中国浦东干部学院博士文库

基于对等计算的信息检索技术

凌 波·著

Information Retrieval Based on Peer-to-Peer Computing

上海社会科学院出版社

中国浦东干部学院博士文库

基于对等计算的信息检索技术

凌　波·著

Information Retrieval Based on Peer-to-Peer Computing

上海社会科学院出版社

图书在版编目(CIP)数据

基于对等计算的信息检索技术/凌波著. —上海：上海社会科学院出版社，2007.9
（中国浦东干部学院博士文库）
ISBN 978-7-80745-095-5

Ⅰ. 基… Ⅱ. 凌… Ⅲ. 计算机网络—情报检索 Ⅳ. G354.4

中国版本图书馆 CIP 数据核字(2007)第 143530 号

基于对等计算的信息检索技术

著　　者：凌　波
责任编辑：徐祝浩
封面设计：王斯佳
出版发行：上海社会科学院出版社
上海淮海中路 622 弄 7 号　电话 63875741　邮编 200020
http://www.sassp.com　E-mail:sassp@sass.org.cn
经　　销：新华书店
印　　刷：上海新文印刷厂
开　　本：890×1240 毫米　1/32 开
印　　张：5
字　　数：128 千字
版　　次：2007 年 11 月第 1 版　2007 年 11 月第 1 次印刷

ISBN 978-7-80745-095-5/G·014　　定价：17.00 元

总 序

PREFACE

创办中国浦东、井冈山、延安干部学院，是党中央从推进中国特色社会主义伟大事业和党的建设新的伟大工程全局出发作出的一项重大决策。

中国浦东干部学院自2005年3月正式开办以来，始终坚持胡锦涛同志提出的切实把学院建设成为进行革命传统教育和基本国情教育的基地、提高领导干部素质和本领的熔炉以及开展国际培训交流合作的窗口的办学目标，秉承“实事求是、与时俱进、艰苦奋斗、执政为民”的办学要求，努力体现“国际性、时代性、开放性”的办学特色，取得了较为突出的业绩，正在向“国内一流、国际知名”的新型领导学院的方向稳步迈进。

办好一所学院的关键是教师。人才是事业之本，人才兴则事业兴、事业旺。“所谓大学者，非大楼之谓也，有大师之谓也”，这句话对中国浦东干部学院来讲同样有指导意义。中国浦东干部学院汇集了一批优秀的教师，他们当中，既有国外学成归来的学子，也有来自国内著名

高校、科研机构的青年才俊。他们有火热的创业激情，有对干部教育培训事业的执著和热爱。他们大多拥有博士学位，在自己所属的学科领域已崭露头角。这支队伍是建设好中国浦东干部学院的人才支持和智力保证。为他们搭建平台，促进他们成长，引领他们发展，是学院义不容辞的职责。

支撑一所学院的基础是学术。学院之称，有学科、学养、学理之意蕴。没有了学术，学院也就失却了原动力和根基。中国浦东干部学院创办以来，坚决贯彻并创造性地执行中央的战略决策和一系列办学要求，明确了教育培训、科学研究、咨询服务、领导测评、网络教育五位一体的功能定位，突出强调了学术研究、学科建设在学院发展中的重要地位。整合学术资源、加强学科建设对学院发展至关重要。

体现一所学院办学水平的重要标志是品牌。品牌汇集了办学的智慧，凝聚了办学的精华，同时也提升了学院的美誉度。中国浦东干部学院以问题为核心，以能力为导向，以现场教学资源为依托，自创办以来培训了大批学员，培训成果显著，一批具有较高质量的课程品牌、教学品牌正在形成。从一所学院的发展来看，既要有教学的品牌，也要有科研的品牌；既要有品牌学员，也要有品牌教员；而这一切都需要长期积累。求木之长必先固其本。积累品牌素材，探寻品牌来源，滋养品牌发展，是学院发展的长远大计。

基于上述认识，我们组织出版了这套中国浦东干部学院博士文库。入选文库的书稿均为学院青年教研人员的博士学位论文，并经过了严格的“双盲”评审。作者根据评审意见和所论问题的发展以及研究的深化，都进行了认真修改，可以说基本反映了所论问题的学科前沿。我们希望，这套分辑出版的文库能开启和激励我们的后续研究，促进学院自身研究特色和学术传统的形成，促进相关学科领域的建设，促进学术交流与繁荣。

文库的出版得到了上海社会科学院出版社领导和编辑同志的鼎力支持和帮助，借此表示诚挚谢意。同时，对为文库的建设作出贡献的评审专家和付出辛劳的同志表示诚挚的感谢。

文库中存在的不足，敬恳广大读者批评指正。

中国浦东干部学院博士文库编委会

2007 年元月

ABSTRACT

Peer-to-peer computing (*Abbr.* P2P) has become a hot spot in the computer research and industrial fields. In a P2P-based system, each peer (node) has equal functionalities and responsibilities, i. e. , a peer can both act as a server of others providing them with data/services, and serve as a client of others consuming their data/services; Furthermore, the interaction among peers can be direct and symmetric; In addition, peers can join in or leave from the system at any time to form dynamic network environment. This sort of systems enjoy many potential merits, including scalability, reliability, high information availability, and efficiently utilizing systems' resources, so that P2P is regarded as a promising technology to re-architecture the future Internet-based applications.

