

摘 要

论文的摘要是对论文研究内容和成果的高度概括。摘要应对论文所研究的问题及其研究目的进行描述，对研究方法和过程进行简单介绍，对研究成果和所得结论进行概括。

摘要应具有独立性和自明性，其内容应包含与论文全文同等量的主要信息。使读者即使不阅读全文，通过摘要就能了解论文的总体内容和主要成果。

论文摘要的书写应力求精确、简明。切忌写成对论文书写内容进行提要的形式，尤其要避免“第 1 章……；第 2 章……；……”这种或类似的陈述方式。

其他要求：

(1) 文字：中文和英文各一份；中、英文摘要及其关键词分别各置一页内；字号小四宋体；

(2) 字数：中文 600 字左右；英文 3500 字符左右；

(3) 字号字体：小四号宋体。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语，能体现论文的主要内容，词组符合学术规范。

关键词 3-5 个，不超过 5 个，每个关键词中间用分号隔开。

关键词：关键词 1；关键词 2；关键词 3；关键词 4；关键词 5

Abstract

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words “the first chapter”, “the second chapter” and the like should be avoided in the abstract.

Other requirements:

(1) Text: one copy in Chinese and one copy in English; the Chinese and English abstracts and their key words are placed on one page respectively; the font size is small four Song typeface;

(2) Number of words: about 600 in Chinese and 3500 in English;

(3) Font size: Times New Roman, small four.

Keywords are terms used in a dissertation for indexing, reflecting core information of the dissertation. It can reflect the main content of the paper, and the phrases conform to the academic norms.

An abstract may contain 3-5 key words, no more than 5, with semi-colons used in between to separate one another.

Keywords: keyword 1; keyword 2; keyword 3; keyword 4; keyword 5

目 录

Abstract	I
目 录	III
第 1 章 引言	1
1.1 选题背景和研究意义	1
1.2 研究内容	1
1.3 论文的组织结构	1
第 2 章 文献综述	3
2.1 本章引言	3
第 3 章 对等网络中宽松约束的一般性搜索的理论模型	5
3.1 本章引论	5
3.2 模型基本假设	6
3.2.1 无偏向性搜索	7
3.2.2 结点特性	8
3.2.3 短时稳态性	8
3.2.4 模型假设的总体叙述	8
3.2.5 宽松约束的一般性搜索性能理论模型	8
参考文献	9
致谢	11
在学期间完成的相关学术成果	13

Contents

Abstract-English	I
Contents	III
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research Contents	1
1.3 Arrangement of the Thesis	1
Chapter 2 A Review of the Literature	3
2.1 Introduction	3
Chapter 3 Theoretical Model of General Search with Loose Con-	
straints in Peer-to-Peer Networks	
5	
3.1 Introduction	5
3.2 Basic Assumptions of the Model	6
3.2.1 Unbiased Search	7
3.2.2 Node Characteristics	8
3.2.3 Short-Term Stability	8
3.2.4 General Description of Model Assumptions	8
3.2.5 Theoretical Model of General Search Performance with Loose Constraints	
8	
References	9
Acknowledgements	11
Relevant Academic Achievements Completed During the Acad-	
emic Period	13

第 1 章 引言

1.1 选题背景和研究意义

1.2 研究内容

1.3 论文的组织结构

第 2 章 文献综述

2.1 本章引言

第3章 对等网络中宽松约束的一般性搜索的理论模型

3.1 本章引论

本章为 P2P 中宽松约束的一般性搜索建立理论模型，以研究此类搜索的效率和带宽开销。根据本章的理论模型可以很好地测算出各种条件下及不同应用中的 P2P 搜索效率和带宽开销，为 P2P 中宽松约束搜索的研究建立了基础。通过模型求解可以得到搜索所需的瓶颈资源（即结点带宽）的理论下限，并可算出不同系统参数下最优的搜索性能以及达到此性能时的最优数据索引分布，从而为 P2P 系统搜索算法的设计、性能优化、性能比较以及可行性分析提供了一般性方法。后面第四章提出的近似最优的实用搜索算法就是直接应用本章模型和结论而设计的。

①

从第2章讨论可知，宽松约束搜索的用途非常广泛，是广域网上多服务器系统的基础功能和服务。为了明确起见，这里重述一下定义：所谓“宽松约束”是指搜索不要求返回全部符合条件的结果，而只要返回一个或若干个（根据用户要求而定）即可；所谓“一般性搜索”是指搜索算法必须是普适的，具体来说就是搜索条件任意和数据存放位置任意。目前尚不能很好地解决此类搜索问题，主要面临的问题是算法网络通信量过大，很容易超过结点和网络的承受能力，造成严重的带宽开销和系统不可扩展（non-scalable）的问题。现有算法中，结构化 P2P 中精确匹配数据标识的数据定位算法无法支持各种非精确匹配的搜索（如子串匹配的搜索）。Gnutella 等非结构化 P2P 系统依靠消息转发的随机搜索方法虽然符合一般性搜索的要求，但是存在严重的性能问题。由于 P2P 巨大的结点数和数据量，不加优化的随机搜索面临“大海捞针”的困境，搜索消息通常要游历很多无关结点并产生大量冗余消息之后才能找到数据。因此，不论使用消息洪泛 [2] 或是性能稍好的随机走步 [45]，普遍认为 Gnutella 式的随机搜索算法不具可扩展性 [47]，当结点较多时带宽约束将造成非常严重的系统瓶颈。带宽约束正是制约此类 P2P 搜索的最大问题。

