

考虑搜索行为与节点内容的 P2P 搜索方法^①

王海波¹ 高涇萍¹ 刘占波¹ 初晓峰² (1.牡丹江医学院 教育技术与信息中心 黑龙江 牡丹江

157011;2.深圳研发中心 中兴通讯 518057 广东 深圳)

摘要: 针对基于兴趣驱动的 P2P 搜索方法在挖掘节点兴趣和扩展搜索兴趣的上下文语义等方面不足,改进 Social-P2P 算法,给出考虑搜索行为和节点内容的 P2P 搜索方法。引入概念格理论,根据节点内容和用户搜索行为建立朋友列表,以朋友列表为形式背景构造概念格,建立兴趣域。搜索消息在概念格内查询,缩短搜索路径和减少搜索消息,概念偏序关系扩展查询消息的上下文语义,增强搜索精确度。实验验证该方法比 Social-P2P 搜索方法和泛洪搜索方法具有更好的召回率和精确率。

关键词: 无结构 P2P 网络; P2P 搜索; 兴趣域; 概念格

Search Method in P2P Considering Search Behavior and Content of Nodes

WANG Hai-Bo¹, GAO Jing-Ping¹, LIU Zhan-Bo¹, CHU Xiao-Feng²

(1.Center of Educational Technology and Information, Mudanjiang Medical University, Mudanjiang 157011, China; 2. ZTE Shenzhen R&D center, Shenzhen 518057, China)

Abstract: There are deficiencies of the interest-based driven P2P search method in discovering interest of node and expanding interest in the context of semantic search. This paper improves the Social-P2P algorithm and puts forward a search method considering the search behavior and the content of nodes. The method introduces the concept lattice theory. According to the node content and user search behavior, a friend list is established, then formal context is extracted from the friend list to build a concept lattice, and the interest domain. It searches for messages within the query concept lattice, cuts down the search path and reduces the search message. Concepts which have deflection order expands the context of semantic query messages and enhance the search accuracy. Experimental results indicate that compared with the traditional flooding method, Social-P2P search method has better recall rate and accuracy.

Keywords: unstructured P2P network; P2P Search; interest domain; concept lattice

P2P(Peer-to-Peer)模式下,服务器与客户机界限消失,核心是 P2P 搜索技术。基于兴趣域 P2P 搜索方法成为无结构 P2P 网络搜索技术研究热点。基于兴趣域 P2P 搜索的文献较多,NeuroGrid^[1]统计历史搜索经验聚集兴趣相似节点,依靠关键字查找,缺乏语义相关性方面扩展能力。Social-P2P^[2]引入社会网

论,依靠人工分类搜索主题,上下文语义缺乏灵活性。文献[3-6]从不同角度建立兴趣域,呈现小世界特性,提高搜索效率。本文综合用户行为与节点的本地内容两视角,给出考虑搜索行为和节点内容的 P2P 搜索方法。根据社会网^[7]与小世界理论建立朋友列表,聚集相似兴趣用户,建立概念格,内容相近朋友聚集为概念。

^① 收稿时间:2009-12-01;收到修改稿时间:2010-01-01

1 Social-P2P搜索算法

2007 年 Lu Liu 等提出 Social-P2P 搜索算法^[2], 引入社会网理论, 模拟人与人之间关系, 建立节点之间联系, 聚集兴趣相近节点, 体现“小世界”特性。Social-P2P 方法如下:

①网络节点接收搜索消息, 查询本地知识索引中与搜索主题相关索引信息。如找到与搜索主题相关目标节点, 修改相关主题索引的权值。否则去②。

②寻找与搜索主题相关主题索引, 转发搜索消息至包含此主题的朋友节点。建立信息目录分类确定主题间相关性。含相关主题的节点权值用 N_{match}/N_{totals} 表示。 N_{match} 表示知识索引中节点与此主题相关连接数, N_{totals} 表示包含此主题的领域中所有主题数。

③将相关连接按权值排序, 向朋友节点转发搜索消息。

2 考虑搜索行为与节点内容的P2P搜索方法

针对 Social-P2P 建立兴趣域视角单一和上下文语义扩展能力不足, 引入概念格理论, 给出考虑搜索行为和节点内容的 P2P 搜索方法。建立朋友节点列表 (friend neighbor node list, FNL), 根据朋友节点和其内容构成的形式背景建立朋友概念格 (friend concept lattice, FCL), 利用概念和概念间偏序关系聚类兴趣相似节点。形成概念扩展消息的语义查询。

