情况和其最具代表性的专长领域。

基本过程如下:获取用户输入的查询主题词,在索引中检索,利用 Lucene 的评分机制计算返回的每篇文档的相关性得分;然后结合专长—文档映射文件,对每个索引文档对应的专长词加上刚才得到的文档得分;最后按专长词得分递减的顺序排序返回专长词列表。输入专家名检索专长信息的过程见图3。

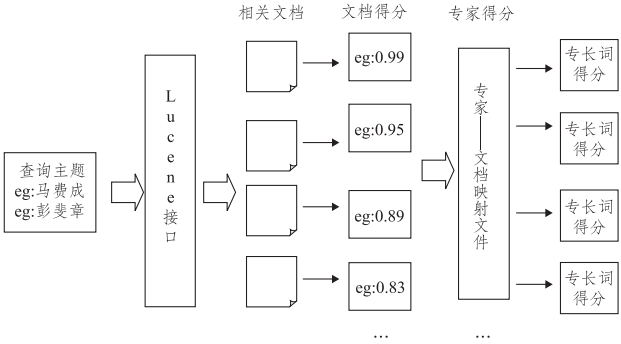

图3 专家专长提示模型基本过程

3 检索实验与评价方法

3.1 实验数据集

本实验采用手工和自动的方法构建了实验数据集,主要包括图情领域专家列表和专家数据集。对于专家数据集,由于图情领域机构大多数是高校和科研机构,因此可以表征专家的信息资源主要有:开设图情专业的院校内部的网页和文档信息、所有相关领域专家在期刊和会议上的发文情况、搜索引擎中关于该组织和相关专家检索得到的所有结果。利用实验开发的网络蜘蛛程序,对上述资源进行了抓取,在搜索引擎 Google 中抓取通过提交专家名结合专家机构的检索式(如:“马费成” and “武汉大学信息管理学院”)所得到的结果集,构建搜索引擎数据集(采集时间:2009 年 3 月,7 万余条记录);抓取万方和 CNKI 数据库中属于图书情报领域的 30 种核心或权威期刊上的所有文章(1994 – 2008 年)的题名、关键词和摘要(采集时间:2009 年 4 月,12 万余条记录)。对于所采集到的数据集的归并将在下文中详述。

3.2 实验过程

由于所采集资源的异构性,首先要对不同数据集做预处理,即对于上文中采集到的搜索引擎数据集和学术数据集,统一按照特定策略规整成编码统一的网页形式,将这两个数据集归并成同构的单一数据集,然后使用基于 Java 的开源组件 Html Parser[10] 对网页内容解析,继而使用 Lucene 对已处理的文档建立索引。

其次是要构建系统构建和运行所必需的一系列辅助文档。根据前面已经建立好了的词表

文件和专家列表,还有专家数据集索引,生成专家—文档映射文件,该文件记录了数据集中每个专家对应出现的文档列表。具体过程为:对于每一个专家,通过 Lucene 使用专家名作为检索词在所有数据集里检索,得到这个专家所对应的所有文档的索引号,去重后,写入专家—文档映射文件中。同时,生成专长—文档映射文件,记录数据集中每个专长词对应出现的文档列表。该映射文件中选取的专长词为本实验所建立的专长词表中的二级词。具体过程为:对于每一个二级专长词,通过 Lucene 对此专长词下的所有下位词(三级词)分别检索,得到这个二级专长词所对应的所有文档的索引号,去重后,写入专长—文档映射文件中。

最后利用武汉大学专家检索系统 WHU-ES 提供的接口,将上述内容整合至系统中并调用 WHU-ES 提供的 Web 检索接口,输入特定的检索主题词,根据检索词的内容,返回专家名或者专长。根据返回的排序后的专家姓名列表或专长列表,结合动态生成的专家聚类网络关系图和专家档案进行结果的判断。

3.3 评价方法

在信息检索评价中,衡量系统检索效果的两个最基本的指标是查全率和查准率,本文主要关注查准率。因为对于专家检索而言,判断返回次序靠前的结果是否准确更有意义。实验采用了 R-precision 和 P@ N 两个指标进行检索评价。P@ N 是指检出结果集前 N 个结果的查准率,可以比较有效地反映系统在真实应用环境下所表现的性能;R-precision 是指检出 R 个相关文档时的查准率,这个指标对测试用户检索的满意度有很好的效果。

4 实验结果评价分析

4.1 实验结果评价

本文共设置了三组实验项目,前两组实验是对比测试专家检索功能:第一组实验是对原始查询的专家检索结果,即不采用专长词表对查询进行扩展,仅仅使用原始查询在 Lucene 中直接进行实验;第二组实验采用了专长词表辅助进行相应查询扩展;第三组实验是对专家专长提示功能进行测试。其中为每组实验分别设计了 10 个查询主题,前两组实验所使用的主题词是专家专长词,主要是词表中的一级词和二级词,如“图书馆学”、“信息检索”、“信息计量学”、“数字图书馆”等,因为查询扩展主要是对一级词和二级词进行的。第三组实验的主题词是专家人名。

