基于共词分析的中文信息检索可视化研究

陈 颖[1,2],白淑琴[3],张学福[4]

(1. 中国科学院 国家科学图书馆,北京 100190; 2. 中国科学院 研究生院,北京 100190;
3. 中国石化集团 河南石油勘探局北京办事处,北京 100073;
4. 黑龙江大学 信息管理学院,黑龙江 哈尔滨 150080)

摘 要:说明了新的信息环境下将共词分析方法与信息检索可视化技术相结合的必要性,设计并实现了基于共词分析的中文信息检索可视化系统,对系统主要功能模块进行介绍,并对未来工作进行展望。

关键词:共词分析;信息检索可视化;中文;信息检索系统
中图分类号:G354 **文献标识码**:A **文章编号**:1007-7634(2009)02-0227-04

Research on the Co-word Based Information Retrieval Visualization System

CHEN Ying[1,2], *BAI Shu-qin*[3], *ZHANG Xue-fu*[4]

(1. *National Science Library, Chinese Academy of Sciences, Beijing* 100190, *China*; 2. *Graduate University of Chinese Academy of Sciences, Beijing* 100190, *China*; 3. *Beijing Office of Henan Petroleum Exploration Bureau of China Petrochemical Corporation, Beijing* 100073, *China*;
4. *Information Management Institute of Heilongjiang University, Harbin* 150080, *China*)

Abstract: This paper explains the need of combining co-word analysis with information retrieval visualization, designs and implements Co-word based Information Retrieval Visualization System, introduces the main modules, finally previews the future.

Keywords: co-word analysis; information retrieval visualization; Chinese; information retrieval system

1 引 言

信息检索可视化是把文献信息、用户提问、各类信息检索模型以及信息检索过程中不可见的内部语义关系,展示在一个低维的可视化空间中,并向用户提供信息检索服务。20 世纪 90 年代后期的信息检索可视化研究日益实用化和智能化,涌现出一批原型系统,如 VIBE, TileBars, LyberWorld 等。此外,国外许多大型搜索引擎和电子商务网站中也出现了功能日趋强大的可视化浏览和检索界面。如可视化元搜索引擎 kartoo[1]实现了对网络站点和页面关系的可视化,能将搜索结果以一个可视、交互的地图表示。

共词分析方法是内容分析法的一种,通过共词方法可发现概念之间的关联,其原理是对一组词两两统计它们在同一篇文献中所出现的次数以此为基础对这些词进行聚类分析,反映出这些词之间的亲

收稿日期:2008-08-27

作者简介:陈 颖(1977-),女,吉林永吉人,讲师,在读博士生,从事数字图书馆,情报分析,信息检索研究; 白淑琴(1968-),女,河南南阳人,馆员,本科;张学福(1966-),男,山东阳谷县人,教授,博士,从事信息检索,信息系统,数字图书馆研究.

疏关系及文章中的概念结构，进而分析这些词所代表的学科和主题的结构变化。较之同引分析，共词分析是对当前发表文献的直接统计，反映的是目前已有论文所集中关心的主题，表现手段上更简单，能更直接地勾勒出当前的学科结构。

网络环境下，海量信息涌现，用户需求多样。建构能更好地揭示信息内容及内容间关联的、便于用户操作和理解的可视化检索系统有着重要意义。因此，本文将共词分析方法与信息检索可视化技术相结合，研究并实现了基于共词分析的中文信息检索可视化系统(The Visualization System of Chinese Information Retrieval Based on Abstract Information，VSCIRAI)。

2 基于共词分析的中文信息检索可视化系统设计

信息检索可视化系统设计时要先明确两个问题：可视化对象的确定和可视化表示形式的选择。可视化对象是指将要在可视化空间显示的内容，主要有文档、数字图书馆、网站和超链接结构等；可视化表示形式是指在可视化空间中以何种形式来代表可视化对象，主要有几何图形，自然实物形式和图标形式等。本文在考察和分析了若干国内外信息检索可视化系统后，将文档摘要作为可视化对象，可视化表示形式采用易于理解的 Radialtree 图形表示。

2.1 系统设计目标

VSCIRAI 将可视化技术应用于信息检索系统中，对文档摘要信息进行检索，运用共词分析方法对检索文档集中概念词进行处理，实时生成概念空间图，从而实现检索过程、检索结果可视化。系统目标主要有以下几个方面。

