Hans Journal of Data Mining 数据挖掘, 2011, 1, 17-20
http://dx.doi.org/10.12677/hjdm.2011.12005 Published Online October 2011 (http://www.hanspub.org/journal/hjdm/)

Information Retrieval Strategy on Unstructured Peer-to-Peer Network

Xi Liu
Huazhong University of Science & Technology, Wuhan
Email: llxfrankcs@gmail.com
Received: Sep. 20th, 2011; revised: Sep. 28th, 2011; accepted: Oct. 8th, 2011.

Abstract: Unstructured peer-to-peer network (P2P) is the main way to implement shared system, but retrieval efficiency is generally low due to search of blindness. In this paper, a new approach to information retrieval based on unstructured P2P systems is presented, by using clustering domain and grouped P2P network based on semantic. The results show that the searching mechanism has good performances on the search success rate and load balancing.

Keywords: Peer-to-Peer Network; Information Recommendation; Semantic Cluster

非结构化对等网络中信息检索策略研究

刘　溪
华中科技大学控制科学与工程系，武汉
Email: llxfrankcs@gmail.com
收稿日期：2011 年 9 月 20 日；修回日期：2011 年 9 月 28 日；录用日期：2011 年 10 月 8 日

摘　要：非结构化对等网络(P2P)是共享系统的主要实现方式，但是由于搜索的盲目性，其检索效率又普遍低下。本文将聚类域和语义分组引入到非结构化对等网络搜索技术中，结合非结构化对等网络中的洪泛搜索机制，提出了基于语义聚类的资源搜索策略。仿真实验的结果表明，该系统所采用的信息检索策略能够有效地提高信息检索的查询成功率，降低网络负载，取得了良好的效果。

关键词：对等网络；信息检索；语义聚类

1. 引言

随着互联网技术飞速发展，网络资源的急剧扩大，传统的互联网应用模式客户/服务器(Client/server，C/S)模式受到严重的挑战，造成大量的网络资源的浪费，对等计算(Peer-to-Peer，P2P)模式以其高可靠性和高可用性已越来越受到人们的青睐。在 P2P 模式下，网络中每个节点即充当服务器为其他节点提供服务，又是客户享受其他节点提供的服务。节点间相互平等，相互共享存储空间，网络带宽等资源。

近年来，随着 P2P 对等计算模式的广泛应用，P2P 网络中共享资源的数量也在急剧增加，查找资源、定位资源的效率直接影响到用户获取资源的速度和网络的带宽利用率。如此如何在网络中准确的定位到搜索资源所在的位置成为影响 P2P 网络发展的关键因素。基于此，大量文献对 P2P 网络中资源定位搜索进行了研究。文献[1]中，利用概率平衡树和匹配路径，提出了一种节点可以根据查询内容进行路由决策的 P2P 语义路由模型，在文档[2]中提出了一种根据文档的语义将节点聚类，但没有考虑向量空间模型。

本文在分析传统的信息检索的相关模型以及存在的不足的基础上，对传统的 P2P 网络信息检索的查询机制进行了改进，将聚类域和语义分组引入到非结构

化对等网络搜索策略中，提出了一种基于语义聚类的资源搜索策略，实验结果表明，算法对系统的性能具有良好的改进作用。

2. 相关工作

2.1. 信息检索中的向量空间模型

资源搜索的实现离不开检索模型，常见的检索模型有布尔模型、向量空间模型、概率模型。其中向量空间模型由于其性能优秀、实现方便已成为流行的检索模型。向量空间模型 VSM(Vector Space Model)最早是由 Salton 提出[3]。在 VSM 中，用 d_j 表示文本，k_i 表示关键词，t 表示文本中关键字的数目，w_{ij} 表示文本 $\boldsymbol{d}_j$ 内容与关键字 k_i 的相关性的权值。如此文本集 D 中一个文档 $\boldsymbol{d}_j$ 的向量表示为：

$\boldsymbol{d}_j = \{w_{1j}, w_{2j}, \cdots, w_{tj}\}$，如果文本中有这个关键字，其权 $w_{ij} \geq 0$，而对于没有出现在文本中的关键字，$w_{ij} = 0$。

