情报科学
第31卷第11期 2013年11月

基于XML的中文全文检索关键技术及其发展
——简评《XML文档全文检索的理论与方法》

苏新宁

(南京大学 信息管理学院,江苏 南京 211102)

摘 要:从全文检索的理论及实践意义出发,讨论了将传统文献信息环境下的全文检索技术应用于网络信息检索的适用性与必要性,阐述了全文检索技术在网络环境下的检索对象—XML数据的结构特征、需求背景、现实意义以及发展方向,随之提出了利用全文检索技术对XML数据资源进行检索的两个难点。对此,《XML文档全文检索的理论与方法》一书从标引理论、索引理论、检索模型理论,到实现XML文档全文检索的技术路线与具体实现都做了深入细致的理论与实践研究。

关键词:全文检索;XML;Lucene;索引理论

中图分类号:G254.9 **文献标识码**:A **文章编号**:1007-7634(2013)11-155-04

Chinese Full-text Retrieval Technologies and Development Based on XML

——A Brief Review of Theory and Methods of XML-based Fulltext Retrieval

SU Xin-ning

(School of Information Management,Nanjing University,Nanjing 211102,*China)*

Abstract: Starting from the theoretical and practical meaning of full-text information retrieval, this essay discusses the applicability and necessity of applying the full-text retrieval technology in traditional information environment into network information retrieval, and illustrates the structure features, need background, practical meaning and development direction of XML data, the retrieval object of full-text retrieval technology in network environment. Furthermore, this essay proposes two difficulties of the XML data resources retrieval by full-text retrieval technology. Aiming to overcome these two difficulties, <<Theory and Methods of XML-based Full-text Retrieval>> comprehensively conducts both theoretical and practical researches on from the index theory, retrieval model theory to the realization method of XML-based full-text retrieval.

Key words: full-text information retrieval; XML, lucene; indexing theory

1 引 言

全文检索是指对文献全文或其主要部分进行索引并提供检索的一种信息技术,能迅速定位信息集合中的某一相关信息[1]。与传统的信息检索技术相比,全文检索技术功能更强大,已成为信息发现、信息分析、信息过滤信息挖掘等应用的支撑技术[2]。虽然,全文检索的理论与方法在传统的文献信息环境下已经发展得非常成熟,但在网络环境下,面对大量未经索引的网络信息,则并不能获得理想的效果。尤其是今天社会大众对网络信息的“即刻满

收稿日期:2013-04-16

作者简介:苏新宁(1955-),男,教授,博士生导师,主要从事信息智能处理与检索,信息分析与科学评价研究.

足"的消费习惯,人们更偏好使用搜索引擎,属于精确检索的全文信息检索方法已不再是主流。然而,当人们需要深入掌握和消化信息时,当信息资源服务于研究和学习时,全文检索仍然是必不可少的工具。因此,如何使全文检索技术更有效地应用于网络信息检索,则必须研究网络信息组织的特点,改善全文检索技术,使之成为适应于网络信息的全文检索技术。

由于有着可扩展性和自描述性等特点,XML得到了越来越多的应用,其结果是产生了大量以XML格式表示的数据和文档。随着XML文档的急剧增长,传统的搜索引擎已经很难满足用户需求,XML信息检索将会成为下一代搜索引擎发展的重要方向之一[3]。目前,基于HTML与文本文档的信息检索技术发展比较成熟,但没有考虑XML文档的结构信息,无法体现XML信息检索的特点。在对XML文本文档进行全文检索时,主要存在两个方面的难点。

