信息检索技术
清华同方光盘协办

编者按： 清华同方光盘股份有限责任公司为发展我国“信息检索技术”，在理论和实践上推动网络信息检索技术的发展与应用，以进一步加快图书情报技术网络化发展进程愿与本刊合作，协办本栏目的工作，为此编辑部代表广大读者对清华同方光盘股份有限责任公司支持我国图书情报领域计算机信息检索技术发展的举措，表示衷心的感谢！

智能检索模型研究

孔 敬[1,2]

[1]（中国科学院文献情报中心 北京 100080） [2]（中国科学院研究生院 北京 100039）

【摘要】 提出了一个智能检索形式框架模型，论述了实例化该模型的建模技术、知识表示和检索算法，对 30 个智能信息检索系统进行了模型框架、知识表示和检索算法的统计分析，总结了三种类型的智能检索模型实例化方案。

【关键词】 智能检索模型 建模技术 知识表示 检索算法 【分类号】 G354

Study on Intelligent Retrieval System Model

Kong Jing[1,2]

[1] (*Library of Chinese Academy of Sciences, Beijing 100080, China*)

[2] (*Graduate School of the Chinese Academy of Sciences, Beijing 100039, China*)

【**Abstract**】 This paper proposes a formal framework model for the intelligent information retrieval. It outlines the typical modeling method, knowledge representation and retrieval algorithm for instantiation of the given formal framework. It provides the statistic analysis of the modeling framework, knowledge representation and retrieval algorithm for 30 intelligent retrieval systems. It summarizes three kinds of solutions for instantiation of the formal intelligent retrieval model.

【**Keywords**】 Intelligent retrieval Modeling method Knowledge representation Retrieval algorithm

1 引 言

信息时代，全球数字信息以宏大的规模不断增长，涌现出海量的分布式异构信息，也导致网络上“信息过载”、“信息迷向”等问题日趋严重。构建智能检索系统实现对信息的高质量和高精度检索，已成为信息检索、人工智能、人机交互、自动化和认知科学等多个领域所关注的研究方向。本文调研分析了以上各领域有关智能检索的文献及实验系统，在此基础上着重研究智能检索模型，探讨智能检索模型实现的方法技术。

2 研究背景与动因

智能检索是信息检索和人工智能研究的一个交迭领域。人工智能研究在 20 世纪 80 年代进入了快速发展期，根据美国 1988 年的统计，在 1987 和 1988 年开发的实用化专家系统分别为 50 和 1400 个[1]。同一时期，人工智能研究成果开始引入信息检索领域，智能检索以及人工智能与信息检索间的关系被热烈讨论。如：Croft (1987)[2]在他的智能信息检索方法中讨论了人工智能领域中的专家系统、知识表示和自然语言处理之于信息检索的应用。在智能检索系统的开发中，研究者们进行了大量的人工智能技术应用尝试。例如：Brajnik (1987)[3]等开发了结合用户建模技术的专家信息检索系统；Bruandet(1989)[4]在开发的智能检索原型系统 IOTA2 中应用了产生式规则构建领域知识模型；Mozer (1984)，Belew (1986)，Wilkinson (1992) 和 Edwin (1995)[5]等分别应用各种神经网络方法开发了信息检

收稿日期：2004 - 10 - 28

索系统;Mejasson(2001)[6]等运用基于案例推理(CBR)的方法开发了用于材料和设计工程的智能检索系统;Setchi(2003)[7]等探索了自然语言处理在信息检索中的应用。20 世纪 90 年代,随着因特网的迅速发展和软件 Agent 技术的兴起,越来越多的研究者开始关注面向 WWW 的智能检索系统设计。Lee(1997)[8]等多位研究者运用 Agent 技术开发了面向 WWW 的智能信息检索系统。此外,多媒体智能检索研究在近年大量涌现,仅 2004 年在 SpringerLink 期刊全文库中就有 3 篇有关文献:Machiraju[9]等开发了应用智能信息检索技术于数字电视节目服务的系统;Čalić[10]等开发了基于内容的多媒体智能检索系统;Gasterato[11]等开发了用于航空图像检索和分类的智能系统。

