基于引文的信息检索可视化系统设计与实现

Research and Implementation on the Citation - based Information Retrieval Visualization System

孙 巍 张学福

（中国科学院国家科学图书馆 北京 100080）（黑龙江大学信息管理学院 哈尔滨 150080）

摘 要 针对新的环境下用户信息需求的特点，设计并实现了具有个性化服务特色的基于引文的信息检索可视化系统，包括系统目标、体系结构、工作流程、核心功能模块及关键技术等，为相关检索可视化系统的进一步研究奠定基础。

关键词 引文网络 引文分析 检索可视化系统

1 基于引文的信息检索可视化系统的开发背景

随着科技的发展，社会信息化、数字化的推进，信息量级数倍增，为了满足信息检索用户日益增长的多样化、个性化检索需求，帮助用户更深刻地揭示信息背后隐藏的信息关联及规律，提高检索效率，可视化技术作为一种新生技术引入到信息检索领域。

信息检索可视化[1]是指把文献信息、用户提问、各类情报检索模型以及利用检索模型进行信息检索的过程中不可见的内部语义关系转换成图形，在一个二维或三维的可视化空间中显示出来，并向用户提供信息检索的技术。引文分析法在揭示信息关联及规律方面具备其他许多方法不可比拟的优越性和独到之处，主要用于研究科学文献结构和科学结构，揭示科学发展史及其规律，评价科研成果以及研究情报用户的构成及行为等[2]。

用户对信息需求的变化以及引文理论、信息检索可视化技术的特点使得研究和构建基于引文的信息检索可视化系统具有重要意义。本文正是在这一新的背景环境下，运用文献间的相互引证关系及现有相关检索可视化技术，研究并实现了具有个性化服务特色的基于引文的信息检索可视化系统（My Information Retrieval Visualization System Based on Citation，简称 My_IRVSC）。

2 My_IRVSC 的总体设计

2.1 My_IRVSC 的系统目标 系统主要实现用户需求选择、引文数据检索、检索结果可视化及检索结果统计分析功能，依次运用需求选择、文献预处理、被引频数统计、相似度计算、建立共引矩阵以及 Radial layout、GraphView 及 DataMountain 的聚类算法等技术。

2.2 My_IRVSC 的体系结构 本系统采用三级 B/S 体系结构（如图 1），浏览器端负责用户与系统间的交互；服务器端包括功能服务器和数据服务器两个层次，功能服务器主要负责与数据服务器、用户界面的交互，对用户提问处理及检索匹配，对信息检索可视化所涉及的中间数据进行处理，并根据处理数据进行共引网络可视化，对检索结果统计分析等；数据服务器负责根据与功能服务器的交互，对数据库数据进行处理。这样既减轻了对浏览器端的性能要求，也减轻了数据服务器的负担，优化了分布数据处理性能，充分发挥了三级 B/S 模型的数据处理能力和安全控制机制。

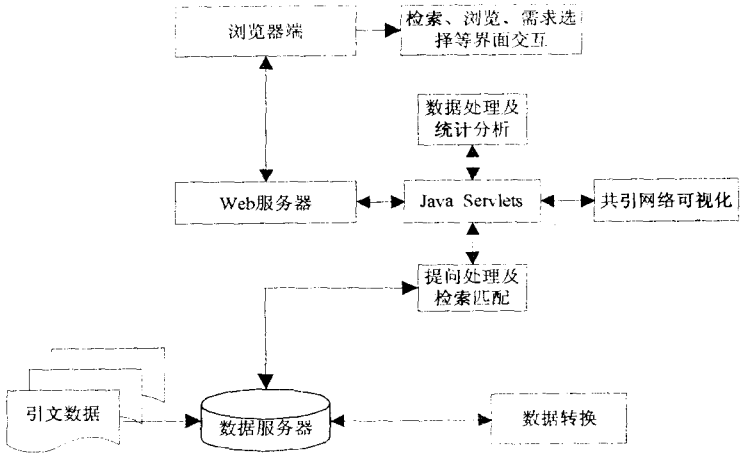

图 1 My_IRVSC 原型系统体系结构图

2.3 My_IRVSC 的工作流程 a. 数据准备。数据服务器中的数据是基于 HTML 格式的文献信息，每一个 HTML 文档记录了一篇文献（包括其引文信息）。所有的 HTML 文档经过引文提取转换程序被转换成 XML 标准格式，为检索及数据处理模块提供数据基础。