(1)现有的全文检索技术对于XML本身的特点与优势应用不够,没有充分考虑其置标所蕴涵的语义信息,致使对大部分面向XML文档的检索仍停留在字面的机械匹配上。

(2)需要对XML查询有一个深入的理解,只有充分理解用户查询的意图与任务,才能忠实的反映出XML文档与查询之间的相关性,才能对查询进行有效的扩展。

基于这两个方面的问题,国家社会科学基金项目"基于中文XML文档的全文检索研究"(批准号:04CTQ005;结项证书号:20070672)对此展开了深入、细致的研究。作为该项目最终成果的专著《XML文档全文检索的理论与方法》(以下简称《XML》),试图深入研究元素与文档词频、元素与上下文相关元素词频的关系,重点探讨XML文档的索引方法,揭示XML文档结构及元素之间的关系,并提出了基于XML的全文检索方法,依据该方法,以XML作为通用数据接口,以Lucene为平台,实现了一个基于XML的全文检索原型系统。其研究将有助于提高XML信息检索的查全率和查准率。

2 全文检索理论体系的构造

全文检索这种情报检索技术最早出现于20世纪50年代。1959年,美国匹兹堡大学卫生法律中心建成的法律情报检索系统是世界上第一个全文检索系统。1973年,美国米德公司面向公众查询的收录有大量以法律、新闻、商业经济、政府出版物等内容为主的大型全文数据库Lexis的投入使用,标志着全文检索领域的诞生。20世纪80年代以来,英文全文检索发展得较为迅速和完善,如今已成为国外文字型信息检索的主流[4]。我国全文检索技术的研究起步于80年代末但发展速度较快。武汉大学陈光祚教授较早开展全文检索技术的研究,主持开发了"湖北省地方志全文检索系统",并倡导用后控词表来改进全文检索技术的检索效果[5]。在20世纪80年代中期,先后有经济日报全文数据库、人民日报全文数据库等几个全文数据库投入使用。从80年代末90年代初开始,我国逐渐在对国外全文数据库进行研究的基础上结合汉字处理的特点,开始了独立开发全文数据库的探索。汉字激光照排技术的发明和广泛使用,为全面开发全文数据库奠定了技术基础。

(1)标引理论。信息的标引主要是给出信息内容的概念主题和类别等。以便于用户从不同的角度用反映提问要求的词汇去检索。

计算机自动标引的理论来源于统计学方法,最初的简单词频统计仅对文献中词汇进行统计,并根据概率论思想进行高中低词汇的概率分布,去除低频词和具有高频率的虚词,余下词作为标引词;加权统计标引的思想,弥补了简单词频统计的不足,提出了根据词汇出现的位置修改词频的方法;另一种相对词频统计的方法——逆文献频率加权标引发(即要求某篇文献中的标引词在整个文献集合中尽可能少的出现)也是统计学理论在自动标引中的具体应用[6]。标引理论是厘清词汇与词汇关系的基础,是实现语义检索的重要依据。

(2)文档描述中的索引理论。文档的描述涉及到在文档集合中每个文档的描述或者描述的结构。换言之,它涉及到了文档的逻辑视图的产生。当信息检索系统在一个文档集合上执行某一查询操作之前,集合中的文档必须通过一些方式被描述。手工或自动建立文档逻辑视图的过程,称之为"索引"。现代索引以信息为基础,用在执行数据查询操作之时,加快了数据排序和搜索的速度。无论是非线性结构还是半结构文档,都是按一定结构组织的相关记录的集合。《XML》一书本着从用户信息需求的角度出发,对信息检索系统的逻辑视图等涉及文档描述的检索问题进行了分析,在索引的理论基础上,提出了文档-索引语词二维矩阵、文档-索

情报科学
第31卷第11期2013年11月

引语词加权矩阵两种文档的描述方法，为XML文档的索引与检索问题的研究提供了理论和方法的支持。

(3)检索模型理论。信息检索模型的理论基础来源于数学，数学中的集合论与布尔代数是构成布尔检索模型的基础。《XML》一书主要从布尔检索模型、概率模型、向量空模型着手，构建了面向XML信息检索的查询模型，为实现基于XML的全文检索系统提供了技术保障。

3 面向XML文档的信息检索技术与方法的发展

全面发展网络环境下的信息检索工具和索引方法，是XML全文检索的重要课题。《XML》中的研究分为两条路线：第一，对传统的信息检索工具和方法进行改造，使之适应网络环境下知识检索的需要；第二，发展新的信息检索工具与方法——面向XML文档的信息检索技术。