以上这些研究主要分为两类:一类是人工智能基础研究,将信息检索作为应用领域。如知识表示、机器学习、推理方法、自然语言理解等基本方法在信息检索中的应用。另一类是结合人工智能和传统信息检索技术来开发智能检索系统。这些研究表明智能检索建模的技术理论研究相对分散和孤立。因此,如何根据智能检索系统的特点,抽象一个核心框架模型,将智能检索的处理机制、方法技术有机地结合起来,就成为本文探讨的主要问题。本文调研了近几年来 30 个智能检索系统,分析了它们的框架结构、知识表示和检索算法,提出了智能检索系统的形式框架和概念模型,评述了实例化该模型的知识表示和检索算法。

3 智能检索模型框架

3.1 信息检索形式模型的相关研究

Baeza - Yates 和 Ribeiro - Neto (1999)[12]提出了一个四元组的信息检索形式化模型:

$$[D,Q,F,R(q_i,d_j)] \quad (1)$$

D 是一个集合,由收藏文档的逻辑视图或表示组成。

Q 是一个集合,由用户信息需求的逻辑视图或表示组成。这些表示也称作查询。

F 是一个框架。该框架用于对文档表示、查询及其关系进行建模。

$R(q_i,d_j)$是一个相关性排序函数(ranking function)。$q_i \in Q, d_h \in D$ 分别表示一个查询和一个文档表示。它定义了文档对于查询 的相关程度顺序。

他们认为构建信息检索模型,首先要考虑文档和用户信息需求的表示。然后,根据这些表示形式,构思一个可以对这些表示进行建模的框架。这个框架同时应该是构造 $R(q_i,d_j)$ 函数的决定基础。例如,对于经典布尔检索模型,这个框架由文档集和标准集合运算符构成。对于经典向量模型,这个框架则由一个 t - 维向量空间和标准向量线性代数运算所组成。对于经典概率模型,这个框架由集合、标准概率运算和贝叶斯定理组成。

在上述模型中,只是描述了信息检索的基础模型,并没有考虑到信息检索的循环过程,针对这一问题,Griffith 和 O' Riorda (2003)[13]提出了一个结合信息检索和合作过滤模型的扩展形式框架。该框架包含了信息检索、合作过滤、相关反馈、会话和会话历史的建模。此外,20 世纪 80 年代末 van Rijsbergen(1986)[14]引入信息检索的逻辑模式。在这种模式中,文档与查询的相关性 w. r. t. 表示为一个隐含的逻辑 $d \rightarrow q$。

上述三个模型中,第一个提供了基本信息检索模型;第二个在第一个模型基础上增加了用户模型,考虑了组用户模型、相关反馈和会话,表达了信息检索的合作性、循环性、动态性和个性化;第三个在信息检索中引入逻辑模型的表达。虽然,这些模型表达缺乏对智能信息检索概念与特性的支持,但对智能信息检索模型的形式化表达提供了参考基础。

3.2 智能检索的形式框架与概念模型

本文着重于描述智能检索的主要特征和核心构成,提出了一个五元组智能检索形式化框架和概念模型的定义。

$$[U,O,D,F,R(D \rightarrow U, sim(U,D))] \quad (2)$$

U 为用户模型,也就是用户知识库,是关于用户个体属性知识的形式化明确描述。形式化指计算机可读。它用于表示用户信息需求,可用一个二元组表示为:$U = <U_c, U_f>$ 。U_c 为用户兴趣主题概念形式化表示。U_f 为用户其他属性的形式化表示,比如用户主动反馈,人机会话历史和用户知识水平等,通常表示为用户信息意图的动态调节因子。

O 为领域模型,也就是领域知识库,是专业领域中一系列概念或术语的形式化规范化表示。它由这些概念(C)及其概念间的关系(R)所构成,用一个二元组表示:$O = <C, R>$。这也是当前流行的本体论(Ontology)的定义。

D 为文档模型,也就是文档知识库,是信息源或称文档的明确描述。它用于表示文本的结构和内容等属性。可用一个二元组表示为:$D = <D_c, D_f>$,D_c 是文档主题概念的形式化表示,D_f 是文档非主题属性的形式化表示,与 U_f 相似,它是文档模型的动态调节因子。U_f 和 D_f 影响因子可作为相似性运算中表达式的加权,或推理运算的附加条件、证据等。