b. 需求选择。用户登录系统后，通过用户检索界面进入需求选择页面，在此进行需求服务选择、组合，系统将选定的组合中的各参数变量传递给检索页面。

作者简介：孙 巍，女，1978 年生，博士生；张学福，男，1966 年出生，博士，教授。

c.初次检索。用户在检索页面输入检索词，根据检索需求提交初次检索请求；Web服务器根据检索需求参数及检索提问，经过java servlets与数据服务器的交互以及提问处理及检索匹配模块，即对用户输入的检索词进行切分、去除停用词、与数据准备阶段处理的引文数据集进行字符串匹配等一系列过程完成引文检索，生成检索结果引文特征词文档集，并将其传给数据处理及统计分析模块，统计分析模块根据用户的统计分析及矩阵中特征词粒度控制服务需求参数，对相应的引文特征词文档集合中各个文档进行预处理，统计被引频数，降序排列，计算其中前 n（矩阵中特征词的个数）个高被引特征词在所有文档集合中的被引频数，生成相应关联引文特征词初始矩阵及共引矩阵，将共引矩阵数据传递给共引网络可视化模块，同时依据引文分析及引文网络理论为用户提供统计分析结果，并通过Web服务器将其传递给客户端浏览器，显示分析结果；共引网络可视化模块根据用户可视化显示需求参数生成用户需要的共引网络图形数据，将其传递给客户端浏览器，显示引文网络可视化图形；另外，系统还提供了用户操作过程信息存储需求服务，将用户每一次检索操作过程的所有参数传递给“我的文档”页面，通过该页面显示用户的操作过程信息及相应的操作错误信息提示。

d. 对检索结果的检索。用户通过分析引文网络中各引文特征词之间的关系以及相应的统计分析结果，从c中得出的引文网络中重新选择特征词，也可以此词为基础重新进行需求选择组合，重复c步骤重新生成新的引文网络及引文统计分析结果，直到用户认为基本满足其检索需求为止。

e.可视化信息检索。用户经过b、c、d步骤后，检索结果被显示在另一页面上。用户通过对检索结果的判断，可以多次与系统交互，通过不同的引文特征词把相关文献检索出来，同时用户可以点击检索结果源文献超链接，获取源文献。

整个工作是一个循环往复的过程，如果用户认为通过可视化检索，已经满足其检索需求，检索过程结束；如果在浏览、理解检索结果的过程中，用户又改变了信息需求，重新提交检索请求，将又一次重复b、c、d步骤的操作。

3 My_IRVSC核心功能模块

My_IRVSC系统主要功能可分解为数据管理模块、系统管理模块、数据检索模块、数据处理及分析模块和共引网络可视化模块（详见图2）。

3.1 **数据管理模块** 该模块包括数据导入及转换子模块，本文研究的原型系统以HTML格式的IEEE中的部分数据为数据源，手动下载至本地实现数据导入。数据转换主要是把每篇源文献的题名及其所有引文的题名、作者、期刊信息分别从HTML格式引文数据中提取出来，自动生成xml文档元素，实现格式的自动转换。

3.2 **系统管理模块** a.权限管理子模块将用户注册信息存储到用户信息数据库中，系统以此来验证用户名及密码，确定合法用户；b. 需求选择子模块主要功能是用户通过注册从系统主页登录到个性化特色的检索页面，通过该页面进入需求选择页面，进行统计分析需求、可视化显示需求、特征词粒度控制需求及操作过程存储需求选择组合，需求选择结果作为隐形表单变量向检索页面传递参数，检索页面接收所有的需求选择组合参数实现需求选择服务；c. 系统的帮助页面与系统主页相同，用户不需注册即可进入该页面。Google外部链接检索为用户提供辅助检索信息；系统的基本功能介绍、操作手册及图表指导用户了解全部的系统功能，为系统使用中经常遇到的问题提供解决方案。