2.1 算法兴趣域的建立

2.1.1 节点内容匹配算法

节点内容兴趣挖掘在于内容表示形式, 提取文档高频词汇为关键字表示节点内容。基于向量空间模型建立节点内容相似度匹配算法, 文档和用户查询式转化成向量并映射到文本向量空间。向量夹角度量用户查询和被检索文档间的相似程度, 节点 n_i 和 n_j 对应向量为 $\vec{n}_i = (w_{i,1}, w_{i,2}, \dots, w_{i,m})$ 和 $\vec{n}_j = (w_{j,1}, w_{j,2}, \dots, w_{j,m})$, 内容相似度公式如式(1), $w_{i,l}$ ($l \in [1, m]$) 为节点 n_i 的第 l 个关键字的数值表示, m 为关键字个数。

$$\text{sim}(n_i, n_j) = \frac{\sum_m (w_{i,m} \cdot w_{j,m})}{\sqrt{\sum_m w_{i,m}^2} \sqrt{\sum_m w_{j,m}^2}} \quad (1)$$

2.1.2 建立 FNL

FNL 是维护与本地节点兴趣相似节点的表, 从两方面考虑兴趣相似度评价标准^[8]: ①朋友节点与本节点存储内容相近程度, 即节点间相同关键字个数。②一段时

间内搜索行为反映节点兴趣, 目标节点命中次数评价节点间兴趣相似度。FNL 节点 n_i 维护关键字集合 M_i 、最近信息命中计数值 QueryHit 和朋友节点关系紧密度 Friend-Weight, QueryHit 表示朋友节点最近一段时间第一跳命中次数。按照 Friend-Weight 排序, FNL 结构如表 1。

表 1 FNL 结构表

节点地址	关键字集合	QueryHit	Friend-Weight
Node ₁	$M_1(m_1, m_2, m_4, m_6)$	3	0.85
Node ₂	$M_2(m_3, m_4, m_r, m_8, m_9)$	8	0.76
Node ₃	$M_3(m_3, m_5)$	2	0.65
Node ₄	$M_4(m_7, m_9, m_{10})$	6	0.64
...	...	...	...

长度为 K 数据窗口 FW(friend-window)存储最近 K 次搜索的关键字, $node_i$ 维护关键字列表, 兴趣匹配算法将命中的新节点 $node_j$ 加入 FNL。如搜索的关键字 mr 所在兴趣域 Cr 包含 FNL 中节点 $node_i$, 且关键字 mr 在 FW 内, 节点 $node_i$ 的 QueryHit 加 1。更改本节点和 $node_i$ 的 Friend-Weight。Friend-Weight 评价公式如下:

$$\text{Friend-Weight}_i = \partial \text{sim}(n, n_i) + (1 - \partial) \text{QueryHit}_i / K \quad (2)$$

∂ 为节点内容相似度权值且 $\partial < 1$, $\text{sim}(n, n_i)$ 为 $node_i$ 与本节点之间内容相似度, 如式(1)。QueryHit_i 为 $node_i$ 在 FW 中命中次数, K 为 FW 的尺寸。综合相似程度和命中频率得到 Friend-Weight_i 值, FNL 前 F 个节点作为朋友节点建立概念格。

2.1.3 更新 FNL

节点内容相似度匹配算法判断新节点是否加入朋友列表, 节点加入朋友列表后, 作为候选朋友节点待更新。一段时间(R 个 FW 数据窗口大小的查询次数)后, 重新排序朋友列表, 前 F 个作为新朋友节点, 删除其他候选节点。

2.1.4 FCL 的建立

以 FNL 为形式背景, 渐进式方法构造 FCL, 不同于 Godin^[9]算法, FCL 更新指形式背景的更新, 包含属性和对象的增加, 既非多概念格横向合并也非多概