F 是一个框架,用于用户、领域、文档及其三者间关系的动态建模。

$R(D \rightarrow U, sim(U,D))$是求得用户信息需求相关文档的算法,包括推理运算 $D \rightarrow U$ 和相似性计算函数 $sim(U,D)$。

此智能检索形式框架中,如图 1 所示,用户模型用来捕获用户的信息需求,预测用户未来的行为,并以形式化

明确表示。领域模型用于规范和扩展用户查询,减除用户查询表达的不完整性和不确定性,同时在文档建模中用于信息源知识内容的抽取与表达。文档模型与用户模型的相关性运算和推理构成智能检索算法。三个模型相互配合完成检索过程,实现知识内容精确检索。三模型知识表达及相互关系的建模构成了检索算法选择与实现的基础。

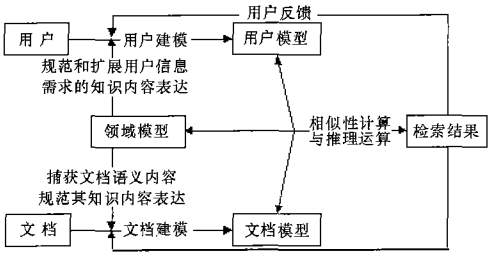

图 1 智能检索形式框架

与前述 Baeza - Yates 和 Ribeiro - Neto 基本检索模型相比,此模型针对智能信息检索特点有以下改进:一是形式框架的核心构成定义为用户模型、文档模型和领域模型,以及带推理机制的检索算法。在信息检索过程中加入了领域知识这个要素的作用,它体现了 Sparck Jones (1983)[15]对智能信息检索系统的定义,即智能信息检索系统为具有知识库和推理能力的计算机检索系统,并将知识库细分为用户知识库、领域知识库和文档知识库;二是以用户模型替代了用户查询,便于表示用户查询表达不确定性和动态性的处理,以及利用用户模型的知识构造适应性界面、主动发现用户兴趣和指导用户检索行为;三是使用包含了相似性计算函数和推理运算的检索算法替代了单纯的相关性排序函数,强调推理机制在智能检索中的作用;四是动态建模与检索。用户模型和文档模型都使用了动态建模机制,表示了信息检索的动态调整过程;五是可表达语义检索,领域模型的建立,增强了用户与文档的语义表达能力,便于实现语义检索机制。六是加入了用户对检索结果处理的反馈机制,表达了信息检索的循环过程。

智能检索形式化模型的实例化技术涉及到建模技术、知识表示和检索算法的选择,以下分别简述之。

4 建模技术与知识表示

建模是对建模对象的有关知识进行发现和明确描述,通过信息抽取,从中发现该对象的隐性知识,然后将这些知识映射到计算机内部,由系统对这部分知识进行维护。本节简述了用户、文档和领域建模的技术及其在信息检索领域中的应用情况,以及常用知识表示方法。

4.1 用户建模

用户建模已是较为成熟的研究领域,从 1997 年第六届到 2003 年第九届国际用户建模大会的论文来看,用户建模在信息检索领域的运用已有大量的理论和实践研究。用户建模按建模粒度可分为组用户建模和单用户建模,按更新方式可分为动态建模和静态建模,按建模内容来源可分为显示建模和隐式建模。其建模方法很多,如人工建模方法、基于逻辑的方法、决策理论方法、贝叶斯方法、神经网络方法、基于案例的方法、基于模型的方法、遗传算法和其他机器学习方法等等。用于建模的内容可概括为:用户静态信息、用户访问文档和人机交互信息。用户知识类型可分为:用户认知能力、用户兴趣内容和用户属性知识等。用户知识表示方法和用户建模方法与推理用户意图的匹配算法相辅相成。前两者是基础,基于不同知识表示和建模技术的用户模型有不同的匹配算法。