① 脚注处序号“□，……，□”的字体是“正文”，不是“上标”，序号与脚注内容文字之间空半个汉字符，脚注的段落格式为：单倍行距，段前空 0 磅，段后空 0 磅，悬挂缩进 1.5 字符；字号为小五号字，汉字用宋体，外文用 Times New Roman 体。

为了解决通信量过大的问题，人们尝试了很多优化措施来改进基础设施和搜索算法的各个方面，包括使用冗余更少和通信方面更“温和”的消息转发算法、采用混合结构、使用超级结点的偏向性搜索、扩散数据索引和缓存搜索结果等等 [42-54,106]。其中最为重要的手段就是主动扩散数据索引，即结点不只存储自身数据，同时缓存其他结点所存数据的索引（参见 2.3.1 节）。由于数据的分布是任意的，搜索不可避免地带有盲目性，平均搜索跳数总是依赖于“知道”目标数据的结点个数而难于进一步提高（详见 3.2 节的分析）。因此在一般性搜索问题中，通过扩散索引来增加数据知名度的方法成为本质上的解决途径。然而索引扩散并不总是有效率的，它也会带来带宽开销。一方面，扩散更多的索引可以使搜索更快地返回，减少了搜索带宽开销；另一方面，由于 P2P 中结点和数据处于不断动态变化之中，当数据失效或更新时（如结点离线、删除或更新数据），数据的索引也相应失效，必须加以更新维护。因此，扩散更多的索引意味着维护开销的增加。于是在带宽开销方面，搜索开销与索引维护开销之间存在着折衷关系（trade-off）。与以往工作中仅考虑搜索开销不同，本章的模型中我们同时考虑搜索和维护两方面，给出了索引扩散方法对搜索整体性能的影响和数学关系。通过模型我们发现索引数量是决定宽松约束一般性搜索性能的至关重要的因素，采用最优索引分布可以很大程度上提高性能，降低系统开销。与一般认为的 P2P 无偏向性搜索难于扩展（non-scalable）恰恰相反，模型显示在最优的索引扩散策略下，基于无偏向性搜索具备很好的可扩展性，其结点负载和带宽开销随系统规模 N （结点数）增长具有 $O(\sqrt{N})$ 的增长关系。这种平方根关系保证了对大规模 P2P 系统很好的适应性。本章剩余部分按照如下方式组织：3.2 节给出模型假设；3.3 节推导出带宽开销和搜索效率的计算公式，给出性能模型；3.4 节通过优化索引分布得到了理论最优的宽松约束搜索，证明了对可扩展性具有重要意义的“平方根关系”；3.5 节对模型结论和意义进行了总结；3.6 节讨论模型适应性并和相关工作进行了比较，最后是本章小结。

3.2 模型基本假设

一般性搜索要解决任意可能的数据存放方式和任意的查询条件下的搜索问题。任意可能的数据存放意味着数据与存放结点之间可以不存在任何相关性，因而无法利用类似兴趣偏好、结点特性或数据存放的规则来指导消息转发或优化性能，单步搜索效率等同于盲目搜索（blind search）。任意的查询条件，意味着只有获得了

完整的元数据（数据中用来被查询的部分）才可以判断出该数据是否符合查询条件，单靠部分元数据不能解决所有可能的查询请求。这样以往针对特定搜索条件的倒排式索引（如关键词倒排表等）无法发挥作用，只能使用正排式的数据索引。这样，一般性搜索限制了可能采取的优化措施，使搜索性能存在理论极限。虽然在特定应用中使用的不一定是严格的“一般性搜索”，其数据与结点之间可能存在一定的相关性，并且可以借助搜索条件的特性来优化性能，但是本章只讨论最一般的情况，目的在于建立基础性的模型，研究和寻找适用于绝大多数应用的宽松约束搜索算法。针对特定搜索条件的高效搜索算法可参见第五章，利用数据和结点的语义相关性来优化性能的研究参见第六章。本节针对 P2P 中一般性搜索的特点给出模型的基本假设，具体包括搜索的“无偏向性”、结点特性以及系统的短时稳态假设。我们首先讨论这三方面问题，最后给出模型假设的总体叙述。

