广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(4):443~445
Vol. 26, No. 4 November 2010

数字档案检索的查询扩展方法*

Search Strategy of Digital Archive Based on Query Expansion

黄明初,钟 威,何拥军,蒙 斌
HUANG Ming-chu, ZHONG Wei, HE Yong-jun, MENG Bin

(1.广西壮族自治区档案局,广西南宁 530022)
(1. The Archives Administration of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530022, China)

摘要:数字档案检索的查询扩展方法以中文分词技术、查询扩展技术、信息聚类技术和数据挖掘技术等相关技术为依托,通过挖掘、整理构建相关词库,将其有机地融合到数字档案信息利用查询中实现查询扩展。查询扩展方法提供相关词检索,明确用户的查询需求,引导利用者根据自己的需求进行检索,使得利用者可以快速地获取自己需要的档案信息,提高数字档案检索的查全率和查准率。

关键词:数字档案 查询 扩展 关键技术
中图法分类号:TP302.1 **文献标识码**:A **文章编号**:1002-7378(2010)04-0443-03

Abstract: Depending on these related technologies such as Chinese Segnentation, Expanding Inquiry, Information Clustering, Data Mining, etc, the inquiry and utilization on the digital archieves information is expanded by exploiting and arranging related word library. The inquiry expansion method provides related words retrieval, defines subscribers' inquiry need, guides diferent users to search under their special needs, and finally helps them quickly obtain the archievs information they're looking for, which increase recall ratio and precision ratio of digital archives.

Key words: digital archive, query, expansion, key technologies

随着数字时代的到来,数字档案在提高工作效率与工作质量方面起着不可忽视的重要作用。然而,在当前的数字档案利用查询方面,由于利用者不熟悉档案业务,在检索策略和检索技巧上缺乏必要的知识,无法用档案关键词来表达所要检索的内容,以及由于存在同义词、近义词、相关词、缩写词等问题,导致检索结果的查全率和查准率不高,偏离利用者的信息需求[1]。如何借助信息技术手段使数字档案得以科学、高效地利用,为各部门的信息化服务,满足人民群众对档案信息的利用需求,已经成为档案信息化建设的一大难题。

针对上述类似检索需求不明确的问题,目前,国内外大多数搜索引擎如百度、谷歌以及一些搜索网站,主要是通过相关词提示帮助用户优化查询方式,明确用户的信息检索需求,但是目前该技术在档案领域还未见有相关报道。本文提出基于中文分词技术、查询扩展技术、信息聚类技术和数据挖掘技术等相关技术,在档案信息领域通过实现查询扩展,提供相关词检索,明确用户的查询需求,引导不同利用者根据自己的需求进行检索,使得利用者可以快速的获取自己需要的档案信息[2]。

1 总体技术路线

收集各种来源的数据资源,应用信息聚类技术和数据挖掘技术进行来源分析处理,经过分词、去停用词等预处理之后,应用相应的查询扩展技术和识别算法从中获取相关词候选集,整合相关词候选集形成相关词集成词库。以形成的相关词集成词库为核心,综合考虑各种利用者的不同利用需求,结合数

收稿日期:2010-08-16
作者简介:黄明初(1952-),男,工程师,主要从事数字档案管理研究。
* 2009年度国家档案局科技项目(项目编号:2009-X-07)资助。

字档案的利用查询分析研究查询扩展的应用实现。详见图1。

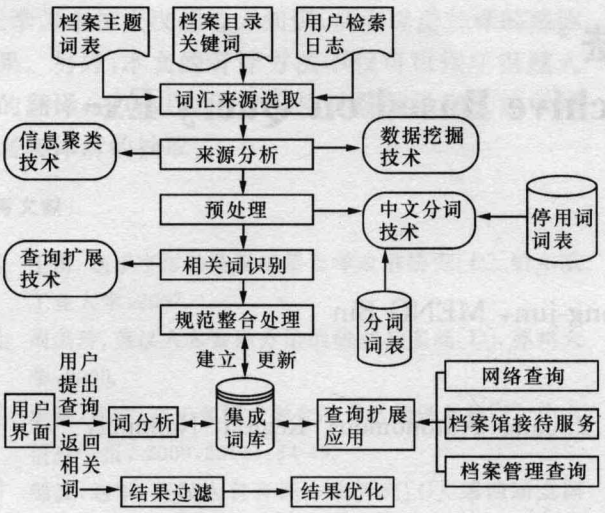

图1 数字档案检索的技术路线

2 关键技术

2.1 中文分词技术

档案目录信息中汇集众多的关键词，在自动构建相关词词库方面，需要从海量的这些目录信息中提取词汇。此外，利用者在检索档案时可能会输入较长的搜索语句，这些语句不一定和实际信息完全匹配，采用改进字符串匹配算法的中文分词技术，即结合档案行业词汇数据库和普通词汇数据库进行正反向切分可以解决这类问题[3]。