4.2 文档建模

文档建模的实质是对作为信息源的多种形式文档如 DOC、XML、HTML、PDF 等格式文档进行知识内容、文档结构、超链接、词频和文档格式与目录路径等的分析与描述。文档建模的概念较少有人提及。但在信息科学各个分支领域,对文档的内容和结构等多种属性的信息抽取技术与方法已有广泛和深入的研究。其中利用领域概念模型来捕获信息源的语义内容已成为近来的研究热点。如 Brasethvik(2001)[16]提出了一个语义文档建模的方法,该方法基于自然语言分析和概念建模来实现语义文档的分类与检索。另一方面,在信息抽取技术未有新的突破之前,如中文词的切分至今也未能很好解决,采用加入文档结构、超链接和文档目录路径等特征的文档模型将补充和优化单纯基于内容和语义文档模型的性能。例如在文档结构分析方面进行元数据提取,从文档数据库中发现结构;利用引文分析,作者共引分析,结合聚类算法进行作者聚类、文档聚类实现部分智能检索[17];使用网页文档的 URL 深度作为是否为内容页的预测因素之一[18],则是将文档目录路径作为建模内容的例子。

4.3 领域建模

领域建模是以公开说明方式构造某一特定域概念、关系以及有关知识的形式化或非形式化明确模型。“概念化”的领域知识定义,包括对应用问题和其解决过程所必需的概念、属性、关系、启发式规则和实例等的定义[19]。领域建模也是一个较为成熟的研究领域,其建模技术与方法在已有 20 多年历史的概念建模国际研讨会中

得以深入研究。2003 年第 22 届研讨会重点研讨了面向未来信息系统的概念建模问题。在领域建模工具方面，近十年兴起的 Ontology 是一种能在语义和知识层次上描述信息系统概念模型的建模工具。Ontology 自被提出以来就引起了极大关注，在包括信息检索的众多领域中广泛应用。如 Tawil A - R(1997)在智能信息检索中利用 Ontology 规范用户查询和捕获信息源的语义内容[20]。形式化 Ontology 的核心技术是描述逻辑(Description Logic)和框架逻辑(Frame - Logic)。关于领域建模技术方法论的研究一直是前沿的理论研究，涉及内容众多，在此不作详细介绍。

4.4 知识表示

人工智能领域常用知识表示方法有：一阶谓词逻辑表示法、产生式表示法、框架表示法、语义网络表示法、面向对象表示法。而实际在信息检索领域中运用的知识表示方法远远不止这几种方法。以下列举了近年来信息检索领域常用的知识表示法。

(1)向量空间模型。基于表示事物特征的关键词抽取特征值。由这些特征值构成一个 t 维向量。信息检索中常将文档的关键词和用户查询词串以 t 维向量的形式表示。

(2)概率模型。在信息检索中，可先建立一个领域分类模型，然后计算所有文档和用户兴趣在这个分类模型上的概率分布，用该概率分布来表达文档和用户兴趣的主题。

(3)产生式规则。产生式规则通常用于表示事物间的启发式关联，其基本形式为：P⇒Q。P 为规则激活使用的条件(或称前提)，Q 则指示规则激活时应该执行的动作(或结论)。基于规则库推理的检索常用产生式来表示用户信息需求和领域知识库。

(4)语义网络或概念图。语义网络(或概念图)用于表示事物(或概念)间的关系，它通常表示为有向图。图的节点指示事物(或概念)，节点间以有向弧连接，而弧上的标签则指示节点间关系。

(5)面向对象。面向对象的方法是将一组数据和与该组数据相关的操作封装在一起成为对象，同时将具有相同特征的对象抽象为类。面向对象的表示法与框架表示法有许多相似之处，如层次分类体系和特性继承机制等。

(6)框架。框架表示法是一种关于事物内部结构化描述的表示法。通常由描述事物各个方面的槽组成，每个槽有多个侧面，侧面又可有多个值。

此外，还有形式逻辑、决策树、神经网络和案例等知识表示法也在信息检索领域中尝试使用。

5 智能检索算法

本文根据智能检索特点，将智能检索运算的算法分类为：基于函数的相似性计算和基于过程的推理算法。常用的相似性计算函数包括：向量余弦夹角、贝叶斯定律和基于相似矩阵的方法等等。下面重点介绍信息检索中常用的推理算法。

(1)推理网络。Turtle H R(1991)[21]提出了使用概率推理网络即贝叶斯网络来表示文档和用户信息需求。在此检索模型中，信息检索被看作成为证据推理过程，关于文本和查询内容的多个证据源被结合起来评估文档与查询的匹配概率。