我们首先来看第一条路线、传统的文献信息环境的信息检索包括三个方面：①信息检索工具(即指人们用来查找信息的辅助设备，多应用于早起对印刷版文献的查找，如各类数据库等)；②信息检索方法(例如布尔检索模型、概率模型、向量空间模型)；③信息检索的产品(即对信息对象有效组织后生成的二次文献进行检索，例如对索引文档的检索、信息检索可视化等)。相应的，对传统方法进行改造可以从三个方面进行：改造传统的信息检索工具；实现自动的信息检索过程；开发并利用新的信息检索产品。《XML》一书在这三个方面都进行了全面深入的理论分析，并提出了创造性的解决方案，在真实数据集上进行了大规模的实现，开发了有效、可行、实用的原型系统。

对传统的信息检索工具改造的目标是使得改造后的网络知识检索系统具有如下三个方面的特征：①是机器可理解的、可换的；②具备自动更新、自动丰富的能力，以及跟上学科知识更新和发展的步伐；③支持对网络信息资源的自动检索，包括自动标引、索引等。《XML》一书提出的改造传统信息检索工具的方法有：①通过集成已有的信息检索方法，面向用户需求，提出了文档-索引语词二维矩阵、文档-索引语词加权矩阵两种文档的描述方法(第三章)；②构建基于XML的信息检索模型(第四、五章)；③XML索引方法的实现(第六章)。所有的这些改造方法都是建立在对已有的信息检索产品——XML文档基础之上的。如果把信息检索系统看作是抽象的概念体系，那么XML文档就是概念体系所对应的实例数据。实例数据是新词汇、新概念、新关系、新结构的来源，将这些新的知识挖掘出来，根据用户需求和实例数据间的对应关系对原有的概念体系进行丰富，并根据实例数据的统计分布和改造原有的概念体系，这就是《XML》一书各项研究的基本思路。

第二条路线，就是发展新的信息检索工具与方法，《XML》一书做出了两方面的尝试：①基于传统知识组织资源构建基于XML的信息检索查询模型，实现语义检索，并增强搜索引擎；②构建基于XML的全文检索原型系统。基于XML的全文检索原型系统实现语义查询和知识浏览，并向搜索引擎提供词汇辅助，实现检索结果的聚类。基于XML的全文检索原型系统的具体功能模块和实现过程在书中的第七章有详细的介绍。

这些研究工作是一个长远研究图景中的第一步，其最终的目标实现网络信息资源与用户需求的自动匹配，这要求网络知识检索系统具备从网络资源(如网页)中直接学习新知识的能力。网络信息资源相当自由和松散，缺乏一致性的结构，这就要求传统信息检索工具的进一步改进，不仅承担检索工具本身具备的报道功能、存储功能和检索功能，还需要拥有丰富的词汇和词汇间的关系，即语义功能。传统检索工具中的词汇都是规范的受控词汇，所覆盖的领域知识在广度与深度上都很不够，《XML》一书中试图从事的一个重要研究，就是借助于XML所具备的优势来改造和丰富已有的检索工具。在《XML》一书完成之际，夏立新教授所主持的国家社会科学基金项目“基于中文XML文档的全文检索研究”已取得了突破性的研究进展，以Lucene为基本实现平台，建立索引和查询索引等，在实现全文检索基本功能的基础之上，再针对于中文XML一些二次开发，从而实现基于XML的全文检索系统。

4 基于XML的中文全文检索系统的具体实现

《XML》一书在第七章详细的介绍了基于XML的中文全文检索系统的具体实现过程：首先论述了全文检索原形的实现平台和关键技术；然后，论述

了全文检索原形系统的实现过程，主要包括：XML解析模块，中文自动分词模块，全文索引模块，全文检索模块和检索界面等的实现过程；最后通过一个实验平台来验证模型的优越性和可行性。