这里用 $\boldsymbol{q}$ 表示查询文本，$\boldsymbol{q}$ 向量表示为 $\boldsymbol{q} = \{w_{1q}, w_{2q}, \cdots, w_{tq}\}$，则 $\boldsymbol{d}_j$ 和 $\boldsymbol{q}$ 的相似度如(1)式所示：

$$\sin(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\boldsymbol{d}_j \boldsymbol{q}}{|\boldsymbol{d}_j| \times |\boldsymbol{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (1)$$

其中 $|\boldsymbol{d}_j|$ 和 $|\boldsymbol{q}|$ 是文档 $\boldsymbol{d}_j$ 和查询向量 $\boldsymbol{q}$ 的模(Norms)。

因为 $w_{ij} \geq 0$，$w_{iq} \geq 0$，则 $0 \leq \sin(\boldsymbol{d}_j, \boldsymbol{q}) \leq 1$

向量空间模型通过计算向量 $\boldsymbol{d}_j$ 与 $\boldsymbol{q}$ 之间的相似度来评价文档 $\boldsymbol{d}_j$ 与查询文档 $\boldsymbol{q}$ 的相关程度，根据相似度的大小对文档进行排序。只要文档与查询部分匹配，就有可能被检出来。为了避免被检索出的文档太多，可以给 $\sin(\boldsymbol{d}_j, \boldsymbol{q})$ 设定一个阈值，相似度大于阈值的文档被检索出来，小于阈值的文档被舍弃。在实际中，一篇文档中出现的不同关键词对该篇文档描述的语义的贡献度是存在差异的，关键词出现的文档数越多，表明该关键词对文档的区分作用越小[4]。

如果用 $freq_{ij}$ 表示关键词 k_i 在文档 $\boldsymbol{d}_j$ 中出现的频率，N 表示文档集中所有文档的总数，n_i 表示文档集中包含关键词 tk_i 的文档总数。根据 $f - idf$ 加权方法，其中 f (Term Frequency)表示关键词在文档中出现的频率，idf 表示所查关键词在文档集中所占比例的倒数，即“逆文档频率”(Inverse Document Frequency)。在文献[5]中给出 $f - idf$ 加权方法的详细介绍。如此，对 w_{ij} 进行修正：

$$w_{ij} = f_{ij} \times \log \frac{N}{n_i} \quad (2)$$

其中 $f_{ij} = freq_{ij} / \max freq_{ij}$, $idf_i = \log \frac{N}{n_i}$

这样，就将关键词在所查当前文档中出现的频率和在所有文档中出现的频率相结合，来确定关键词在当前文档中的权重 w_{ij}。

2.2. 常见的 P2P 信息检索方法

根据 P2P 网络结构的不同，现有的 P2P 信息检索方法主要有三类，基于集中式目录服务器的信息检索、基于结构化 P2P 网络的信息检索和基于非结构化 P2P 网络的信息检索。

早期的 P2P 应用软件(Napster)主要采用集中式目录服务器，目录服务器中保存着所有资源的标识符和索引信息。当有用户需要检索信息资源时，用户首先连接目录服务器，目录服务器通过资源标识符查找资源所在位置，并返回给用户节点，用户直接定位到资源位置节点，下载所需资源。由于有目录服务器的存在使搜索速度快，也节省了网络带宽，但也限制了系统中的节点数量，容易出现单点失效的故障。

结构化 P2P 网络没有目录服务器，节点间具有相对稳定和规则的拓扑结构，资源的放置也是由精确的算法分配到特定的节点上。结构化 P2P 网络通常采用分布式哈希表(Distributed Hash Table，DHT)构建，资源定位准确，搜索迅速，可以保证一定的效率，其代表性研究项目主要包括 CAN、Tapestry、Chord、Pastry 等[6]。但结构化 P2P 网络不支持非精确匹配的查询，维护也相当困难，当有一个节点非正常离开网络后，将导致拓扑结构的变化，需要复杂的步骤纠正系统的错误。目前结构化 P2P 网络还缺乏在 Internet 中大规模真实部署的实例。

