<<信息检索>>

图书基本信息

书名：<<信息检索>>

13位ISBN编号：9787115235756

10位ISBN编号：7115235759

出版时间：201008

出版时间：人民邮电出版社

作者：David A.Grossman,Ophir Frieder

页数：230

译者：张华平

<<信息检索>>

内容概要

本书是“信息检索”课程的优秀教材，书中对信息检索的概念、原理和算法进行了详细介绍，内容主要包括检索模型与算法、检索实用策略、跨语言信息检索、查询处理、融合结构化数据和文本、并行信息检索以及分布式信息检索等，并给出了阐述算法的大量实例。

本书有一定的广度和深度，而且所有的内容都用当前的技术阐述，是高等院校计算机及信息管理等专业本科生和研究生的理想教材，对信息检索领域的科研和技术人员也是很好的参考书。

作者简介

David A.Grossman 佐治亚梅森大学博士，现在伊利诺伊理工大学计算机系任教。
曾在美国政府部门高级技术服务中心和研究发展办公室担任项目经理。
主要研究领域包括信息检索、结构化和非结构化数据集成以及数据挖掘。

Ophir Frieder 乔治敦大学教授、计算机科学系主任。
曾任伊利诺伊理工大学计算机系首席教授、学院数据检索实验室主任。
ACM会员，IEEE和美国艺术与科学研究院高级会员。
他在数据检索系统、通信系统、高性能系统结构等方面均有深入的研究。

<<信息检索>>

书籍目录

章节摘录

8.4 P2P信息系统　　现在，我们来关注一个新兴的领域，它是互联网领域和信息检索的交叉领域，即P2P体系结构。
P2P体系结构是分布式环境，根据其定义，认为网络上的每个节点都是潜在的信息源（服务器），也是需求信息的客户端（客户端），同时也是信息传播的中间路由器（路由器）。
每个节点都是独立的，而且系统以纯粹无中心的方式运行。
而在信息检索系统领域，所提供的资源都是以可检索数据的形式呈现的。
P2P系统最主要的特点就在于其天然的随机性以及耐久性。
P2P系统可以从容地处理系统中节点的加入与离开。
这些节点提供的资源都可以根据需要在系统中动态添加或者删除。
此外，单一节点的故障不会导致整个系统崩溃。
P2P运动的起源通常要归功于Napster（它是一个音乐文件共享系统），尽管Napster实际上依靠的还是依据中心集中式方式而实现的。
也就是说，Napster并不是以完全无中心的形式存在的，因此，这并不是一个纯粹意义上的P2P体系。
但是，Napster确实为用户提供了P2P的功能，因为用户可以与他人动态地共享文件。
Napster从性能和可靠性的角度看存在一些争议，除此之外，Napster的集中式实现模式最终也注定了它会遇到法律问题。
现在，Napster再也不能以其原来的形式存在了。
Napster的灭亡给P2P技术爱好者带来了教训。
作为回应，他们创建了Gnutella协议[V0.4，2004]，这是真正的P2P，是许多当今P2P研究的基础。
（Gnutella协议[V0.6，20041的后续版本也存在，并扩展了P2P的体系结构，引入了层次结构。
该协议及其应用稍后讨论。
）　　基于Gnutella（版本0.4）协议的系统一般只提供了最原始的搜索能力。
也就是说，它们一般依赖于名称精确搜索，而名称精确搜索往往通过子串匹配来实现。
具体来说，如果查询中的所有词是某个文件元数据的子串，那么查询就匹配该文件。
匹配的文件按照其元数据与查询的相似度进行分组，最终返回到客户端。
目前，我们还没有可为大家接受的方法来对这些分组进行排序。
用户选择其中的一组结果，从相应的服务器下载相关文件[Rohrs，2000;Rohrs，2001]。
大多数基于Gnutella协议（版本0.4）建立的P2P信息检索系统依然存在其他问题，其中包括：网络全局泛滥问题、搜索结果不确定而且精度差。
因为从定义上看，P2P信息检索系统是无中心的，所以对于每一个检索请求，信息将被发送到所有有可能含有相关文档的节点上。
因为每个节点搜索与文档排序的计算能力是有限的，所以任何潜在的相关文档都会发送到请求的节点上。
给定潜在相关文档的数目，结果信息的网络总流量一般会大大高于已有的网络资源。
这种情况就称为网络全局泛滥。
为了减少信息传输量，文献[Yu等人，2003]研究了一种结果过滤和合并的技术，而且节点一般都会收集邻居节点的信息。
因此，查询请求节点所需的潜在结果到达中间节点时，只将合并和过滤过的结果返回到请求节点。
层次P2P网络的最新合并技术可参见文献[Lu和Callan，2004]。
……

版权说明