(1)以用户为中心。由用户主导检索过程，用户可与系统实时交互，如定制检索界面，选择检索方式等，从而实现个性化服务[2]。

(2)检索扩展与精化。利用共词分析方法扩展、精化检索词，实现相关反馈，使用户逐渐明确并发现新的信息需求。

(3)实时信息检索可视化。系统能根据用户的需求，检索出相关文档，并能通过可视化显示技术将检出文档集中概念词共现信息及文档间关系展示给用户。用户可改变浏览和检索信息的视角，修正或再次检索，操纵整个检索过程。

(4)具备标准化、可扩展接口。系统采用模块化设计，利用跨平台的开发工具实现，提供标准化接口，可处理大容量大范围数据集，并提供功能扩展和系统升级。

2.2 系统模型设计

将传统的信息检索系统涉及的信息进行视图映射和转换，使可视化处理能服务和受控于信息检索任务，从而实现信息检索的可视化。模型参见图 1。

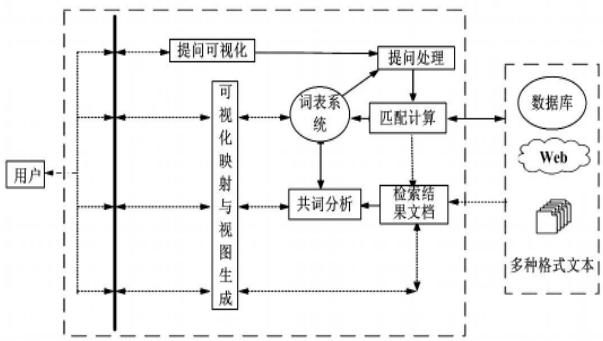

图 1 基于中文的可视化信息检索模型

图 1 中实线箭头表示传统信息检索中的信息传递与处理以及人机交互过程，虚线箭头代表经可视化处理的检索过程。根据用户的需求，信息检索中涉及的检索提问、词表处理(如未登录词识别，同义词扩展，排除停用词)、匹配计算过程和对检索结果的处理操作等，都能够进行可视化映射、生成相应视图，并通过图形用户界面反馈给用户。在此过程中，用户通过图形界面上的实时交互，对可视化处理中的各种参数进行控制，向系统传达新的检索指令，以便于查看各种视图效果及调整检索参数和策略。实现了检索过程和检索结果的可视化。其中共词分析部分又可细分为数据抽取，构造共词矩阵或共词向量和数据分析三个步骤。

2.3 系统主要功能模块设计

VSCIRAI 主要功能模块如图 2 所示。纵向可看作系统主要流程，模块间存在一定的协作关系，较粗的双向箭头表示模块间协作较为紧密。

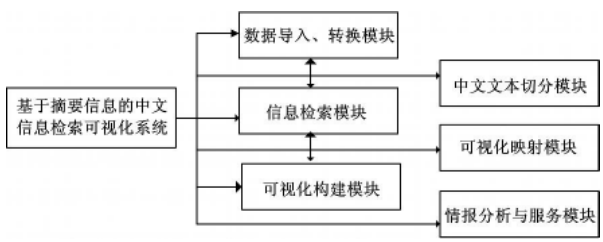

图 2 VSCIRAI 主要功能模块

各模块功能描述如下。

(1)数据导入、转换模块。

本模块与信息检索模块在功能上密切配合。数据导入部分负责经由检索将符合用户需求的文档提取出来,并抽取出摘要信息,作为系统基础数据。数据转换部分负责将不同数据库中不同格式、不同字段的数据统一转换成本系统所处理的 XML 格式,使系统独立于数据库,对不同数据格式提供统一接口。

(2)信息检索模块。

是系统核心模块。主要提供文档检索和基于概念空间的概念检索,并利用可视化技术,将检索过程及结果实时呈现给用户,支持用户与系统的交互反馈。文档检索主要包括初级检索和高级检索,支持同义词扩展检索,支持检索词初步处理功能,支持检索结果排序功能。概念检索负责与映射处理模块、可视化构建模块协作,对检索结果文档集中文档的摘要信息进行共词分析,形成概念空间图,展现给用户。进而,用户可基于概念空间图来明确、调整信息需求,发现新的检索词,重构和优化查询式;可通过词共现关系并结合用户背景知识来推理词的语义关系,在一定程度上实现了概念检索、语义检索。