非结构化 P2P 网络对网络拓扑和文件的位置没有精确的控制，节点的拓扑关系一般较为松散，节点间无需维护拓扑信息，扩展性和容错性好，是构建 P2P 网络的首选。代表性研究项目主要包括 Gnutella、

Freenet 等[7]。但网络中资源的搜索一般采用洪泛和随机游走算法实现，把查询消息散布于整个网络，使网络开销大，带宽利用率低。因此，设计出优秀的非结构化路由算法，是非结构化对等网络研究的热点[8]。

3. 基于语义聚类的资源搜索策略研究

在非结构化对等网络中，节点网络规模巨大，随着节点的增加，网络资源也在不断膨胀中，这就个资源搜索带来了巨大的挑战。因此，如何准确的定位资源所在的位置，将是 P2P 网络急需解决的问题。如果在信息检索前，对资源进行分类，具有相似资源的节点组成特定的虚拟区域，搜索是根据相应的查询关键字在对应的区域类搜索，可大大节省搜索的效率。在这里本文利用聚类算法，将具有相同和相似资源的节点组成特定的聚类域，聚类域的形成将使搜索具有明显的方向性。本文根据节点建立聚类域，将大大改善搜索的性能。

本系统建立是基于以下假设：

1) 每个节点上的共享文档具有较高的相似性。

2) 查询请求可利用向量空间模型表示成向量形式。

这样每个新节点加入到网络中后，首先利用向量空间模型对文档进行语义提取，在各个节点上建立文档索引表，建立节点的特征向量，寻找特征向量相似的节点发送申请加入请求，节点根据语义相似度阀值以及聚类域分组成员是否已达到上限来判断是否给出应答请求信号。当新节点加入聚类域后，聚类域内所有成员更新自己的节点列表。

聚类域形成后，域中成员根据文本聚类算法求的域的中心，进而得到聚类域的语义特征向量。当有查询请求时，查询请求首先根据向量空间模型表示成向量形式，查询请求特征向量与聚类域特征向量进行匹配，找出相似性的特征向量，在聚类域中进行查询。

算法步骤描述：

Step 1：利用 VSM 模型对节点中文档进行处理，得到所有文档的 VSM 向量空间模型。得到节点的语义信息，建立本地的数据库和路由表。

Step 2：通过节点语义信息找寻语义相似的节点，对相似的节点进行聚类，形成相关的聚类域。

Step 3：当节点发起关于某一资源的查询时，首先节点对邻居节点数据库进行资源查询，查找满足条件的文档返回。

Step 4：节点将查询消息洪泛至组内其他成员，聚类域中成员对查询消息进行响应，直到查询成功和失败的条件被满足。

4. 实验分析

为了对系统进行定性分析，将本文中语义聚类的非结构化 P2P 网络与 Gnutella[9]网络相比较。Gnutella 采用洪泛搜索机制进行查询，查询消息在各个节点之间进行随机转发，直到找到资源所在位置或者所设阀值 TTL = 0 为止。由于客观条件的限制，本文只通过仿真实验重点对检索效率进行了测试。仿真实验在一台 PC 机上进行，仿真中的文档集来自于维基百科，包括体育、娱乐、教育、科技等相关类别，文档集中包括不同的关键字大约 523 个，设置 300 个节点，每个节点管理大约 100 个文件。对查询成功率和带宽利用率两个方面进行了详细的分析。

