

●张玉峰 李 敏

# 网络信息检索结果的智能处理方法 \*

**摘 要** 对网络检索结果信息的智能处理方法,包括知识提取方法和知识集成方法。知识获取模式是实现智能化信息服务的有效途径。参考文献 13。

**关键词** 网络信息 检索结果 知识获取 知识集成

**分类号** G250.73

**ABSTRACT** In this paper, the authors think that intelligent processing methods of network information search results include the one of knowledge acquisition and the one of knowledge integration, and the pattern of knowledge acquisition is an effective approach to the realization of intelligent information services. 13 refs.

**KEY WORDS** Network information. Search results. Knowledge acquisition. Knowledge integration.

**CLASS NUMBER** G250.73

## 1 网络信息检索结果处理的现状分析

网络信息数量海量化,形式多样化,内容复杂化。如何在信息海洋中有效获取和利用信息与知识呢?国内外众多的搜索引擎应运而生。著名的英文搜索引擎有 YAHOO、INFOSEEK、EXCITE、AL TAVISTA 等,中文搜索引擎有 SOHU、GO YO YO、YAHOO 等。这些网络搜索工具为用户获取有效信息做出了卓越贡献。但是随着数字信息资源的开放化、多媒体化和异构分布化,现有的搜索机制显得被动无力。网络信息检索的原始结果仍是大量信息的堆积,不能满足广大用户的要求。被困在毫无头绪的信息海洋中的用户终会因厌倦而丧失对这种查询途径的信心。所以如今的搜索引擎均对检索结果采用一定的处理技术,以使其能突出重点,更为有序化。

### 1.1 相关度排序

#### 1.1.1 词频排序

这是一种基于内容的排序方法,即若某站点或页面包含的某关键词越多,则对于这一主题的相关度就越高。然而由于各站点的质量水平不一致,关键词出现的多少并不能完全代表它的相关度。而且一些站点经营者为了某种目的,在本站点的正文中有意加大某些关键词的使用量,这种欺骗行为同样会使搜索引擎得出错误结论,给用户带来信息垃圾。

#### 1.1.2 访问量排序

许多站点都设有自己的访问量计数器,以此显示自己的重要性及吸引力。搜索引擎会记录它所搜索到的页面的

被点击次数,从而判断该页面被访问的频率。这是一种由公众集体确认网站重要性的方法,具有一定的客观性与公正性。但也有些站点经营者通过点击奖励、网上游戏等方式获得高访问量。

#### 1.1.3 基于超链接的排序

Web 中蕴含着丰富的结构信息。页面之间的超链反映了某种引用关系,通过构造 Web 页面引用图,综合考虑页面的被引用次数以及所引用页面的相关度来判断本页面的重要性<sup>[1]</sup>。

### 1.2 转换过滤

这是按照一定的条件对搜索结果再进行优化、过滤的过程,如信息格式的支持与转换<sup>[2]</sup>,采用信息过滤技术减少重复信息和垃圾信息<sup>[3]</sup>,应用聚类技术对检索结果进行联机聚类等<sup>[4]</sup>。

随着互联网的发展壮大,用户的信息需求也在不断改变。一般来说,互联网用户的知识水平层次较高,他们对网络信息检索的要求也与一般信息用户不同。他们需要的是经过深层加工、信息量充分、内容精准、易于理解吸收的检索结果。这样既可以很好地解决问题,又可以节省宝贵时间。但如今的搜索引擎对网络信息的检索及其结果处理并不能满足广大用户日益提高的要求。因此需要寻求更高层次的检索结果处理方法。

## 2 网络信息检索结果的知识提取

网络搜索引擎是按照一定的搜索程序对信息进行辨别与搜集的,因此它具有机械性,不能超越设计者的思维模式。

\*本文为国家社会科学基金资助项目“基于学习的智能检索机制研究”(编号:01BTQ011)成果。

网络信息用户在缺乏恰当或必要的概念和语词符号时,使得检索提问具有一定的模糊性,降低了检索结果的满意度。因此,通过对用户需求的分析研究,跟踪用户的兴趣爱好,建立面向领域的用户需求模型库,对网络信息检索结果进行一定程度上的知识提取是必要的。

## 2.1 知识提取方法

人们通过网络搜索引擎所得到的检索结果只是表层知识内容,而对蕴含其中的潜层知识无法获知。用户希望在信息检索结果中,既可以得到表层知识,又能获得潜层知识。所以需要检索结果实施潜层知识的提取。

所谓知识提取,即知识获取,就是对信息源进行智能处理,从中抽取出所需知识,并将其转换成计算机程序利用的表示形式,如事实、规则及框架的集合。其中知识源主要指专家、书刊、数据库和网络信息资源。