(2)模糊推理。有多种模式，其中最常用的是基于模糊规则的推理。模糊规则的前提是模糊命题逻辑组合(合取、析取和取反操作)，用作为推理的条件；结论是表示推理结果的模糊命题。所有模糊命题成立的精确程度均以相应语言变量定性值的隶属函数来表示。

(3)神经网络。将用户的查询内容和文本表示为因果关系，利用神经网络的方法来进行用户需求和信息资源的匹配运算。常用的神经网络方法有：多层感知器、射线基本函数、贝叶斯网络和 Kohonen 网络。

(4)基于规则的推理(RBR)。知识一般表示为产生式规则库。规则在运行过程中环环相扣，形成复杂的推理网络，知识在推理网络中得以传递，进行相应的分析和判断。

(5)基于案例的推理(CBR)。知识表示为大量已有问题及其解答的案例。从案例中推导出未知问题的解答。

此外，信息检索领域中使用的检索算法还有框架推理、聚类算法和自然语言处理等等。

6 案例分析

本文分析了 30 个系统案例，选自于 IEEE/IEE Electronic library、ACM Digital Library、Elsevier ScienceDirect、ProQuest 博士论文全文库、SpringerLink 期刊全文库、中国学术期刊全文库、维普中文科技期刊库和中国科学院学位论文库。选择标准为 1997 - 2003 年发表，其系统设计进行了较为详细描述，自标引为智能检索或智能信息检索的文献。本文统计分析了前述用户模型、文档模型、领域模型、知识表示、检索算法在这些系统的应用情况。

结果表明，这些系统中约 63% 构建了用户模型，57% 构建了领域模型，57% 构建了文本模型，同时构建二种以上模型约有 70%，但同时构建三种模型则仅有 13%。此外，只是运用了一些推理方法，三种模型均未采用约有 7%。这说明在智能检索系统中，综合构建这三种模型的设计思想还未得到足够重视。并且，各模型现有研究成果在智能检索系统开发中应用的层次和深度较浅，例如用户建模的许多成熟技术与成果未得到充分运用。

在采用知识表示法方面，加权关键词向量用的最多，约占 40%，其次为概率模型，约占 17%。然后是语义网

络(概念图)、面向对象、产生式规则,各占 10%。其他用到的知识表示方法还有决策树、框架等。

在检索算法运用方面,约 27% 构造了相似性计算函数,其中多为向量夹角余弦函数,也有其它矩阵运算函数和贝叶斯公式。80% 的系统都采用了不同程度的推理运算。其中模糊推理最多(5 例)。其次是神经网络 3 例,聚类算法、基于案例的推理、基于规则的推理和自然语言处理各为 2 例。其它推理方法不再列出。

30 个案例系统可归纳为四种类型:个性化服务系统、推理算法研究系统、语义检索系统和问题解答系统。多数系统是智能检索系统某一环节解决方案的实验系统。

综合以上分析,本文总结了以下几种智能检索模型实例化方案。

(1)成熟型

方案特点:采用已广泛应用的建模技术和检索方法,如:基于特征抽取的简单易用的知识表示方法,成熟的相似性函数与推理方法。

用户模型和文档模型知识表示:加权关键词向量或概率模型。

领域模型知识表示:加权关键词向量、层次结构的术语词典或面向对象的类。

相似性函数:向量夹角余弦函数或贝叶斯公式。

推理算法:模糊推理或聚类算法。

简评:基于特征抽取的知识表示法和检索算法,方便实现,检索速度较快。但缺乏对语义内容的理解,检索的精度有待提高。

典型案例:TU(2000)[22]等开发,基于知识分类和体系结构的智能信息检索 Agent。

(2)语义型

方案特点:接受用户自然语言的查询表达,基于语义内容的检索,查询多语种信息。

用户模型和文档模型知识表示:语义网络或概念图。

领域模型知识表示:语义网络 + 规则知识库 + 语言词典库。

推理算法:基于规则推理 + 神经网络方法 + 知识查询操作语言 + 自然语言处理。

简评:基于语义内容的检索,采用自然语言处理技术和语义表示法更能精确表示用户意图。但其最大缺点就是响应时间慢。实际应用可结合第一种类型的检索方法处理大数据集,得到预处理结果集,再根据用户兴趣,锁定小部分数据集作语义内容检索,提高查询精确度。典型案例:Setchi(2003)[7]等开发,基于深度自然语言处理的信息检索系统。

