

基于节点兴趣非结构化 P2P 网络搜索机制研究

庄 伟

(南京邮电大学 自动化学院 江苏 南京 210023)

摘 要: 随着网络用户以及网络资源的增长, P2P 网络, 一种在对等者 (peer) 之间分配任务和工作负载的分布式应用架构, 引起了广泛关注。由于具有较好的可用性、可扩展性, 非结构化 P2P 网络成为 P2P 网络研究的热点。现今对于非结构化 P2P 网络资源搜索算法的研究主要是在洪泛算法的基础上进行一定的改进, 但这些算法存在明显的问题: 一是算法在搜索过程中有一定的盲目性, 搜索效率不高; 二是搜索过程中会向所有邻居节点发送消息从而产生大量的冗余消息, 易造成网络阻塞。针对上述问题, 提出一种基于节点兴趣的非结构化 P2P 覆盖网络拓扑结构, 在相似度较高的节点之间建立二叉搜索树来降低查询消息转发的盲目性并提高搜索效率。仿真结果表明, 与传统的洪泛算法相比, 提出的非结构化 P2P 网络搜索机制查询时间更短, 查询消息量更少, 搜索效率更高。

关键词: P2P 网络; 拓扑结构; 兴趣相似度; 二叉搜索树

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2018)05-0068-05

doi: 10.3969/j.issn.1673-629X.2018.05.016

Research on Search Mechanism of Unstructured P2P Network Based on Node of Interest

ZHUANG Wei

(School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Peer-to-Peer (P2P) network, a distributed application architecture that divides tasks and workloads between peers, has recently attracted lots of concern since the growth of network users and network resources. Unstructured P2P networks have better usability and scalability, which makes it become the focus of P2P network research. Unstructured P2P network resource search algorithm is mainly based on improved flooding algorithm, but these algorithms exist some obvious problems. One is the central idea of the blind search algorithm, the search efficiency is not high; the two is the search process will produce a message sent to a large number of redundant messages from neighbor nodes, easy to cause network congestion. For these issues, we raise a binary searching tree based on interest unstructured P2P overlay network topology. A logical connection between similar nodes is used to reduce the blindness of message forwarding and to improve the search efficiency. The simulation shows that compared with the traditional flooding algorithm, the proposed unstructured P2P network search mechanism has shorter query time, less query message and higher search efficiency.

Key words: P2P network; topology structure; interest similarity; binary searching tree

0 引 言

在互联网技术发展的初期阶段, 计算机网络模型主要是 Client/Server (客户端/服务器) 模式, 在这种模式中, 由于计算机计算存储性能均较低, 节点工作主要依赖于中央服务器。中央服务器是整个网络的核心部分, 是网络中资源或者服务的提供方, 这种模式对中央服务器的要求很高并且使中央服务器的负载较大, 一般只适用于中小规模的网络。随着互联网技术、计算机存储能力的快速发展和网络资源的急剧增加, 中小

规模的网络已经不能满足用户的需求, 用户对于网络技术的要求也越来越高, 在此基础上, 一种新型的 P2P 对等网络应运而生^[1]。P2P 网络淡化了中央服务器的概念, P2P 网络中各个节点的地位和作用都是相同的, 且每个节点既能向其他节点进行资源或者服务的请求, 又可以享受其他节点提供的服务。

目前 P2P 网络的拓扑结构可根据耦合程度分为结构化 P2P 网络^[2]和非结构化 P2P 网络^[3]。结构化 P2P 网络主要有 Tapesstry^[4]、Chord^[5]等, 网络中所有

收稿日期: 2017-04-21

修回日期: 2017-08-23

网络出版时间: 2018-02-08

基金项目: 国家自然科学基金 (61374180)

作者简介: 庄 伟 (1992-), 男, 硕士研究生, 研究方向为复杂网络系统。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20180207.1809.038.html>