知识提取的方法有很多种,包括:将人类专家的经验知识精炼转换为共享知识的方法;从经验数据、实例、数据库以及出版物中获取知识的各种学习方法等。提取方法大致可以分为采访、机助采访、机器学习、机器感知等<sup>[5]</sup>。

自动化知识提取方法之一是机器学习,即根据搜索引擎捕捉到的地址找到具体内容所在页面,并且根据一定的经验规则对其内容进行分析、综合、挖掘和归纳,最终获取知识。机器学习包括获取新知识、精炼知识库、智能查找、理论创新等多方面的学习能力<sup>[6]</sup>。机器学习方法包括记忆学习、传授学习、演绎学习、归纳学习和类比学习等。机器归纳学习是系统根据有关的数据或实例,应用归纳推理推导出一般性规则或结论。它可以迅速产生知识库,是一种实验成功的、有效的自动化学习方法。本文采用基于实例的机器归纳学习方法,实现对检索结果的知识提取。

## 2.2 知识获取模式

检索结果的知识提取,是一种多维学习处理,需要学习专家的有关经验知识和用户模型知识,作为结果处理的支撑知识,这里不加细说。本文主要围绕检索结果的知识提取过程,探讨其知识获取模式。知识获取应当建立在有序的检索结果上,依据用户需求,选择重要结果或全部结果,执行智能处理。

知识获取模式为:首先依据用户需求模型,从搜索引擎检索结果中选取典型实例集;利用规则库中的知识,对典型实例集执行机器归纳学习,产生粗略知识集;进而检测知识元素与用户需求的概念隶属度,求出最终的知识集合,将其反馈给用户,并把它汇入知识提取库,以供以后知识提取参照之用。知识集最终反馈给用户后,若用户满意则作最后的提交,否则修改用户模型,更新用户需求模型库。

### 2.2.1 规则库

规则库是系统执行推理解决问题的核心部分,主要包括实践经验知识,即启发式知识。它描述专家关于知识提取的智能行为与决策规则<sup>[17]</sup>。系统不断地从领域专家、知

识工程师及用户那里获取关于知识提取的经验规则,构造规则库,用于支持机器归纳推理。

### 2.2.2 用户模型库

通过询问、观察、分析和学习处理模型化用户的偏好特点和背景知识,建立用户模型库。基本步骤为:通过交互获得用户信息,首先建立用户类的原型模型,进而逐步建立个性化用户模型。

### 2.2.3 机器归纳学习

首先对已知的检索结果实例进行预处理,转换为适当的形式,例如正例集和反例集。然后采用数据驱动方法搜索规则库,依据归纳算法执行归纳推理,产生新的结论。归纳过程中,可根据用户模型评估精炼新知识。然而此时所得到的知识集合不一定是理想的结果,可能是分散的知识元素,必须对各知识元素进行有效地集成与整合。

## 3 网络信息检索结果的集成

由于网络信息检索结果的数量大、质量低,影响了人们对信息的利用,对知识的创新。因此信息集成与知识集成被正式提出,其目的是通过共享标准的综合数字信息来实现资源的高效组织。“集成”一词有综合、结合的含义,然而在信息集成与知识集成中不是如此简单。

### 3.1 信息集成与知识集成

信息集成是指针对某一特定领域或特定组织机构,以信息为对象、信息资源为本体、集成服务为动力、网络技术为手段、协同作业为方法,以构建资源保障体系为目标,把信息资源诸要素有机地联结成一个整体的动态过程,也是优化要素、体系重构的过程<sup>[8]</sup>。

知识集成是指将分散的知识元素依据一定的逻辑规则有机地结合在一起,使知识有序化、层次化,从而高效地利用信息资源,有利于知识创新,促进不同思想的撞击以及观点的集成<sup>[9]</sup>。知识集成的难度比较大,因为不同的信息源会提供关于同一信息的不同观点,或者它们是来自同一群集(Population)的不同样本<sup>[10]</sup>。

知识集成与信息集成有明显区别<sup>[11]</sup>。信息集成是为了信息的有效利用,对数据库进行检索、排序、统计等。其过程是确定性的、可重复的结构化过程。知识集成是为了知识的创新而对知识进行挖掘、共享和集成。其过程是不确定性的、不可重复的非结构化过程。但是二者之间又有密切联系。信息集成以知识集成为目的,知识集成以信息集成为基础。网络搜索引擎对网络信息资源的搜索及处理,实质上就是一种信息集成工作,并通过对检索结果信息的排序、转换、过滤等处理,为知识集成打下基础。