2.2 查询扩展技术

利用研究的中文分词技术对用户输入的查询关键词进行分词分析，通过整理后的查询关键词，产生一组和关键词高度相关的词。例如关键词“土地改革”，查询扩展模型将产生“土地改革运动、土地改革法、土改”等扩展相关词。此外，在用户检索式的基础上提供扩展相关词，通过检索式的重新构建可以进一步完善检索式，更好地表达用户的需求，达到扩大检索范围和缩小检索范围的效果。

2.3 数据挖掘技术

依托数据挖掘技术构建相关词库，使得语义上相关的词可以很方便地被关联在一起。这些相关词以特定的方式组织在相关词库中，使得某一个给定的词可以快速准确的找到其语义相关的词。

2.4 信息聚类技术

通过信息聚类技术获取与查询表达式相关的词。经过相关度计算，在检索结果页面的上方或下方提供相关检索词，类似谷歌的“相关搜索”。此外，在检索结果页面的左侧提供聚类浏览导航体系，类似谷歌的“百宝箱”。

3 词库构建

3.1 集成词库来源分析

档案信息领域的检索词库来源有很多方面，其中主要包括档案主题词表、档案目录中的关键词、各地档案馆依据馆藏档案自行建立的词汇、地方性词汇、档案文献集、用户检索日志等。(1)档案主题词表。目前国内含档案专有名词的分类词表和主题词表有很多，经过分析比较，选取其中常用的、有代表性的主题词表作为基础，可以选择《国务院公文主题词表》、《中国档案主题词表》等，采用人工批量录入的方式将两者结合使用。(2)档案目录关键词。经过长期的实践和积累，档案管理机构根据丰富的馆藏和多样的档案利用需求，形成了多种类型、不同结构的档案目录数据，利用中文自动分词技术对档案目录数据库中存储的大量真实的档案目录信息进行分词，将目录信息分解为基本关键词汇，再对分词结果进行了停用词处理，去掉没有实际意义的停用词。(3)各地档案馆馆藏档案。馆藏档案来源于本地区各机关及企事业单位，综合反映了这个地区政治、经济、科学和文化各方面历史活动的真实面貌，涵盖从清代、民国、革命历史、建国后各个时期，涉及文书、科技、照片、实物、音像等各种类型的档案，各综合档案馆可以依据各自的馆藏档案实际内容自行建立词库。(4)地方性词汇。各地综合档案管理机构所保管的馆藏档案均具有地方特色，如广西档案馆，常用词会建立有“广西壮族自治区”、“广西区政府”、“广西区党委”、“东盟博览会”等；广西桂林市档案馆则建立有跟桂林相关的地方词以及所属桂林的各地县名称等，针对这种情况可由档案管理员自行建立补充，从而使有地方特性、新词及生僻词可以被有效识别，使词库中包含更多的档案专业领域关键词和主题词。(5)档案文献集。档案期刊登载了对应全国各地档案工作的新情况、新进展、新动态和新成果，同时它还传播着档案学相关知识，从具有专业权威性的档案中文核心期刊中，如《中国档案》、《档案与建设》、《档案学研究》、地方档案期刊等，精心选取一定数量的文章组成档案文献集合，应用自然语言处理技术对文献集合进行自动化处理，从中挖掘出具有一定关系的词汇。(6)用户检索日志。用户检索用词是用户检索意图的体现，检索日志中包含着大量的历史访问信息，其中包含着潜在的用户反馈，反映了用户视角的词汇关系，通过对日志中的数据进

行提取、分解、合并的预处理,包括数据净化、信息识别等,最终获取相关词汇集成词库。

3.2 数据的关联整合

不同的词来源从不同侧面反映了词汇间的关系。首先使用人工批量录入的方式从选定的档案主题词表、各地档案馆馆藏档案及收集地方性词汇中提取基础词构建基础特征词库,收集档案目录关键词、档案文献集和用户检索日志几种词来源,经过分词、去停用词等预处理之后,将从中获取到的词汇经过与基础特征词库的规范整合处理,最终形成扩展相关词库。各种来源的词汇关联整合流程见图 2。