While file sharing has become the most popular research and development topic in the P2P computing field, existing systems can just support semantics-free sharing of large granularity and inefficiently utilize their own resources. To address these challenges, with taking the text files as sharing objects, the concept of peer-to-peer based information retrieval has been

proposed. Furthermore, to address the challenges related to its key techniques, an extensive study to this new type of systems has been conducted and following contributions have been achieved:

(i) A three-layer based architecture has been proposed, which is made up of structured layer, unstructured layer and application layer. This new type of architecture has inherited the advantages of both structured and unstructured architectures while eliminated their respective disadvantages as well, so that the potential merits of peer-to-peer computing can be exploited to efficiently support information retrieval in the dynamic P2P environment.

(ii) Based on the comprehensive analysis on the existing resource location and query routing schemes, a peer-clustering based resource location mechanism and a self-adaptive routing strategy have been devised, which can efficiently utilize systems' resources and effectively satisfy users' demands.

(iii) With the systematically studing on the challenges related to ranking and merging the answers retrieved from different peers in a P2P-based information system (e. g. PeerIS), a deep insight to its underlying diathesis has been obtained. Furthermore, a fully distributed ranking and merging strategy has been designed, while relevant issues have been addressed.

(iv) An analysis on retrieving optimization and peer dynamics for the P2P-based information retrieval systems has been conducted. Specifically, a high level cost model for this sort of

systems has been proposed and an agent-based strategy has been divised to obtain the coefficients of the model. Furthermore, with taking the dynamics of peers as a factor of P2P systems' cost, a fuzzy cost analysis on peers' dynamics has been carried out. Specifically, by employing the fuzzy theory, the peers' behavior and identify their reliability has been depicted and captured. Consequently, the goal of retrieving optimization has been defined, i. e., conducting peers must satisfy the demand of reliability during the retrieving processing, while the response time must be shortest and the systems' resources must be efficiently consumed.

(v) Based on the key techniques described above, PeerIS: a prototype of P2P-based information retrieval system, has been implemented.

In a word, this book has detailed the design, key techniques and experimental results of a P2P-based information retrieval system, which have been realized in PeerIS. All these contributions are achieved through a comprehensive study on the related theories, existing technologies and experimental results. Moreover, the experimental results have verified such a P2P-based information retrieval system can realize fully semantic information retrieval and sharing of fine granularity, and efficiently utilize the system's resource as well.

Keywords: peer-to-peer computing, information retrieval, resource locating and query routing, retrieved result ranking, retrieval optimization.

前　言

PREFACE

对等计算(Peer-to-Peer computing,简称P2P),自2000年中期以来迅速成为计算机研究界和工业界关注的热点。在对等计算系统(简称P2P系统)中,每个节点都拥有对等的功能与责任,即每个节点既可以充当服务器向其他节点提供数据或服务,又可以作为客户机享用其他节点提供的数据或服务;节点之间的交互直接对等;任何节点可随时自由地加入或离开该系统,形成一个真正动态的网络环境。这类系统具有许多潜在优势,如系统的可扩展性、鲁棒性、信息可用性、系统资源利用率高以及能够满足某些特殊应用需求,因而P2P被认为是未来重构基于Internet应用的前沿技术之一。

虽然当前基于对等计算的研发几乎集中于文件共享应用,但是现有的P2P文件共享系统大多存在仅支持粗粒度(文件水平)、弱语义(甚至缺乏语义)的共享以及系统的效率低等局限性。为了应对这些挑战,我们提出了基于对等计算的信息检索,深入研究了这类系统的关键技术所面临的挑战,并取得了以下成果:

(1)提出了三层构架的体系结构,由下至上分别为:结构化层、非结构化层和应用层。这种新型的体系结构集成了当前流行的结构化和非结构化两种不同的体系结构,既充分发挥了它们的优点,又消除了它们的不足,能够更充分地发掘对等计算的潜在优

势，因而能有效地支持在动态的对等计算环境中进行的信息检索。

（2）在综合评析当前对等计算系统所采用的资源定位和查询路由策略的基础上，提出了基于节点聚类的资源定位机制和自适应查询路由策略，使基于对等计算的信息检索系统不但能够高效地利用系统资源，而且能够有效地满足用户需求。

（3）系统地研究了当前基于对等计算的信息检索系统在检索结果排序和合并方面所面临的挑战，提出了一种全新的分布式检索结果排序和合并策略，并解决了与之相关的问题。

（4）进行了检索优化和节点动态分析。提出了一种与对等计算系统特性相适应的代价分析模型和获得模型中各个代价因子系数的办法；把节点动态性置于该模型之中，应用模糊集理论刻画和捕捉节点的行为模式，进行了节点模糊可靠性分析，以确定节点的可靠性。把检索优化的目标扩展为：保证检索处理执行时间最短和系统资源消耗最少；同时保证执行节点在整个检索处理过程中具有最高的可靠性。

（5）基于上述关键技术，研发了基于对等计算的信息检索原型系统：PeerIS。

总之，本书详细论述了实现基于对等计算的信息检索系统的关键技术和测试结果。本书的工作是建立在对相关理论和已有技术的详尽分析以及大量的实验测试结果之上的。实验结果表明，我们提出的基于对等计算的信息检索技术不但能够支持语义丰富的信息检索与共享，而且能高效地利用系统资源并有效地满足用户需求。

目 录

CONTENTS

CONTENTS