(3)问题求解型

方案特点:面向专业构建领域专家,采用求解问题的知识型方法,解答用户咨询问题。

用户模型和文档模型知识表达:框架、案例或规则。

领域模型知识表示:案例知识库 + 规则知识库。

相似性函数:向量夹角余弦函数或贝叶斯公式。

推理算法:基于案例的推理(CBR)和基于规则的推理(RBR)。

简评:CBR 和 RBR 均为求解问题的知识强方法。与 RBR 相比,CBR 在知识表示、知识复用等方面有优势,常用于工程设计、医疗诊断领域。实际系统中常用两种方式混合推理。

典型案例:Montani(1999)[23]等开发,基于万维网的知识管理和决策支持系统,用于医疗诊断咨询。当用户输入患者病历时,系统通过贝叶斯公式分类统计方法匹配病案库中的案例供用户选择,然后再根据用户选择,使用基于规则的推理方法,匹配相应规则和有关参数得出最终结果。

7 结 语

本文研究了智能检索系统形式模型和模型实例化的方法与技术,提出了智能检索系统形式框架的核心组成有三个基本模型(用户模型、领域模型、文档模型),以及集成这些模型的检索算法(相似性计算与推理机制)。在智能检索概念模型中,明确定义了三个模型,描述三模型之间的相互作用和检索机制,为智能检索系统的设计提供了良好的参考模型。简述了智能系统模型的技术与方法,结合 30 个智能信息系统案例的统计分析,评述了用户模型、文档模型、领域模型、知识表示和检索算法在智能检索系统的应用情况,总结了几种类型的模型实例化方案。然而,这些实例化方案构筑在他人开发的智能检索系统上,所提出的智能检索形式模型与方法需要亲自开发实验系统验证。因此未来的工作拟采用 Agent 技术开发实验系统,对智能检索建模技术、知识表示法和检索算法进行实验研究。

参考文献:

1 高济,朱森良,何钦铭. 人工智能基础. 北京:高等教育出版社,2002:8

2 Croft W B. Approaches to intelligent information retrieval. Information Processing & Management, 1987,23(4): 249 - 254

3 Brajnik G, Guida G, Tasso C. User modeling in intelligent information retrieval. Information Processing & Management, 1987, 23(4): 305 - 320

4 Bruandet M. Outline of a knowledge - base model for an intelligent information retrieval system. Information Processing & Management, 1989, 25(1): 89 - 115

5 Cortez E M, Park S C, Kim S. The hybrid application of an inductive learning method and a neural network for intelligent information retrieval. Information Processing & Management, 1995, 31(6): 789 -

813

6 Mejasson P, et al. Intelligent design assistant (IDA): a case base reasoning system for material and design. Materials & Design, 2001, 22(3): 163 - 170

7 Setchi R, Tang Q, Cheng L. Information Retrieval Using Deep Natural Language Processing. In: Palade V, Howlett R. J, Jain L C, ed. KES 2003, LNAI 2773. Berlin Heidelberg: Springer - Verlag, 2003, 879 - 885

8 Lee C, Chen Y. An embedded visual programming interface for intelligent information retrieval on the Web. In: Knowledge and Data Engineering Exchange Workshop, 1997. Proceedings. 1997 : 46 - 53

9 Machiraju C, Kanda S, Dasigi V. Application of Intelligent Information Retrieval Techniques to a Television Similar Program Guide. In: Orchard R et al. ed. IEA/AIE 2004, LNAI 3029. Berlin Heidelberg: Springer - Verlag, 2004: 788 - 796

10 Čalič J, et al. ICBR - Multimedia Management System for Intelligent Content Based Retrieval. In: Enser P et al. ed. CIVR 2004, LNCS 3115. Berlin Heidelberg: Springer - Verlag 2004: 601 - 609

11 Gasteratos A, Zafeiridis P, Andreadis I. An Intelligent System for Aerial Image Retrieval and Classification. In: Vouros G A, Panayiotopoulos T, ed. SETN 2004, LNAI 3025. Berlin Heidelberg: Springer - Verlag, 2004: 63 - 71