开发的基于XML的全文检索原形系统的设计目标是：充分考虑了XML置标所蕴涵的语义信息。在建索引时，按照不同的置标，将不同置标的内容分开建立索引，提高标引的深度[7]。同时对多种语种信息内容（比如同时包含中文和英文的网页内容）进行索引。并提供多个字段的组配检索，布尔逻辑检索等，以方便用户更有效地查询信息。

和其他的全文检索系统相比较，《XML》一书开发的基于XML的全文检索模型有两大特色。

（1）索引结构的特点。从索引结构的优化入手，在索引结构的设计及组织方式上充分考虑了XML置标所蕴涵的语义信息。如，根据XML文档的结构，将有关作者的信息和标题信息分开建立索引，查询时，可以根据用户的要求分开查询，也可以灵活的组合查询。

（2）集成了一个性能优良的中文自动分词模块。系统在建索引和查询时，通过调用自动分词模块，将XML文档的内容处理前进行一次自动分词的预处理。中文自动分词模块的集成，使索引文件压缩的更小，同时使系统的查准率得到了很大的提高，从总体上提高了本模型的系统效率。

这些实践说明，实用的全文检索系统将是理论与技术的结合，建立在传统信息检索理论的技术之上，结合XML研究的最新研究进展，利用自动分词、自动标引、搜索、机器学习、数据挖掘等技术，实现网络环境下的知识检索。

5 结 语

从检索技术本身来看，如何更好地解决语词的切分问题以及语义理解、句法理解问题，提高全文检索系统的检索性能，乃至实现具有学习、分析、理解、推理机制的智能化和基于知识库推理机制的信息检索系统，才是未来全文信息检索系统发展的主要趋势。

Web搜索引擎的全文检索技术发展至今，目前多是基于词语匹配层次上，对于用户的检索要求不甚清楚，或是无法发掘原文的潜在的信息，网络信息的检索效果一直没有显著提高，因此要从根本上改善网络信息检索工具的检索效果，可从网页项目和结构的标准化，以及从索引机制的完善做起。XML作为一种标准在出版和数据交换中的应用，为更好的信息检索创造了很好的机会。XML能够用结构化、机器可读的格式来显示数据的语义[8]。一些机构已经开始定义标准模式来获得许多域的语义，而一些内容提供商开始用XML和标准模式来发布信息。于是，有关专业人士指出“，XML将引发对Web查询技术、Web数据库技术及Web数据交换技术的全面革新”[9]。如何将关键词检索发展为知识检索，如何利用XML语言的半结构化特点发展出更加智能，面向语义理解的检索系统，是信息检索研究主要方向。《XML》一书所取得的学术成果，将大大推进中文全文检索技术理论体系的构建及其技术的创新，为实现基于语义的知识检索及知识检索的可视化打下坚实的基础。

参考文献

1 夏立新，王忠义.基于XML的全文检索原型系统的设计与实现[J].现代图书情报技术,2007,(8):67-70.
2 张云秋,吴正荆. 网络全文检索系统的实现技术及其未来发展[J].情报科学,2003,(10):1080-1083.
3 贾素玲,王　强.XML技术应用[M].北京：清华大学出版社,2007:3.
4 YueShan Chang , MinHuang Ho , ShyanMing Yuan. A unified interface for integrating information retrieval [J]. Computer Standards & Interfaces ,2001,(23) : 325-340
5 陈　睿,陈光祚,谢新洲. 湖北省地方志全文检索系统(上)[J]. 情报理论与实践,1991, (2):29-30.
6 夏立新,金　燕,方　志.信息检索原理与技术[M].北京：科学出版社,2009:13-16.
7 夏立新, 庄青青, 陈卓群. 论XML文档的索引结构设计[J]. 情报科学,2007,(9):1378-1383.
8 方　志,夏立新,刘启强.中外全文检索研究的现状及趋势[J].图书情报知识,2006,(9):71-73.
9 王弘蔚,肖诗斌. 一种基于NativeXML的全文检索引擎[J].情报学报,2003, (5):550-556.

（实习编辑：赵红颖）