(3)中文文本切分模块。

本系统涉及词频统计分析并利用切分出的词进行匹配检索,切分效率及精度对系统有重要的影响。因此,本模块是系统的核心及难点。

中文文本词切分涉及许多问题:如分词规范,分词算法的选择与优化,词典编制与使用和运行效率等。本模块设计时,结合系统具体要求和现有技术及资料条件综合考虑了上述问题。采用中国科学院计算技术研究所的汉语词法分析系统 ICTCLAS,并在其基础之上再次开发,以适应本系统的需求来完成切分词,去停用词,切分歧义,未登录词识别等功能。

(4)映射处理模块。

主要功能是对检索结果文档集中文档的摘要信息进行共词分析,利用可视化映射算法,对相关数据进行处理,为实现可视化图形准备数据。主要包括:文本同义词替换、词频统计、相似度计算、形成概念共现矩阵及其变换矩阵等。

(5)可视化构建模块。

是系统的核心模块,功能主要是实现概念空间图的实时生成,并在一定程度上实现了概念聚类。可视化构建过程是实时的,并可与用户动态交互。

(6)情报分析与服务模块。

本模块体现了本系统的另一特色。即利用共词分析及相关情报分析方法,基于检索结果及可视化概念空间图分析主题、领域、概念间语义关联等,以期发现概念间潜在关联,学科热点、学科结构等,为用户的更一步的情报需求提供帮助[3]。

3　基于共词分析的中文信息检索可视化系统实现与评价

3.1　系统实现

基于扩展性、共享性和安全性的考虑,本系统采用三层 B/S 架构,由数据服务器、WEB 服务器和 WEB 终端组成。系统体系结构如图 3 所示。开发环境为:CPU:P2.40G,内存:1GB,操作系统:Windows XP,IDE:Eclipse +JDK1.4.2,编程语言:Java、JSP、HTML、JavaScript、XML,Web 服务器:Tomcat 及其插件。

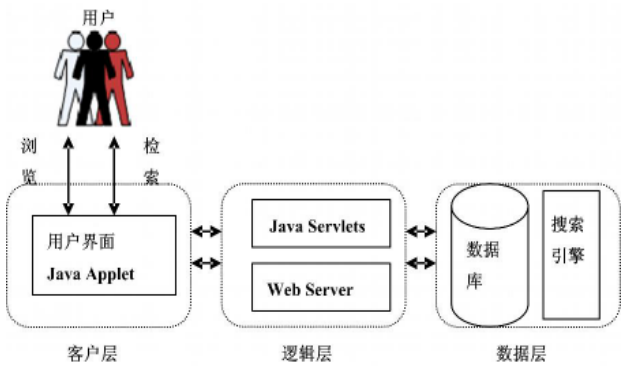

图 3　VSCIRAI 体系结构图

3.2　系统测试与评价

3.2.1 系统测试

测试数据来自于中国科学文献服务系统(ScienceChina),共 873 条记录,格式为 XML,语言为中文,学科为计算机。计算机科学发展迅速,新名词术语大量涌现,较适于通过中文自然语言处理来动态生成概念空间。

以“计算机”为检索词对系统进行测试,结果表明各项功能较好地达到系统设计目标。其中:①中文文本切分、对象过滤功能测试中,系统较好地完成了中文切分词功能,切分粒度较为合理,切分歧义问题解决较好,停用词的排除效果基本达到系统要求;②未登录词识别功能测试中,系统能将识别出的中、英文未登录词存入相应文档中,便于在领域专家的指导下,依据一定规则,定期对新词表进行统计分析,

实现词表的动态更新和维护；③概念提取与和自动标引功能测试中，系统提取出 1067 个概念词，统计总词频为 3427 次，并可依据齐普夫定律或累积频次自动选择标引词；④可视化概念空间图动态生成功能测试中，系统依据概念* 概念矩阵，利用可视化映射算法及 RadialTree 可视化显示技术实时生成可视化概念空间图，该概念空间图支持用户的交互操作，如点击，拖动，鼠标悬停等。参见图 4 及其变形图 5。其它功能如文档检索功能，矩阵生成功能和人机交互功能暂不详述。