12 Baeza - Yates R, Ribeiro - Neto B. Modern Information Retrieval. New York: ACM Press, 1999:23

13 Griffith J, O'Riordan C. A Formal Framework for Combining Evidence in an Information Retrieval Domain. In: Palade V, Howlett R J, Jain L C, ed. KES 2003, LNAI 2773/2003. Berlin: Springer - Verlag. 2003: 864 - 871

14 van Rijsbergen C J. A non - classical logic for information retrieval. The Computer Journal, 1986, 29(6): 481 - 485

15 Sparck Jones K. Intelligent retrieval. In: Jones, K P, ed. Intelligent Information Retrieval: Proceedings of Informatics 7. London: ASLIB, 1983:136 - 142

16 Brasethvik T, Gulla J A. Natural Language Analysis for Semantic Document Modeling. M. In : Bouzeghoub Z, Kedad E, Métais Eds. NLDB 2000, LNCS 1959/2001. Berlin, Heidelberg: Springer - Verlag, 2001 : 127 - 139

17 He Y, Hui S. Mining a Web Citation Database for author co - citation analysis. Information Processing and Management, 2002, 38 : 491 - 508

18 Zhu T, Greiner R, Haubl G. Learning a Model of aWeb User's Interests. In: Brusilovsky P, et al. Eds. UM 2003, LNAI 2702. Berlin, Heidelberg: Springer - Verlag, 2003: 65 - 75

19 徐振宁,张维明,陈文伟. 基于 Ontology 的智能信息检索. 计算机科学, 2001, 28(6): 21 - 26, 44

20 Tawil A - R, Behrendt W. Requirements for components of an intelligent information retrieval model for the WWW. In: Intelligent World Wide Web Agents (Digest No: 1997/118), IEE Colloquium on, 17 March 1997:1 - 7

21 Turtle H R. Inference Networks for Document Retrieval. UMI, 1991

22 Tu H, Hsiang J. An architecture and category knowledge for intelligent information retrieval agents. Decision Support Systems, 2000, 28 (3): 255 - 268

23 Montani S, Bellazzi R. Integrating case based and rule based reasoning in a decision support system: evaluation with simulated patients. In: Proceedings of the 1999 AMIA Annual Symposium. Philadelphia: Hanley and Belfus, Inc. 1999 : 887 - 891

(作者 E - mail:kongj@mail.las.ac.cn)

(上接第 49 页)

这主要是由于 Scirus 搜索的是种子表内的站点,如大学网站等,而这些学术性、教育性网站上仍可能存在着非学术内容,如校园内有关行政人员的信息、BBS 上的言论等。所以,Scirus 检索的结果中仍然会有无关信息,甚至是人们在网络上表达不满情绪的一些激烈用语。

4 结 语

虽然 Scirus 是一个"年轻的"专业科学搜索引擎,毕竟它是从 2001 年开始启用的。而与 Google 相比,Scirus 对于专业的科学信息的搜索具有明显优势。Scirus 通过独特的倒置金字塔工作流程来准确锁定科学信息,这一定程度上保证了 Scirus 搜索的专业性和查准率。这也是其能够快速准确找到人们所需科学信息的原因之一。但是,Scirus 还有一些地方有待改进,如开发多种语言的搜索界面,过滤检索结果中的无关信息、优化在线词表、完善主题分类等。

参考文献:

1 http://www.searchenginewatch.com(Accessed Mar. 5,2004)

2 SCIRUS(Elsevier 科学出版社开发的科技文献门户网站)使用指南. http://www.lib.zjut.edu.cn/free_bus/1217SCIRUSElsevier.htm (Accessed Mar. 17,2004)

3 Femke Markus. Scirus: The search engine for scientific information only. ACS Chicago,27 August 2001. http://www.lib.uchicago.edu/cinf/222nm/presentations/222nm029.pdf (Accessed Mar. 17,2004)

4 http://www.scirus.com(Accessed Mar. 15,2004)

5 SCIRUS White Paper: How Scirus Works. http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf (Accessed Mar. 16,2004)

6 张捷,王娟萍. 科学搜索引擎——SCIRUS 的检索模式与评述. 津图学刊,2003(4):70 - 72

(作者 E - mail:jennifercn2004@sohu.com)