念格纵向合并,修改 Godin 算法得到 P2P 朋友概念格合并算法。

邻居节点的形式背景为 $K=(G,M,I)$, G 为对象集合, M 为属性(关键字)集合, G 与 M 之间关系为, 其与新增节点形式背景 $K_1=(G_1,M_1,I_1)$ 合并后的形式背景 $K^*=(GG_1, MM_1, II_1)$, 由 K^* 构造概念格 $L(K^*)$, 可以在 $L(K)$ 基础上构造 $L(K^*)$, 用 Godin 算法生成新 FCL。

2.1.5 FCL 的更新

1) 对象的删除

节点退出 FCL 删除相应对象, FCL 删除一个对象时, 节点可能具有其它节点所没有的属性(关键字), 删除这些属性对原格结构无影响, 带来形式背景外延与内涵同步更新问题, 修改原格的概念之前修改形式背景内涵集合, 去除相应属性, 然后在原 FCL 中追加对象。概念格外延指朋友节点的 URI , 内涵指每个节点内容的关键字表示。

更新后的 FCL 中的概念划分为四种类型: 不变概念、删除概念、更新概念、冗余概念。更新或删除的概念与对象 x^* 有关, 在删除对象 x^* 时, 从包含修改对象的最小概念出发, 减少遍历的节点数。使用队列数据结构 CQueue, 对概念格由下到上的层次遍历。

2) 属性增加

FCL 的形式背景中节点增加关键字, FCL 的属性增加, 节点不增加和退出, 只扩充形式背景的内涵或修改个别概念的外延。在基于属性的渐进式构造概念格算法 CLIF_A^[10] 基础上建立 FCL 属性增加算法。CLIF_A 算法针对一个形式背景增加一个属性的情况, 网络节点更新内容有两种情况: ①新加入的属性 m^*M , M 是形式背景 $K=(G,M,I)$ 的内涵, 可采用 CLIF_A 算法更新 FCL。②新加入的属性 m^*M , 只是该属性的对象集发生了变化, 对相关概念进行修改即可。CLIF_A 算法考虑第一种情况, 增加第二种情况的处理。

3) 属性的删除

在 FCL 属性增加算法基础上建立 FCL 属性删除算法。FCL 中属性的删除有两种情况: ①删除属性后 m^*M , M 是形式背景 $K=(G,M,I)$ 的内涵, 与对象的删

除类似, 分情况修改父概念与子概念的连接。②删除属性后 m^*M (m^* 还在 K 中), 只修改涉及到属性 m^* 的概念。

2.2 考虑搜索行为和节点内容 P2P 搜索方法

1) 初始化, FNL 为空, 以物理邻居节点作为朋友节点加入 FNL。以 FNL 节点集合 N 为形式背景 $K=(G,M,I)$ 的外延, 以关键字集合 M 为形式背景 K 的内涵, 建立 FCL。设定 TTL 初值为 R 更新时间计数器 Tr 。

2) 节点发出查询消息二元组 (URI, m_r) , URI 为查询节点 URI , 以关键字 m_r 进行查询并将 m_r 添加到 FW 中, 窗口向右移动一步。先查询本地, 然后查询 FCL 中属性为 m_r 的概念。在 FCL 中的最大下界中查询该概念格内涵中是否包含 m_r , 如 FCL 内涵中没有 m_r , 以泛洪方式向所有朋友节点发查询消息。如 FCL 内涵中含 m_r , 从 FCL 的最小上界开始深度遍历 FCL, 记录遍历路径。在遍历路径中去除最小上界与最大下界概念, 根据剩余概念情况将各概念的内涵组成新扩展查询消息。扩展的查询消息直接向概念外延节点转发, TTL 计数, 减至零时停止转发, Tr 开始计数并减 1。

3) 朋友节点收到转发消息在本地查询, 如有关键字 m_r , 向查询节点发出反馈消息二元组 (URI, m_r) , URI 为命中节点的 URI , 通知查询节点命中, 继续在 FCL 中查询同 2), 不向查询节点转发查询消息(避免构成回路), 消息转发给相似概念的网络节点, TTL 值减 1。

4) 查询节点收到反馈消息后, 在本地 FNL 中查找地址为 URI 的朋友节点, 节点 QueryHit 值加 1, 更新 Friend-Weight。如本地 FNL 中没有此节点, 则根据式(1), 计算本节点与命中节点 n_i 的 $sim(n, n_i)$, 如 $sim(n, n_i) > \varepsilon$, n_i 加入到 FNL 的末尾, 否则抛弃。