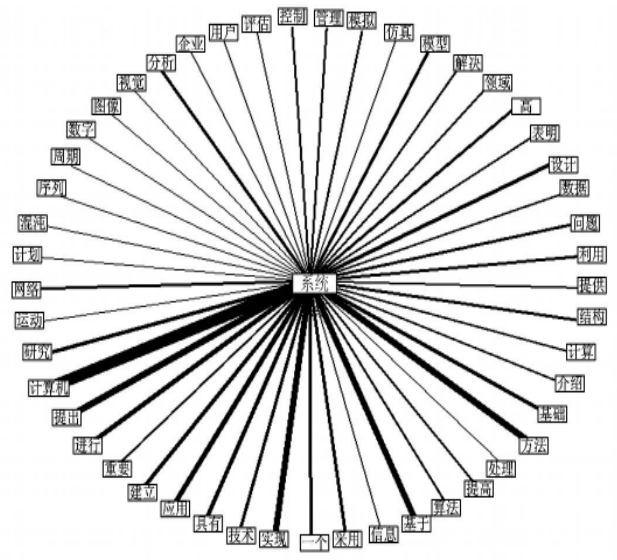

图 4 概念空间图 1

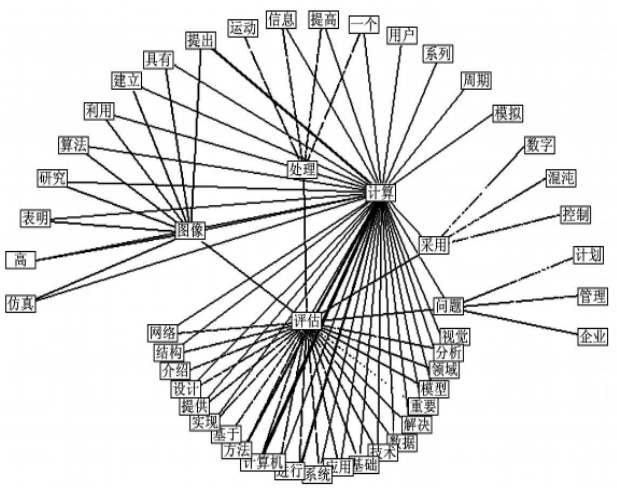

图 5 概念空间图 2

3.2.2 系统评价

依据上述系统测试结果并参照其它相应指标，从三个角度对系统进行评价。

⑴用户界面角度。主要依据 White 和 McCain 从信息可视化角度提出的对信息检索可视化界面进行评估的 8 个指标，测试结果表明基本达到要求。

⑵信息检索角度。主要的评估指标是查全率和查准率，由于海量信息条件下查准率比查全率更为重要，因此，更注重查准率的评估，测试结果表明由于 VSCIRAI 能提供概念间关联信息，有助于用户明确检索需求，其查准率显著提高。