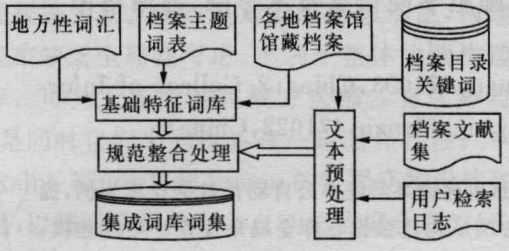

图 2 词汇关联整合流程

4 应用方式

在构建好集成词库的基础上,综合考虑各种利用者的不同需求,将查询扩展的相关词服务应用在档案信息检索上,利用检索词的相关词对检索进行修正,帮助用户重新构造更加有效的查询方式,从而减少多余检索步骤,在检索扩展中提高查全率的同时保证查准率。档案的利用查询主要包括普通用户网络查询、档案馆接待服务查询、档案管理查询。

(1)普通用户的网络查询。普通用户通过局域网或互联网查询档案,因缺乏档案专业知识和必要的检索技巧,应采用类似谷歌、百度等简洁方便的查询方式。应用查询扩展技术后可使得查询系统具有一些联想的能力,让普通用户可以通过与输入的查询关键字相关的档案信息词汇一步一步地深入查找,直到找到想要的内容(图 3)。

(2)档案馆接待服务查询。档案馆为来访的利用者提供档案查阅服务的时候,除了可采用上述所述的普通用户网络查询方式让利用者自己搜索权限范围以内的档案信息外,有需要的时候负责接待的工作人员还要为利用者代查。因工作人员熟悉档案业务,采用专业性较强的查询方式可以让他们更快更准确地查找档案,但由于档案数量非常庞大以及利用者描述不清楚等因素也会影响查询效果,查询扩展技术的应用可以减少此类情况发生的概率,例如利用者对人名的描述不清晰时,使用同音词或近音词匹配极有可能找到所需的档案,通过数据挖掘获取一些生僻的词汇作为相关词,另一些看似无关的实际是相关度很大的词汇也作为相关词显示,拓展了工作人员的视野,进一步丰富了他们的专业知识,服务质量亦不断提高。

(3)档案管理查询。基于编研、鉴定、管理、著录、统计等各种不同目的的档案管理查询,不同的目的查询需求不同,如档案编研查询的目的是把所有相关的信息都检索出来,对查全率的要求较高。而档案鉴定则对查准率要求较高,拥有分词和联想能力的查询扩展技术能弥补普通数据库查询的缺陷。

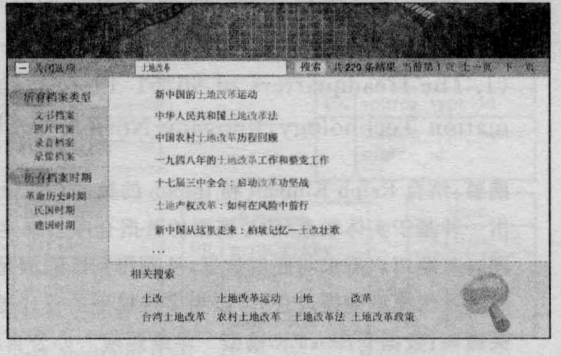

图 3 普通用户网络查询的应用效果

5 应用结果对比

从图 4 传统查询方法和查询扩展方法的对比查询结果可以看出,本文提出的查询扩展方法结合了档案领域的专有名词,并依据检索日志中的历史访问信息,提取了检索日志内的潜在的用户反馈信息,通过多来源的相关词来引导用户进行检索,所以,比传统查询方法有着更好的查准率和查全率。

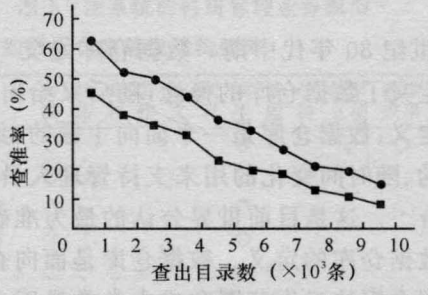

图 4 查询扩展方法与传统查询方法的查询效果对比

●:查询扩展方法,■:传统查询方法。

参考文献:

[1] 杨公之.档案信息化建设实务[M].北京:中国档案出版社,2003.

[2] 章成志,徐小琴.信息检索系统的相关词提示技术与评测[J].情报理论与实践,2007,30(1):100-104.

[3] 刘开瑛.中文文本自动分词和标准[M].北京:商务印书馆,2000.

(责任编辑:邓大玉)