### 3.2 网络信息检索结果的知识集成方法

信息检索结果经过以上的知识获取后,得到一个知识元素的集合,这些知识元素可能是离散的、无序的,需要进一步的知识集成处理。

知识集成可包括对知识元素的排序、计算、分析与综合、分类整合、知识抽象及知识推理等。由于需要集成的知识具有复杂的语义和结构,而且知识的来源不同,因此各知识元素既具有分散性,在语义结构上还存在一定的耦合:同一知识内容可以有多种表达方式,而不同的知识内容可能会在其某一侧面有相同的表达。因此,对知识的集成是一项复杂繁琐,并且精确性要求高的工作。

### 3.2.1 结构化集成

由广义的知识定义得知,数值性知识、符号性知识、样本性知识是主要的异构性知识。它们具有不同的特征,单纯将它们集成在一起有一定的难度。可以通过进行知识描述形式之间的相互转换达到知识集成的目的<sup>[12]</sup>。对于异构性知识的集成,可以利用全局性的知识表达模式,对知识进行格式转换、处理及组合性集成。随着需求的增长,可以不断扩展、完善该模式。但是由于缺少语义内容上的集成,所得到的结果缺乏深度。

### 3.2.2 基于特征集成

无论是表层知识还是潜层知识,均有不同的特征。其特征可以分为:描述性知识与操作性知识;独立性知识与半独立性知识;关于概念定理的知识与关于图像声音的知识等。相关知识元素具有相同的特征值,可以通过构建知识特征库,进行特征自动识别,将知识聚类集成。

### 3.2.3 模型驱动集成

应根据用户的实际要求及各知识元素本身的客观特征,进行不同层次的知识集成。该方法需要构建不同层次的集成模型,解决在一定的要求下集成什么、如何集成的问题,然后通过模型映射实现知识集成。

### 3.2.4 面向对象的知识集成

对于多种不同的知识集成要求,需要考虑知识元素之间存在的多种关联,实现多元集成。采用面向对象的集成模式较为理想。该集成模式可以表达更多相关的语义,支持抽象机制,具有可扩充性和可重用性,有利于知识的重构和知识创新。

## 3.3 知识集成标准化

知识集成的结果需要具有通用性,否则只是局部解决了知识离散的问题,同样会不可避免地给知识共享带来障碍。对于知识集成来说,集成标准化极为重要,尤其是在知识定义、知识格式(Formats)以及互操作网络协议(Interoperable Network Protocols)等方面<sup>[13]</sup>。只有各方面同时达到标准化,才能得到有效的、高质量的知识集成产品。

## 4 结束语

现代智能技术的水平仍然有限,但已经给人们带来智力上的解放。机器智能化的最终目标不是要完全取代人类的大脑,而是要成为它的辅助工具,使其有更大的空间扩展智能,更深层次的发掘潜能。搜索引擎的检索结果对于大多数人来说仍然是无序的、不易接受的,对其进行知识提取和集成等智能化处理,将产生不可估量的效果与效率。这是人们梦寐以求的理想,也是实现智能化信息服务的重要途径。

## 参考文献

- 1, 4 王继成, 萧嵘等. Web 信息检索研究进展. 计算机研究与发展, 2001, 38(2)
- 2 孙建军. 网络信息资源搜集与利用. 南京: 东南大学出版社, 2000
- 3 傅忠谦, 王新跃等. 个性化网上信息过滤智能体的实现. 计算机应用, 2000, 20(3)
- 5, 6, 7 张玉峰. 智能情报系统. 武汉: 武汉大学出版社, 1991
- 8 霍忠文, 李立. 把握“占有”重点“集成”. 情报理论与实践, 1999, 22(5)
- 9 Davide Ravasi, Gianmario Verona. Organising the process of knowledge integration: the benefits of structural ambiguity. Scandinavian Journal of Management, 2001, 17(1)
- 10 B. Scotney, S. McClean. Efficient knowledge discovery through the integration of heterogeneous data. Information and Software Technology, 1999, 41(9)
- 11 顾新建, 祁国宁. 知识集成初探. 计算机集成制造系统 CIMS, 2000, 6(1)
- 12 肖人彬, 苏牧等. 一种实现异构知识集成的新方法. 华中理工大学学报, 1999, 27(2)
- 13 Judy E. Scott. Organizational knowledge and the Intranet. Decision Support Systems, 1998, 23(1)

张玉峰 武汉大学信息管理学院教授. 通讯地址: 湖北武昌武汉大学. 邮编 430072.

李 敏 武汉大学信息管理学院硕士研究生. 通讯地址同上.

(来稿时间: 2001-09-26)