3.2.1 无偏向性搜索

如果 P2P 中所有结点都以相同或相近的概率接收到搜索请求，那么称此搜索算法为“无偏向性搜索（unbiased search）”。无偏向搜索可看作是对等网络搜索的基础，也是应用最多的一种。这是由于对等网络中单结点能力较弱，只有让所有结点均摊巨大的总体搜索负载，才可能支撑起高强度的搜索算法。如果搜索消息不能近似均匀地散落在所有结点上，那么负载重的结点很容易发生过载。均衡负载的思想正是对等网络存在和发展的基础，即集合众多微小的力量（大量弱结点）形成巨大的总体服务能力。实际应用中，只要搜索算法在选择邻居结点转发消息时没有特殊的偏向性，那么就可以近似认为是无偏向性搜索，譬如最常用的洪泛和随机走步就是无偏向性算法。模型中我们假设搜索算法是无偏向性的。对于使用了超级结点的偏向性搜索，通常可认为是将结点划分为不同类别。同一类别的结点运行相同的算法和协议，因此各类别内部仍然是无偏向性搜索，符合模型假设。此时可针对每个类别的结点群体分别使用模型，从而得出整体算法的性能（详见 3.5 节）。由于数据分布是随机的，且结点收到搜索消息的概率彼此相同，因此无偏向性条件下单步搜索的成功概率不会优于盲目搜索的性能。虽然搜索途中遇到的结点可带来历史信息并指导后续遍历，但由于巨大的结点数和数据量以及数据随机存放的特点，小量历史信息的作用非常有限，且与特定应用的性质相关。因此模型中对此不作考虑，仍然以盲目搜索作为无偏向性搜索单步性能。

3.2.2 结点特性

P2P 中结点总处于不断地动态变化之中,随时有结点加入和离开系统。动态性显著地影响系统性能和索引的有效性。按照 P2P 中通常的假设,结点的行为(在线或离线)独立于其他结点,并且所有结点具有统一的动态特性以及在线时间分布,这种分布一般用指数分布来刻画。具体来说,设结点的平均在线服务时间(session time)为 T_{session} ,则单个结点在线时间长度符合参数 $\lambda=1/T_{\text{session}}$ 的指数分布。由于结点动态变化彼此独立,因此一段时间之内发生的动态变化次数符合泊松分布,其参数与指数分布参数一致。

3.2.3 短时稳态性

我们假设 P2P 系统具有短时稳态特性,即一段不长的时间中系统的性质(如结点平均在线时间 T_{session} 、数据总量、数据访问频度分布)以及系统规模(结点数 N)不会发生显著的变化。尽管长时程中系统性质可能会明显变化,但是这种变化总是靠缓慢的积累而产生的。因此,可以用一个稳态模型来描述 P2P 系统,而 P2P 系统的长程变化可以用同一个稳态模型的不同参数取值点来刻画。

3.2.4 模型假设的总体叙述

综合以上三方面就得到本章理论模型的基本假设。具体而言,模型假设 P2P 系统和一般性搜索算法具有如下特点:一般性搜索算法对结点不具有偏向性,所有结点以相同或相近的概率收到搜索请求;结点在线时间可用独立同分布的指数分布来描述;P2P 系统在相对较短的时间段内可看作是稳态,系统主要参数保持不变。以上假设是针对 P2P 系统及一般性搜索的特点而做出的,具有广泛的适应性,可用来研究目前大多数 P2P 宽松约束搜索系统的性能。关于模型适应性将在 3.6.1 节做进一步讨论。

3.2.5 宽松约束的一般性搜索性能理论模型

本节建立无偏向性搜索的理论模型,模型统一解决无偏向性搜索的带宽开销计算、带宽开销的理论下限、最优的搜索性能以及最优索引分布。模型中用到的符号及说明参见表 3.1。

参考文献

致 谢

衷心感谢导师 ××× 教授和物理系 ×× 副教授对本人的精心指导。他们的言传身教将使我终生受益。

在美国麻省理工学院化学系进行九个月的合作研究期间，承蒙 Robert Field 教授热心指导与帮助，不胜感激。

感谢 ××××× 实验室主任 ××× 教授，以及实验室全体老师和同窗们学的热情帮助和支持！

本课题承蒙国家自然科学基金资助，特此致谢。

在学期间完成的相关学术成果

学术论文:

- [1] Yang Y, Ren T L, Zhang L T, et al. Miniature microphone with silicon- based ferroelectric thin films[J]. Integrated Ferroelectrics, 2003, 52:229-235.
- [2] 杨轶, 张宁欣, 任天令, 等. 硅基铁电微声学器件中薄膜残余应力的研究 [J]. 中国机械工程, 2005, 16(14):1289-1291.
- [3] 杨轶, 张宁欣, 任天令, 等. 集成铁电器件中的关键工艺研究 [J]. 仪器仪表学报, 2003, 24(S4):192-193.
- [4] Yang Y, Ren T L, Zhu Y P, et al. PMUTs for handwriting recognition. In press[J]. (已被 Integrated Ferroelectrics 录用)

专利:

- [5] 任天令, 杨轶, 朱一平, 等. 硅基铁电微声学传感器畴极化区域控制和电极连接的方法: 中国, CN1602118A[P]. 2005-03-30.
- [6] Ren T L, Yang Y, Zhu Y P, et al. Piezoelectric micro acoustic sensor based on ferroelectric materials: USA, No.11/215, 102[P]. (美国发明专利申请号.)