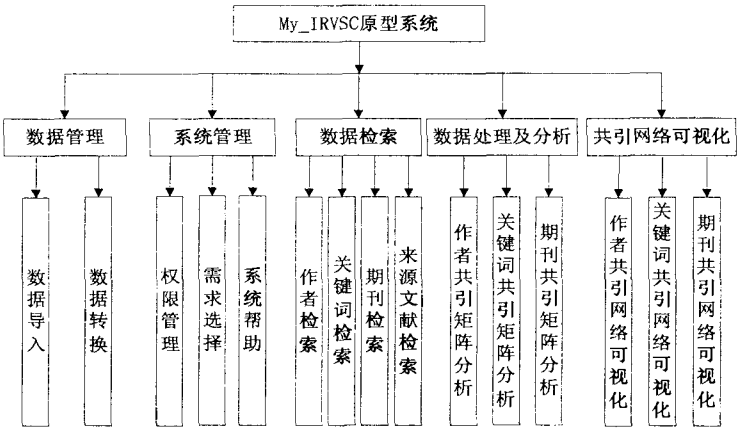

图2 My_IRVSC原型系统主要功能模块

3.3 **数据检索模块** 该模块主要是根据接收用户检索需求参数，通过检索页面实现对源文献、引文中的题名关键词、作者及期刊的检索。用户通过多次重新进行需求选择以及与检索页面的交互完成检索的全过程。此过程中，既可以进行初次检索，还可以对检索结果实现再次检索。该模块还实现了在另一页面显示源文献的原文、引文关键词、作者及期刊等检索结果。

3.4 **数据处理及分析模块** 数据处理模块主要是根据接收的用户统计分析需求参数及矩阵中特征词粒度控制需求参数对引文中的题名关键词、作者及期刊进行数据处理，完成文本预处理、被引频数统计、相似度计算、生成共引矩阵以及相关被引频数统计图表四个功能。

由于本文研究的系统在作者及期刊检索上实现的是全词检索，通过数据转换模块已初步实现了引文中的作者及期刊的文本预处理，该模块在对这两项进行数据处理过程中没有对其进行文本预处理。

3.5 **共引网络可视化模块** 该模块包括特征词的聚类和可视化呈现两个部分。在特征词聚类方面，根据数据处理及分析模块生成的特征词节点数据实现特征词聚类；在可视化呈现方面引入了Prefuse工具提供的事例程序，系统根据所接收的用户可视化需求参数，实现共引关键词及共引作者、共引期刊可视化显示。

4 My_IRVSC系统开发的关键技术

4.1 文档预处理 My-IRVSC系统开发过程中的文档预处理包括对HTML源文档以及对题名关键词的文本文档的预处理两部分。对HTML文档的处理需要对HTML文档进行Tag查找、字符串匹配、不规则符号替换等;对引文题名中的关键词文档的预处理包括去除引文题名中的标点符号,进行英文词切分,提取检索词词干,排除停用词,对用户输入检索词进行同义词替换等。

4.2 文档格式转换 My-IRVSC开发过程包括HTML向xml以及xml向文本文档两种格式的转换工作。

在对HTML文档的预处理的基础上,最终实现以"#"分隔xml文档中各元素的题名、作者、期刊。该系统选用将HTML源文档转换成xml文档格式的数据是基于下面两点考虑:第一,XML文档格式是比较通用和获得国际认可的标准格式,可以方便地与数据库(如SQL Server)之间进行数据转换;第二,不用考虑数据库操作以及数据库与java程序的数据传递操作。

为了给共引网络可视化模块及数据处理及分析模块提供数据基础,数据检索模块的最终数据是以文本形式存储的。系统运用了Java数据绑定的开放源代码Castor框架,该框架是一种代替xml文档模型的强大机制,能用来处理日益复杂的文档,它是主要关心文档数据内容的应用程序。通过Java castor[3]数据绑定技术的编组和解组功能来实现语料库XML文档的重新映射,实现了xml格式向txt格式的转换。

4.3 数据检索 My-IRVSC系统主要实现了题名关键词、作者、期刊的检索,对作者、期刊的检索是通过用户提问与xml数据完全匹配的方法实现的,因此主要的检索技术是对引文题名关键词的检索,具体的实现步骤如下:

步骤1,对引文关键词的扩展。通过查找同义词词典,将用户输入的i个检索词替换成相应的i组同义词环。

步骤2,基于词干查找。利用波特算法,将引文题名中的关键词与上一步替换后的检索词进行基于词干的查找。

4.4 相似度计算 数据处理及分析模块需要对每两个文档的相似度进行计算,基本算法如下:

步骤1,读取被引频数统计过的文档,对任意两个文档进行比较,抽取共有关键词。

步骤2,依据明氏(Minkowski)距离公式[4],计算任意两个文档的关联性:

$$D(X,Y)=\left(\sum_i |X_i-Y_i|^r\right)^{1/r}$$

步骤3,将计算结果以矩阵形式写入文件。

4.5 生成矩阵 数据处理及分析模块需要生成相应的特征词共引矩阵及各种被引频数统计表。生成共引矩阵及其变形矩阵的算法为:

步骤1,分别统计题名关键词、作者、期刊被引的总频数,降序排列,存入高被引降序数组,并生成相应的图表;

步骤2,统计高被引降序数组中任意两个特征词在一个源文档中的共引频数。生成特征词*源文档矩阵,特征词*特征词矩阵;

步骤3,将权值大于用户指定值的特征词*特征词矩阵存入新矩阵;

步骤4,将显示权值大于用户指定值的共引特征词及其共引频数值写入可视化数据文档。

4.6 嵌入可视化工具 系统在共引网络可视化模块中引入Prefuse工具提供的事例程序RadialGraphDemo,RadialGraphView及DataMountain类,系统可以根据所接收的用户可视化需求参数,以RadialGraphDemo(图3,左下),RadialGraphView(图3,右上)类实现共引关键词及共引作者、共引期刊可视化显示,另外,Datamountain(图3,左上)类可实现对共引期刊的可视化显示。

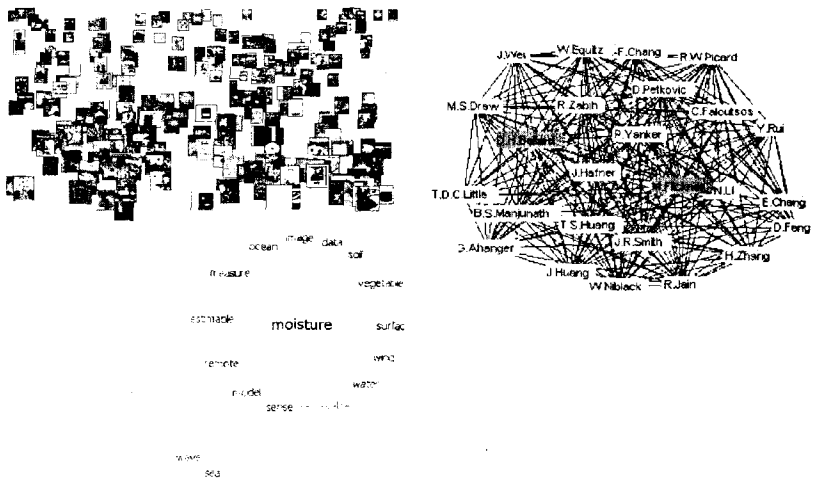

图3 三种共引网络可视化图

5 结论与展望

经粗略的系统功能测试,得出My-IRVSC系统是一个集检索、可视化及统计分析功能于一体的具有个性化特色的信息检索可视化系统。

检索功能方面,My-IRVSC提供了以作者、期刊、关键词为检索入口的多种检索途径,能检索出引用过该检索词的文章的作者、期刊、关键词及该文章的源文献,并可通过链接获取源文献,同时通过多种可视化图形与用户进行交互,为用户提供实时可视化检索服务,检索结果显示形式清晰,信息组合的逻辑性强;可视化方面,My-IRVSC运用了PFNET映射算法,及多种显示技术对检索结果进行可视化,并 (下转第76页)

态分布)和全局分布信息(文献集合,动态分布),也就是说,当我们设计加权方案时,既要考虑局部权重,又要考虑全局权重。同时,还需要消除文献长度和语义信息缺乏等不对称分布问题的影响。本文在向量空间模型基础上对这些问题进行了探讨,在后续研究中,我们主要探讨三个方面的问题:基于概率检索模型的词加权技术及其优化策略,分布式环境下的加权技术及其优化策略,以及相关反馈技术在查询加权中的应用。

参 考 文 献

1 Robertson, S. Understanding Inverse Document Frequency: On Theoretical Arguments[J]. Journal of Documentation, 2004, 60(5): 503 - 520

2 Salton G, Buckley C. Term - Weighting Approaches in Automatic Text Retrieval[J]. Information Processing & Management, 1988, 24(5): 513 - 523