4.1. 查询成功率

查询成功率是节点发送的查询消息数和收到查询成功的消息数之比，查询成功率越高说明搜索算法越好。图 1 给出本系统与 Gnutella 网络在查询成功率上的对比。

从实验结果可以看出，随着查询数的增加，本系统相对于 Gnutella 网络在查询成功率上的优势得到体现。这说明信息推荐服务确实有助于改善信息检索的效果。

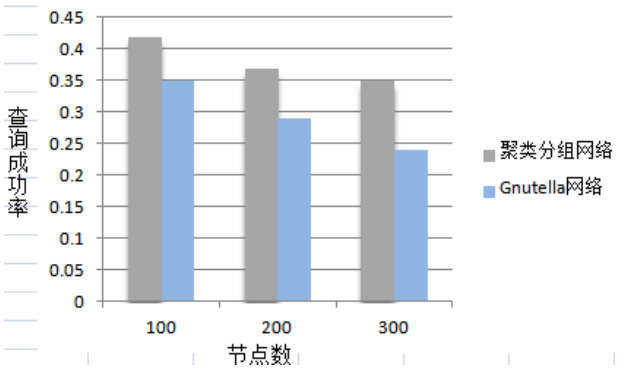

Figure 1. Clustering inquires the success rate compared
图 1. 聚类分组查询成功率比较

4.2. 带宽利用率

对等网络的主要性能瓶颈来自于网络带宽和时延，高效率的检索技术要尽量以最少的带宽消耗获得最满意的检索效果。网络带宽利用率是指复制一个文件需要的时间和网络带宽，它也是衡量副本创建策略的一个重要因素。对于一个给定的网络，网络带宽利用率越低说明副本策略越好。由于仿真实验无法直接测量出查询执行过程中的实际带宽利用率，于是我们可以利用信息检索过程中的消息量来衡量系统带宽利用率。图 2 给出本系统与 Gnutella 网络在带宽利用率上的对比。

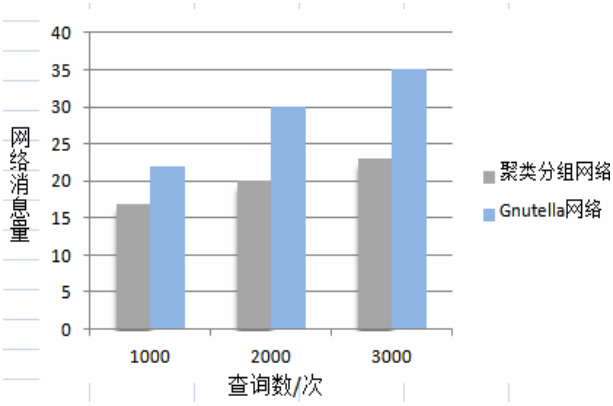

Figure 2. Bandwidth utilization efficiency
图 2. 带宽利用效率

5. 结论

资源搜索策略是对等网络研究的难点，本文将语义聚类和向量空间模型引入对等网络，可以更准确快速的定位用户所查询的文档。避免了非结构化 P2P 网络中基于洪泛机制的盲目搜索，也克服了结构化网络的缺乏有效的基于内容的搜索问题，实验结果分析表明，本文所提出的算法具有很大的实用性。

参考文献 (References)

[1] 许立波, 于坤, 吴国新. 基于匹配路径和概率平衡树的 P2P 语义路由模型.软件学报, 2006, 7(10): 2106-2117.
[2] P. Triantallou, C. Xiruhaki, M. Koubarakis, et al. Towards high performance peer-to-peer content and resource sharing systems. In: Proceedings of the Conference on Innovative Data Systems Research (CIDR), 2003.
[3] G. Salton. A. Wong. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613-620.
[4] 凌波, 吕永成, 周水庚等. P2P 信息检索及其优化策略[J]. 计算机科学, 2006, 33(8): 173-177.
[5] G. Salton, M. McGill. An introduction to modern information retrieval. New York: McGraw-Hill, 1983.
[6] 李绍滋, 曹阳, 周昌乐. 基于非结构化的 P2P 信息检索关键技术研究[J]. 智能系统学报, 2006, 1(2): 74-78.
[7] 王建荣. 对等网络中的查询搜索机制与信任模型研究[D]. 天津大学, 2008.
[8] 刘勇. 大规模对等资源共享关键技术研究[D]. 电子科技大学, 2009.
[9] Gnutella. The annotated gnutella protocol specification, 2009. http://rfc-gnutella. sourceforge. net/developer/stable/index.html