节点按照某种结构进行有序组织形成一种结构规则的网络。每个节点只存储特定的信息或者特定信息的索引,当用户需要在结构化网络中搜索信息时,必须知道这些信息可能存在于哪些节点中。但结构化 P2P 网络维护成本较大,并且节点之间的物理结构比较容易破坏,所以结构化 P2P 网络不适用于大规模网络部署^[6]。非结构化 P2P 网络主要有 Gnutella^[7]等,在非结构化 P2P 网络中,每个节点只存储自身的信息。当用户需要从非结构化 P2P 网络中搜索信息时,用户预先并不知道该信息存储在哪些节点上。非结构化 P2P 网络采用的都是基于洪泛算法的消息传递机制,洪泛算法搜索成功率不高并且在消息传递过程中会产生大量的冗余消息,因此非结构化 P2P 网络中资源的定位和搜索性能的提高是关键问题。

针对非结构化 P2P 网络的搜索问题,文献[8]提出了洪泛算法。当一个成员节点请求资源时,会向所有邻居节点发送请求查询包,并给出查询深度 TTL,邻居节点收到查询包后,首先搜索本节点的资源,若未找到查询请求需要的资源则转发给自己的所有邻居节点,以此类推,直到找到资源或者 TTL 为 0 为止。文献[9]利用无标度网络度分布的幂率特性,在搜索过程中将查询请求传递给度值最大的邻居节点。文献[10]采用只传递给一定比例的相邻节点的策略,在彼此相似度较高节点之间建立关联,节点之间相似度高则更有可能包含查询请求需要的资源。文献[11]提出将节点中内容的受欢迎程度与节点的度值相结合作为确定资源搜索路径的搜索算法。文献[12]提出通过完全二叉树重新组建网络拓扑结构并利用二叉树的特性提高资源搜索的效率。

研究表明,非结构化 P2P 网络中每个节点都会表现出一定的兴趣倾向,而当两个节点之间的兴趣倾向相近时,这两个节点所包含的资源之间也会相应地表示出一种相关性^[13-14]。文中考虑了非结构化 P2P 网络的拓扑结构与网络中节点之间的兴趣倾向,引入兴趣二叉搜索树的概念,通过节点之间的兴趣相似度构建多个搜索二叉树,将兴趣倾向相近的节点组建成一棵搜索二叉树。当节点在资源搜索时可以将查询请求转发给与自己直接相连的节点,这样有利于减少网络搜索过程中的冗余消息,同时有利于提高资源搜索的成功率。

1 基于兴趣的二叉搜索树

1.1 节点兴趣相似度的计算

1.1.1 节点的兴趣域

在非结构化 P2P 中,每个节点存储的资源信息都会表现出对某个领域的兴趣倾向,文中采用经典的向

量空间模型(vector space model, VSM)^[15]表示节点的兴趣域。向量空间模型利用关键词在资源中出现的次数构成向量来表示节点的兴趣域模型,是信息获取领域中的一种经典算法。采用特征向量的形式表示节点的兴趣域,可以将节点之间相似度的比较转化为特征向量之间相似度的比较。

节点兴趣域表示方法:选取节点资源中出现次数最多的 k 个关键词,关键词 k_i 在节点中的权值取为关键词 k_i 在节点所包含资源中出现的次数。

1.1.2 节点兴趣相似度

节点中的数据用 VSM 模型表示出来后,会通过相似度函数来计算两个向量之间的相似程度。文中使用皮尔逊相关系数^[16]来衡量两个向量之间的相似程度。对于节点 X 和节点 Y ,节点的兴趣域通过 VSM 模型分别表示为 $X = \{x_{i1}, x_{i2}, \dots, x_{im}\}$, $Y = \{y_{j1}, y_{j2}, \dots, y_{jm}\}$,其节点之间的相似度通过皮尔逊相关系数表示为:

$$\text{Sim}_{x,y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (1)$$

其中, N 为 X 与 Y 中相同元素的个数。将 $\text{Sim}_{x,y}$ 取绝对值,即可将兴趣相似度范围限制为 $[0, 1]$ 。相似度 $\text{Sim}_{x,y}$ 计算出的数值越高,表示节点之间的兴趣相似度越大,反之越小。当节点 X 与节点 Y 之间的资源完全相同时, $\text{Sim}_{x,y}$ 的值为 1。

1.2 二叉搜索树

二叉搜索树是指或者为空树,或者具有如下性质的二叉树:如果它的左子树不为空,则左子树上的所有节点的值都小于其根节点的值;如果它的右子树不为空,则右子树上的所有节点的值都大于其根节点的值,并且它的左右子树也分别为二叉搜索树;不允许存在重复的节点。