(5) 如 Tr 值减为零, 查询节点更新朋友列表, 节点将 FNL 中的节点与候选的节点重新按 Friend-Weight 值排序, 取前 F 个节点构造新朋友列表。如 Tr 值不为零, 则转到(7)。

(6) 根据新 FNL 构造形式背景, 按照更新后的形式背景利用概念格的维护算法动态的删除概念格的属性与对象, 再将新加入的对象与属性动态的构造新的概念格 FCL。

7)反馈结果排序后交给用户。

假设节点 n 的 FNL 中有 5 个朋友节点, 以 5 个朋友节点的关键字集合为 {a,b,c,d,e} 建立形式背景如表 2。

表 2 根据 FNL 建立的形式背景表

M \ G	a	b	c	d	e
1	0	1	0	1	1
2	0	1	0	1	0
3	1	0	1	0	1
4	0	1	1	0	0
5	1	0	0	0	0

FCL 中的查询示例如图 1 所示。

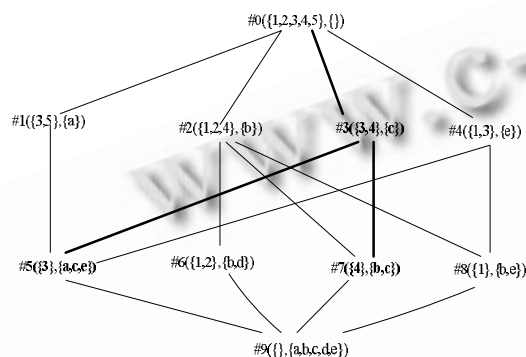


图 1 FCL 中的查询示例图

3 仿真实验

实验采用东京大学 Sam Joseph 教授开发的通用 P2P 模拟器—NeuroGrid。数据来自搜狐分类目录资料库, 随机抽取 1000 个主题作为关键字。主题随机分配给 2000 个文档, 每个文档只有一个关键字。搜索算法与 Gnutella^[1]算法和 Social-P2P 算法^[2]比较。实验硬环境 CPU (Pentium4 1.7G HZ)、内存 (1.0GB)PC 机, 模拟器 neurogrid0.2.1。

以下三个标准衡量算法性能:

1)召回率

$$\text{召回率} = \frac{\text{检索到正确结果数}}{\text{网络中所有正确结果数}} \quad (3)$$

2)搜索范围

$$\text{搜索范围} = \frac{\text{搜索过程中访问节点数}}{\text{网络节点总数}} \quad (4)$$

3)一跳命中率

$$\text{一跳命中率} = \frac{\text{搜索中一跳产生的命中数}}{\text{本次搜索中所有的命中数}} \quad (5)$$

3.1 算法的召回率

如图 2 所示, Gnutella 方法将搜索消息转发至 TTL 范围内节点, 召回率 91%左右。随搜索进行本文给出的算法召回率逐渐提高接近 Gnutella, 通过建立兴趣域将相似兴趣节点聚集一起, 提高搜索成功率, Social-P2P 与其相比效果较差。

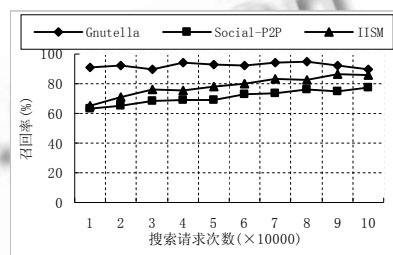


图 2 召回率比较

3.2 算法的搜索范围

如图 3 所示, Gnutella 方法搜索 95%左右节点, 泛洪方式转发查询消息, 产生消息冗余。本文给出的算法和 Social-P2P 算法都在以往搜索过程中收集节点信息, 均实现有目的转发, 避免盲目搜索带来的网络开销, 缩小搜索范围。