3 Lan M, Sung SY, Low HB et al. A Comparative Study on Term Weighting Schemes for Text Categorization[J]. Proceedings of 2005 IEEE International Joint Conference on Neural Neural Networks, 2005(1): 546 - 551

4 Cummins R, O' Riordan. Evolving General Term - Weighting Schemes for Information Retrieval: Tests on Larger Collections[J]. Artificial Intelligence Review, 2005(24): 277 - 299

5 Papineni K. Why Inverse Document Frequency[C]. NAACL. Proceedings of the North American Association for Computational Linguistics, New York, NY: Association for Computational Linguistics, 2001: 25 - 32

6 Singhal A. Term Weighing Revised[D]. Ithaca, NY, USA: Cornell University, 1997

7 Greiff WA Theory of Term Weighting Based on Exploratory Data Analysis[C]. Croft WB, Moffat A, van Rijsbergen CJ et al. Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98). New York, NY, USA: ACM, 1998: 11 - 19

8 Wistsch HF. Global Term Weighting in Distributed Environments [J]. Information Processing and Management, 2007, 43(2): 1 - 13

9 Sparck Jones K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval[J]. Journal of Documentation, 1972, 28(1): 11 - 21

10 Manning CD, Raghavan P, Schütze H. An Introduction to Information Retrieval [M]. Cambridge, England: Cambridge University Press, 2007: 1 - 461

11 Aizawa A. An Information - theoretic Perspective of tf - idf Measures[J]. Information Processing and Management, 2003, 39(1): 45 - 65

12 Kang BY, Lee SJ. Document Indexing: a Concept - based Approach to Term Weight Estimation[J]. Information Processing and Management, 2005, 41(5): 1065 - 1080

13 Shehata S, Karray F, Kamel M. Enhancing Text Retrieval Performance Using Conceptual Ontological Graph[C]. In ICDM Workshops, 2006: 39 - 44

14 Singhal A, Buckley C, Mitra M. Pivoted Document Length Normalization[R]. Pivoted Document Length Normalization, TR95 - 1560. Ithaca, NY, USA: ACM SIGIR, 1996: 21 - 29

15 Singhal A, Salton G, Buckley C. Length Normalization in Degraded Text Collections[C]. Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, Ithaca, NY, USA: Cornell University, 1996: 149 - 162

16 Zakos JA. Novel Concept And Context - based Approach for Web Information Retrieval[D]. Indiana, USA: Griffith University, 2005

17 Zhai C, Lafferty J. A Study of Smoothing Methods for Language Models Applied to Information Retrieval[J]. ACM Transactions on Information Systems, 2004, 22(2): 179 - 214

(责编:梅王京)

(上接第 72 页)且可以通过实时交互的方式对动态图形的参数加以控制,多种共引网络以及揭示引文之间引证关系的可视化形式,生动直观;系统能够提供完全实时的检索可视化功能,通过多种可视化及分析功能,辅助用户检索以及对检索结果进行统计分析,进一步降低了用户的认知负担。此外,My-IRVSC 系统具有个性化特色,为用户提供了需求选择服务。系统通过用户对一系列系统的检索、可视化、统计分析等功能的选择组合,在系统所提供的多样化服务的范围内,最大程度地满足了用户的个性化信息需求,在一定程度上加强了与其他方式服务的集成。

由于时间、条件等方面的限制,本文研究还存在如下不足:

a. 本文采用了实际数据库中的部分资源对原型系统进行了验证,然而缺乏在真正的实际数据库资源条件下的系统、全面的验证,这是下一步工作中研究的重要内容;b. 系统中采用的各种可视化显示技术还有许多不便和问题,需要进一步考察相关算法和技术,并对其进行改进,以更好地适应用户认知环境下的信息检索可视化。

参 考 文 献

1 张 进. 论情报检索可视化过程中信息节点的歧义性问题[J]. 情报学报, 1998, 17(3): 175 - 179

2 胡利勇, 陈定权. 同引分析与可视化技术[J]. 情报科学, 2005(4): 532 - 537

3 Java 中的 XML: 使用 Castor 进行数据绑定[EB]. http://www.jspcn.net/htmlnews/1104931873609l590.html[2006 - 4 - 10]

4 孙建军, 成 颖. 信息检索技术[M]. 北京: 科学出版社, 2004: 204

(责编:阳)