1.3 兴趣二叉搜索树的构建

根据二叉搜索树的结构特征,在构建兴趣二叉搜索树非结构化 P2P 覆盖网络拓扑结构时,首先要确定二叉搜索树的根节点。在确定根节点之后,其他节点依据与根节点的兴趣相似度的值在兴趣二叉搜索树中确定自己的位置。在构建过程中,为了保证二叉搜索树的性质,兴趣二叉搜索树应满足如下性质:左孩子与根节点的相似度应该小于根节点中设定的相似度阈值;右孩子与根节点的相似度应该大于根节点中设定的相似度阈值。在兴趣二叉搜索树中允许存在重复节点,文中右孩子与根节点的相似度可以相等。

兴趣搜索二叉树的构建步骤如下:

(1) 确定二叉搜索树的根节点。在建立基于兴趣二叉搜索树非结构化 P2P 覆盖网络时,从根节点开始构建。根节点的选取应该综合考虑节点的负载能力、节点的计算能力、稳定性和带宽等多方面因素。为了模拟仿真,将网络中的节点按照兴趣相似度等分为 N 个兴趣块,并在每个兴趣块中创建二叉搜索树,其中 N 是所创建的二叉搜索树的个数,即为根节点的数目。

(2) 根节点向网络中的普通节点发送广播,广播中需要包含根节点自身的兴趣域 R ,在普通节点接收到根节点的广播后会计算自身与根节点之间的兴趣相似度 $Similarity$,并且将 $Similarity$ 按照其大小进行排序,选择计算获得的 $Similarity$ 值最大的两个节点向该根节点发送请求加入的信息,该信息中包含计算获得的 $Similarity$ 值。

(3) 根节点接收到两个普通节点的请求加入的信息后,会分别将该信息中包含的 $Similarity$ 值与根节点本身设定的阈值相比较,如果该 $Similarity$ 值大于设定的阈值,则将该普通节点作为根节点的右孩子节点,否则作为根节点的左孩子节点;如果 $Similarity$ 的值等于设定的阈值,则将该普通节点作为根节点的右孩子节点。

(4) 根节点的左右孩子节点分别再发送广播,重复第二步,直到所有节点都被分配到相应的二叉搜索树中。

在构造兴趣二叉搜索树的过程中,节点被分为两类,一类是根节点,另一类是普通节点。对于两类不同的节点,节点中存储的信息也不相同。

根节点作为一个搜索二叉树中的超级节点,应该维护并存储整棵搜索二叉树的有关信息,所以根节点应该存储的信息如下:

(1) 子节点 IDS: 记录该根节点的所有子节点的 ID。

(2) 根节点 IDS: 记录网络中其他根节点的 ID。

(3) 本地资源: 记录本节点可以被搜索到的资源。

(4) 兴趣域: 记录整棵树的兴趣域。当查询包需要的资源不在根节点的兴趣域但是在该孩子节点中搜索到的时候,则将查询包加入根节点的兴趣域中。在初始时,兴趣域仅存储根节点的本地资源,随着搜索的进行会动态更新兴趣域。

普通节点应该存储的信息如下:

(1) 本地资源: 记录本节点可以被搜索到的资源。

(2) 兴趣相似度: 记录本节点与根节点之间的兴趣相似度的值。

2 搜索算法

当基于兴趣二叉搜索树的非结构化 P2P 覆盖网

络拓扑结构创建完成后,资源在兴趣二叉搜索树上的搜索分为两个阶段: 第一阶段是在与查询请求相似度最高的兴趣二叉树中搜索; 第二阶段在兴趣二叉搜索树中没有搜到所需要的资源后,将查询请求根据兴趣转发因子转发到其他兴趣二叉树中搜索。

2.1 兴趣二叉搜索树内部查询

在创建基于兴趣二叉搜索树的非结构化 P2P 覆盖网络拓扑结构后,节点资源的搜索首先会直接根据查询资源的兴趣倾向在相应的二叉搜索树中查询。在相应兴趣二叉搜索树中搜索资源的步骤如下:

(1) 查询请求包 Q 向所有的根节点发送广播搜索请求,选择与查询请求包 Q 兴趣相似度值最大的节点作为开始搜索的起始节点。

(2) 在根节点中查询搜索本地资源,查找成功则直接返回结果,否则转到第 3 步。

(3) 根据兴趣二叉搜索树的特点,选择根节点的右孩子节点进行消息转发,如果该节点的右孩子节点不存在则不进行转发。

(4) 如果接收到查询请求的所有节点均被访问过,则直接跳转到第 6 步,否则跳至第 5 步。

(5) 查询节点的本地资源,如果查找成功则返回结果,否则跳转到第 6 步。

(6) 判断生存时间 TTL 的值,如果 TTL 已经减小到 0,则表示搜索结束,否则跳转到第 3 步。

2.2 兴趣二叉搜索树之间查询

当在某个兴趣二叉搜索树中没有查询到所需要的资源时,需要考虑跳转到下一个兴趣二叉搜索树,对于下一个兴趣二叉搜索树的选取,文中综合考虑了查询反馈情况和两个根节点之间的兴趣相似度情况,提出了兴趣转发因子的概念。

2.2.1 查询反馈比

在非结构化 P2P 网络搜索过程中,考虑到以前历史的查询结果,返回查询反馈最多的节点更有可能包含搜索需要的资源。所以,选择查询反馈比最高的节点进行消息转发是一个相对较好的选择。

定义 1: 假设节点 $peer$ 的一个邻居节点为 $neighbour$,则将节点 $neighbour$ 的反馈包数与节点 $peer$ 的所有邻居节点总的反馈包数的比值称作查询反馈比:

$$Sim_{back} = QueryHits_{neighbour} / \sum QueryHits_{peers} \quad (2)$$

其中, $QueryHits_{neighbour}$ 表示邻居节点 $neighbour$ 的反馈包数; $\sum QueryHits_{peers}$ 表示所有邻居节点的反馈包数。

在非结构化 P2P 网络中,每个节点都保存一个 $PeerList^{[17]}$ 的列表结构,其中包含了节点的所有邻居节点对查询反馈的响应信息。该列表结构如图 1 所示。

Peer_ID	Total Query Feedback	QueryFeedback	Query Request
---------	----------------------	---------------	---------------

图1 PeerList 列表结构

其中, Peer_ID 是节点所有邻居节点的 ID; Total Query Feedback 是节点获得所有邻居节点的总反馈包数目; Query Feedback 是从 Peer_ID 返回的反馈包数目; Query Request 是节点发送到 Peer_ID 的请求包数。

PeerList 列表结构会随着搜索的进行不断地动态更新, 更新方法如下: 如果节点为被请求节点, 则查找该节点本地资源, 如找到需要资源, 则返回一个查询反馈包, 否则传递请求资源, 并更新该节点 PeerList 中 Query Request 的值。如果节点接受查询反馈包, 则更新节点 PeerList 中 Total Query Feedback 和对应邻居节点的 Query Feedback 的值。

2.2.2 兴趣转发因子

兴趣转发因子 Sim 是综合考虑查询反馈比 Sim_{back} 和节点兴趣相似度 Sim_{xy} 的一个参数。Sim 越高的节点更有可能包含搜索所需要的资源。

$$Sim = \alpha Sim_{back} + \beta Sim_{xy} \quad (3)$$

其中, $\alpha + \beta = 1$ 。

在节点搜索过程中, 下一个兴趣二叉搜索树选取的方法是计算所有根节点中兴趣转发因子并选择结果最大的根节点代表的兴趣二叉搜索树作为下一个需要进行资源搜索的兴趣二叉树。

通过使用节点的兴趣转发因子作为兴趣二叉搜索树跳转的依据, 兴趣二叉搜索树之间转发查询请求的步骤为:

(1) 如果在根节点 R 的兴趣二叉搜索树中未找到所需要的资源, 则计算与其他根节点之间的兴趣转发因子, 选择兴趣转发因子最大的节点进行查询请求的转发。

(2) 将计算获得的兴趣转发因子从大到小排序, 选择兴趣转发因子最大的根节点作为下一个需要进行搜索的兴趣二叉搜索树。

(3) 跳转到第 1 步。

3 仿真分析

为了评价算法的优劣, 实验采用 PeerSim1.0.5 仿真模拟器, 运用 Java 编程建立仿真环境。在此环境下, 对传统的搜索算法与文中提出的搜索算法进行分析比较。为了增加仿真的真实性, 非结构化 P2P 网络中的根节点和资源的分布都遵循 Zipf 分布。网络中共产生 10 000 个节点, 选取 100 个根节点, 查询的生成时间设置为 10, 通过以下两个衡量标准对文中算法与传统洪泛算法进行分析比较。