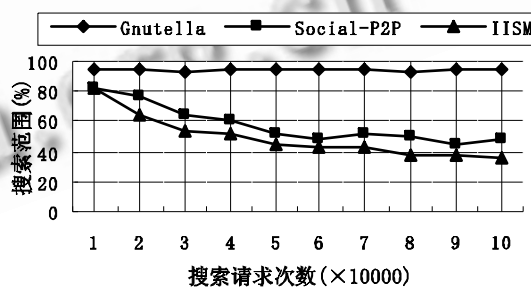


图 3 搜索范围比较

Social-P2P 方法记录历史搜索经验聚类相似兴趣节点, 未考虑节点内容反应节点兴趣这一因素。Social-P2P 搜索第二阶段的节点选择方式与概念格方法相比不能较好的扩展搜索消息的上下文语义, 本文给出的算法收敛更快。前 30000 次搜索, 本文给出的算法搜索范围收敛速度明显快于 Social-P2P 搜索方法, 随着搜索次数的增加, 聚类到一定程度, 二者搜索范围收敛速度减慢, 趋于平稳。

3.3 算法的一跳命中率

如图 4 所示, 前 30000 次搜索本文给出的算法和 Social-P2P 的一跳命中率不断升高, 最后趋于平稳。Gnutella 的一跳命中率在 5% 上下。本文给出的算法从用户与节点内容角度聚集兴趣相似朋友节点建立朋友列表。兴趣相近节点聚集程度高于 Social-P2P 方法。搜索前 30000 次, 本文给出的算法搜索一跳命中率曲线比 Social-P2P 变化的更快, 趋于平稳后, 有 3% 左右差距。本文给出的方法搜索路径长度更小, 降低网络负载, 减少消息转发量。

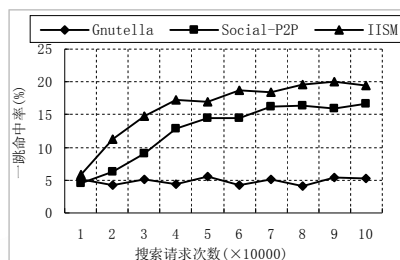


图 4 一跳命中率比较

实验结果表明, 本文给出的搜索方法更有效聚集相似兴趣节点, 概念格比分类目录更能体现搜索消息上下文语义。

4 结论

通过引入概念格理论, 对 Social-P2P 方法进行改进, 给出考虑搜索行为和节点内容的 P2P 搜索方法。从用户与节点内容两个视角体现节点兴趣。概念格理论引入基于兴趣驱动的 P2P 搜索方法中。实验证明此方法与 Social-P2P 搜索方法和传统泛洪式搜索方法相比, 提高搜索效率减轻网络负载。

参考文献

- 1 Joseph S. NeuroGrid: Semantically routing queries in peer-to-peer networks. Proc of the International Work-shop on Peer-to-Peer Computing. 2002(2376): 202 – 214.
- 2 Liu L, Antonopoulos N, Makin S. Social Peer-to-Peer for Resource Discovery. Proc. of the 15th Euromicro International Conference on Parallel, Distributed and Network-Based Processing. 2007:459 – 466.
- 3 Crowcroft J. Application level programmable Inter-network environment, 2005. <http://www.cs.ucl.ac.uk/research/alpine/alpine.html>.
- 4 Sripanidkulchai K, Maggs B, Zhang H. Efficient content location using interest-based locality in peer-to-peer systems. INFOCOM.2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, 2003:2166 – 2176.
- 5 王菁, 张焕杰, 杨寿保, 高鹰. 利用集合差异度实现基于内容聚类的 P2P 搜索模型. 中国科学院研究生院学报, 2007, 24(2): 241 – 247.
- 6 Tan YH, Li YP, Chen ZP, et al. Research and Implementation on Routing Scheme Based on Interest Mining in Unstructured P2P Systems. 7th International Conference on Web-Age Information Management Workshops, 2006:2 – 2.
- 7 Crespo A, Garcia-Molina H. Semantic overlay networks. [2009-8-2]. <http://www.db.stanford.edu/crespo/publications/op2p.pdf>.
- 8 杨舰, 吕智慧, 钟亦平, 张世勇. 一种基于兴趣域的高效对等网络搜索方案. 计算机研究与发展, 2005, 42(5): 804 – 809.
- 9 Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on Galois (concept) lattices. Computational Intelligence, 1995, 11(2): 246 – 267.
- 10 李云, 刘宗田, 陈峻. 基于属性的概念格渐进式生成算法. 小型微型计算机系统, 2004, 25(10): 1768 – 1771.
- 11 Gnutella. [2009-8-6]. <http://www.gnutella.com>, 2001.