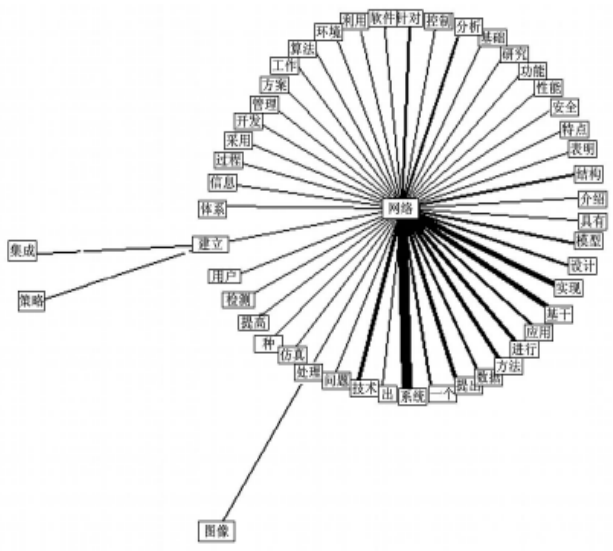

图 6 VSCIRAI 以“网络”为种子概念的概念空间图

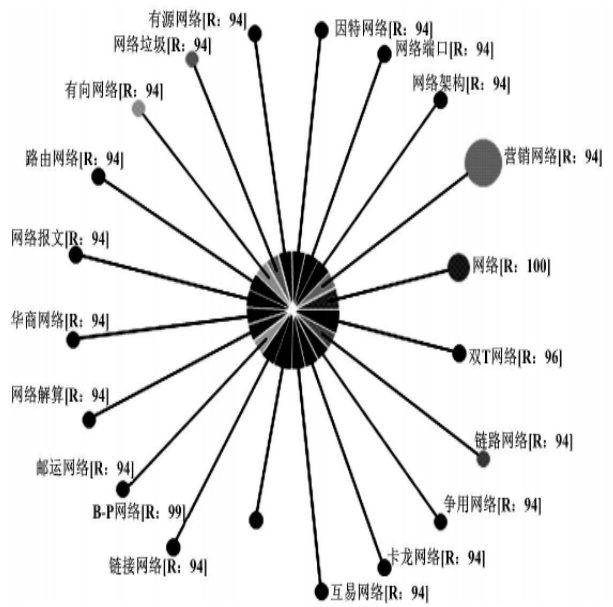

图 7 CNKI 以“网络”为检索词构建的相似词图形

⑶概念检索角度。将提取出的概念及概念间关系与《全汉语主题词表》及中国学术期刊全文数据库(CNKI)[4]的概念词典进行对比，以“网络”作为种子概念构建基于词共现的概念空间图，如图 6 所示。CNKI 以“网络”作为检索词，提供的相似词图形显示如图 7 所示。结果表明，VSCIRAI 主要有以下优势：①从概念语义角度揭示了概念间关联，有助于用户获得与种子概念“网络”有关联的概(下转第 235 页)

6 结　语

本文基于知识管理的视角，构建了企业知识创新能力的评价指标体系，应用了模糊综合评判模型对企业知识创新能力进行了评价。并根据评价结果给出了提高企业知识创新能力的建议。由于从知识管理角度评价企业知识创新能力的研究尚处在全新阶段，无论是评价体系构建和评价方法的选择方面，还是知识管理与企业知识创新的关系方面均有许多问题需要进一步探索。同时，对知识管理与企业知识创新的相互作用机制，企业知识创新系统内外部学习机制及其博弈模型等仍需要进一步的研究。

参考文献

1 吕　巍. 知识优势[M]. 北京：机械工业出版社，2002：25-38.
2 刘助柏. 论知识创新能力[J]. 机械工程学报，1999，(1)：6-10.
3 李芳春. 浅析企业知识管理的内涵[J]. 河南商业高等专科学校学报，2003，(9)：34-35.
4 郭　韬，楼　瑜. 基于复杂性理论的企业知识创新系统研究[J]. 情报杂志，2008，(4)：112-113.
5 詹湘东. 基于知识管理的区域创新能力评价研究[J]. 科技进步与对策，2008，(4)：118-119.
6 赵光州，赵立龙. 区域创新体系的知识管理[J]. 经济问题探索，2004，(3)：42-45.
7 范德成. 基于客户知识管理的企业技术创新能力评价[J]. 决策参考，2008，(4)：70-71.
8 林　山，蓝海林. 现代企业知识创新研究述评[J]. 科技管理研究，2005，(6)：37-38.

（责任编辑：孙晓明）

（上接第 230 页）念信息；②支持用户与概念空间图的实时交互，如调整共现概念数量、概念共现阈值等；③概念间关系强弱清晰可见，便于用户依据具体需求选择新概念词来优化、完善检索；④能及时反映新名词术语；⑤概念空间[5]图能反映用户输入检索词以外的概念间的关系及概念间的间接关联，利于用户了解所关注的其它概念间关联。而 CNKI 虽然基于概念词典计算概念间相似度，但存在概念间关系固定，不能依用户检索条件的不同而发生相应变化，提供的相似词数量固定为 20 个且词间关系揭示不深入，在反映新名词术语上存在时滞等不足。

4 结　语

本文设计实现了基于共词分析的中文信息检索可视化系统，并进行测试与验证。由于时间及数据集等因素，系统仍存在不足，这些不足也是下步工作的方向。主要有：缺乏对不同专业领域、大数据集数据的系统、全面验证；所采用的 Radial tree 显示技术仍存在某些问题，如不能显示共频次数，共现概念多时数据显示不清晰等，需进一步完善和探索更好的显示技术；需结合实际的词表深入研究基于词共现原理来辅助用户检索和自动标引；共现概念的可视化显示粒度的分析和测试仍需深入等。

参考文献

1 Kartoo 网站[EB/OL]. http://www.kartoo.com/，2008-04-02.
2 Nicholas J. Belkin，Colleen Cool，Adelheit Stein & Ulrich Thiel. Cases，Scripts，and Information-Seeking Strategies：On the Design of Interactive Information Retrieval Systems[EB/OL].http://citeseerx.ist.psu.edu/viewdoc/summary doi=10.1.1.48.490，2008-04-16.
3 王曰芬，宋　爽，苗　露. 共现分析在知识服务中的应用研究. 数字图书馆[J]，2006，(4)：29-34.
4 CNKI 网站 [EB/OL]. http://dlib.cnki.net/kns50/，2008-04-16.
5 邓珞华. 图书情报数学[M]. 长春：东北师范大学出版社，1983：35-37.

（责任编辑：孙晓明）