3.1 搜索成功率

非结构化 P2P 网络中资源的搜索成功率计算方法为:

$$\text{搜索成功率} = \frac{\text{成功搜索次数}}{\text{总搜索次数}} \times 100\% \quad (4)$$

实验结果如图 2 所示。

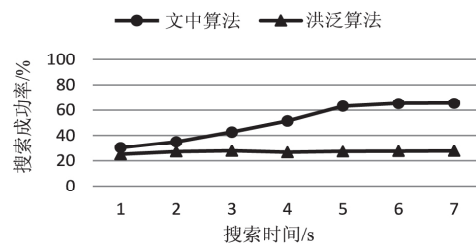


图2 搜索成功率对比曲线

从图中可以看出, 文中算法的搜索成功率总体高于传统的洪泛算法的搜索成功率, 并且随着搜索跳数的增加, 文中算法搜索成功率也随之增加。这是由于文中搜索算法在搜索过程中构建了兴趣二叉搜索树, 节点在选择下一跳节点时选择的是更可能包含搜索所需要资源的节点进行消息的转发, 并且在搜索过程中不断更新节点的兴趣域, 使得节点与查询包的相似度越来越大, 减少了不必要的跳转, 提高了搜索成功率。

3.2 平均路径长度

平均路径长度是指在多次搜索中的搜索路径的平均值, 可以用来衡量搜索过程中的时延问题。在搜索过程中, 平均路径长度越大, 表明搜索的路径越长, 搜索的响应时间也就越长, 搜索过程中的时延也就越大。所以在搜索过程中, 搜索到所需要资源的平均路径长度越小越好。

实验结果如图 3 所示。

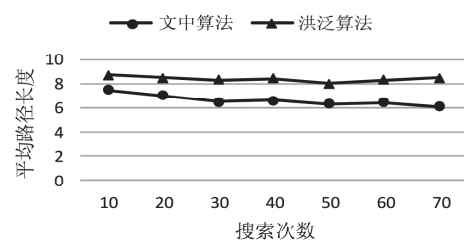


图3 平均路径长度对比曲线

从图中可以看出, 文中算法的平均路径长度会随着搜索次数的增加不断下降, 而传统的洪泛算法则基本不受影响。这是由于在传统的搜索算法中, 搜索次数的增加并不会引起节点内容的改变, 每次搜索都相当于一次新的搜索, 而在文中搜索算法中, 一方面将网络中兴趣倾向相近的节点构建成二叉搜索树, 使得节点资源在搜索时可以在更少的跳数之内找到资源, 另一方面如果在某兴趣二叉搜索树中找到资源, 则会将该资源添加到该兴趣二叉搜索树兴趣域中, 即文中算法可以充分利用历史搜索结果, 以后搜索遇到相同

资源时可以直接返回结果。

4 结束语

针对非结构化 P2P 网络中节点连接的随机性,提出了一种基于兴趣二叉搜索树的非结构化 P2P 覆盖网络拓扑结构,将内容相似度较大的节点构建成一棵二叉搜索树,利用二叉搜索树的特性,提高了搜索效率。通过 PeerSim 仿真,与传统的洪泛算法相比,该算法可以大大提高资源的搜索成功率,减少搜索的平均路径。

参考文献:

- [1] TEWARI S ,KLEINROCK L.Proportional replication in peer-to-peer networks [C]//International conference on computer communications. [s.l.]: IEEE, 2006: 1-12.
- [2] 王学龙,张璟.P2P 关键技术研究综述[J].计算机应用研究, 2010, 27(3): 801-805.
- [3] GAETA R ,GRANGETTO M.Identification of malicious nodes in peer-to-peer streaming: a belief propagation-based technique [J].IEEE Transactions on Parallel & Distributed Systems, 2013, 24(10): 1994-2003.
- [4] ZHAO B Y ,KUBIATOWICZ J D ,JOSEPH A D.Tapestry: an infrastructure for fault-tolerant wide-area location and routing[R].California: University of California at Berkeley, 2001.
- [5] 王挺,吴晓军,张玉梅.基于遗传算法的双向搜索 Chord 算法[J].计算机应用研究, 2016, 33(1): 46-49.
- [6] FURNESS J ,KOLBERG M. Considering complex search techniques in DHTs under churn [C]//Proceedings of the 2011 IEEE consumer communications and networking conference. Washington DC, USA: IEEE Computer Society, 2011: 559-564.
- [7] ADYA A ,BOLOSKY W J ,CASTRO M ,et al.Farsite: federated, available, and reliable storage for an incompletely trusted environment [J].ACM SIGOPS Operating Systems Review, 2002, 36(SI): 1-14.
- [8] FERREIRA R A ,RAMANATHAN M K ,AWAN A ,et al. Search with probabilistic guarantees in unstructured peer-to-peer networks [C]//Proceedings of the fifth IEEE international conference on peer-to-peer computing. Washington DC, USA: IEEE Computer Society, 2005: 165-172.
- [9] ADAMIC L A ,LUKOSE R M ,PUNIYANI A R ,et al. Search in power-law networks [J].Physical Review E, 2001, 64(4): 046135.
- [10] RAMANATHAN M K ,KALOGERAKI V ,PRUYNE J. Finding good peers in peer-to-peer networks [C]//Proceedings of the 16th international symposium on parallel and distributed processing. Washington DC, USA: IEEE Computer Society, 2011: 24.
- [11] TAKEDA D ,SUGAWARA S.A content searching scheme using popularity and link degree of nodes in unstructured P2P networks with cache routers [C]//International conference on complex, intelligent, and software intensive systems. Washington DC, USA: IEEE Computer Society, 2016: 590-594.
- [12] 何可,吴晓军,张玉梅.基于节点兴趣的非结构化 P2P 网络拓扑结构研究[J].计算机工程与应用, 2016, 52(9): 102-107.
- [13] 谭义红,陈治平,林亚平.基于兴趣挖掘的非结构化 P2P 搜索机制研究与实现[J].计算机应用, 2006, 26(5): 1164-1166.
- [14] HSIAO H C ,SU H W.On optimizing overlay topologies for search in unstructured peer-to-peer networks [J].IEEE Transactions on Parallel and Distributed Systems, 2012, 23(5): 924-935.
- [15] BUCKLEY C.Implementation of the smart information retrieval system [R].Ithaca, NY, USA: Cornell University, 1985.
- [16] STIGLER S M.Francis Galton's account of the invention of correlation [J].Statistical Science, 1989, 4(2): 73-79.
- [17] 刘璇.非结构化 P2P 网络基于马尔科夫链的资源搜索算法研究[D].北京: 北京交通大学, 2015.
- [18] Multidisciplinary Optimization, 2008, 37(4): 395-413.
- [12] 伊进田,刘云连,刘丽,等.一种高效的混合蝙蝠算法[J].计算机工程与应用, 2014, 50(7): 62-66.
- [13] 王丽,王晓凯.一种非线性改变惯性权重的粒子群算法[J].计算机工程与应用, 2007, 43(4): 47-48.
- [14] 盛孟龙,贺兴时,王慧敏.一种改进的自适应变异蝙蝠算法[J].计算机技术与发展, 2014, 24(10): 131-134.
- [15] LAAMARI M A ,KAMEL N.A hybrid bat based feature selection approach for intrusion detection [M]//Bio-inspired computing: theories and applications. Berlin: Springer, 2014: 230-238.

(上接第 67 页)

optimization algorithm: theoretical foundations, analysis, and applications [M]//Foundations of computational intelligence. [s.l.]: [s.n.], 2009: 23-55.

- [10] PAN T S ,KIEN D ,NGUYEN T T ,et al.Hybrid particle swarm optimization with bat algorithm [C]//Proceeding of the eighth international conference on genetic and evolutionary computing. Nanchang, China [s.n.], 2015: 37-47.
- [11] WANG Yong ,CAI Zixing ,ZHOU Yuren ,et al.Constrained optimization based on hybrid evolutionary algorithm and adaptive constraint-handling technique [J]. Structural and