对于前两组实验,采用的评价指标是 5-prec、P@ 10 及 P@ 20,对各个查询主题的各项指标得分取平均值,得到的评测结果见表 2。

表 2 专家检索的评价结果

实验类别 \ 评价指标	5-Prec	P@10	P@20
原始查询的专家检索(扩展前)	0.562	0.433	0.327
基于专长词表的专家检索(扩展后)	0.714	0.542	0.412

由于本实验研究的是专家的专长,一般来说,某个专家的专长不可能太多,因此,在第三组实验中,笔者采用评价指标 1-prec、3-prec、P@ 5 及 P@ 10 进行了相应的评价,对各个查询主题的各项指标得分取平均值,得到的评测结果见表 3。

表 3 专家相关专长提示评价结果

实验类别 \ 评价指标	1-Prec	3-Prec	P@5	P@10
专家相关专长提示	1	0.823	0.74	0.61

4.2 实验结果分析

使用专长词表进行查询扩展后,原来单一的查询主题在检索系统中会被扩展为多个相关主题,如在本实验中,检索“信息检索”领域的专家,将会从“XML 检索”、“索引模型”、“多媒体检索”等多个方面进行查询扩展,进而返回更全面和更精确的结果。对于专家专长提示而言,使用专家名进行检索后,将会基于专长词表,返回结果得分靠前的专长词(专长词表中的二级类目词)以自动识别专家专长领域。

从前两组实验的结果看出,相对于未使用

词表进行原始词查询的检索实验,基于专长词表进行查询扩展后,各评价指标的检索效果都有所提高。从系统平均的 P@5 和 P@10 的值来看,前 5 个结果中一般可以检索到 3 – 4 个相关专家,前 10 个结果中也有 5 – 6 个相关结果。

从第三组实验看出,使用专长词表辅助进行相关专长提示的效果是不错的,特别是平均第一查准率为 100%,说明返回最靠前的专长词比较准确地反映了专家的专长。从 P@5 和 P@10 的值来看,前 5 个结果中准确反映专家专长的专长词有 3 – 4 个,前 10 个结果中也有 5 – 6 个左右的结果是准确的,说明采用专长词表辅助进行专长提示的方法是可行的。

后两组实验的 3-Prec 和 5-Prec 的平均值都比较高,即在已检出 3 或 5 个相关文档时保持了比较高的查准率,说明系统返回的前几项结果的准确率是相当高的。这样的结果对用户提高对专家检索系统的满意度是有价值的。

5 结语

本文以武汉大学专家检索系统 WHU-ES 为平台,借鉴基于相关文档集的归并排序法和基于词典进行查询扩展的方法,对图情领域专家检索进行实验与评价,包括专家排序和专长识别两个方面。从实验结果可以看出,利用专长词表可以有效地提高专家检索的查准率。同时,基于专长词表进行的专长识别也取得了较好的效果。但还有一些问题需要解决,首先是专长词表的规范性问题,主题词表的编制是图情领域研究的一个热点和难点。一方面,手工编制词表精确可靠,但是需要投入大量的人力物力,在构建效率上非常低,且更新困难。另一方面,就目前的技术水平,完全自动化地构建词表,精确地探测词汇之间的等同关系、属分关系和相关关系等是不可能的。因此,本文采用了半自动的利用向量空间模型进行文本聚类的词表构建方法。对于三级词类目,基本做到了自动化构建,且在系统的测试中取得了较好的效果。但是在前期一二级类目的划分中,受到主观因素影响比较大,因此词表的科学性和规范性、严谨性等还不理想。同时,从系统的实际运行情况来看,扩展后造成的检索速度和系统响应速度问题,也有待在以后的研究中进一步改善。

参考文献:

[1] 360doc. 显性知识和隐性知识相互转换的过程[EB/OL]. [2009-06-19]. http://www.360doc.com/showweb/0/0/555973.aspx.

[2] 陆伟, 韩曙光. 组织专家的检索系统设计与实现[J]. 情报学报,2008(5):657 – 663.

[3] Balog K, Azzopardi L, De Rijke M. Formal models for expert finding in enterprise corpora[A]. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval[C]. ACM Press, 2006:43 – 50.

[4] Westerveld T. Correlating Topic Rankings and Person Rankings to Find Experts[A]. Proceedings of the 15th Text Retrieval Conference, 2006.

[5] 许婷. 信息检索结果优化研究[D]. 武汉:华中师范大学计算机系, 2007.

[6] 魏铁进. 管理科学主题词表[M]. 北京:机械工业出版社, 1996:420 – 428.

[7] 国家图书馆. 中国图书馆分类法[EB/OL]. [2008-05-19]. http://clc.nlc.gov.cn.

[8] Salton G, Buckley C. Term-Weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513 – 523.

[9] Apache software Foundation. Apache Lucene-Overview [EB/OL]. [2008-06-19]. http://lucene.apache.org/java/docs/index.html.

[10] SourceForge. HTML Parser[EB/OL]. [2008-05-29]. http://htmlparser.sourceforge.net/.

陆 伟 武汉大学信息资源研究中心研究员,博士,教授,博士生导师。通讯地址:武汉珞珈山。邮编 430072。

刘 杰 武汉大学信息管理学院 2008 级硕士研究生。通讯地址同上。

秦喜艳 武汉大学信息管理学院 2007 级硕士研究生。通讯地址同上。

(收稿日期:2009-09-26)