P2P 社区的信息检索算法设计

王欣惠
WANG Xin-hui

北京农业职业学院 信息技术系,北京 102442
Technological Information Department,Beijing Vocational College of Agriculture,Beijing 102442,China
E-mail:wang-xinhui@163.com

WANG Xin-hui.Algorithm design of information retrieval based on P2P community.Computer Engineering and Applications, 2010,46(3):134-136.

Abstract: In order to effectively locate resourse and search information in P2P community,a hybrid P2P topology is adopted and the document searching in P2P community is divided into three processes:local retrieval,intra-search and inter-search.In local retrieval,the new term weight calculation method is designed to solve the problem about txt search in homogeneous collection.In intra-search and inter-search,a part of peers which are more relevant to the query to conduct the retrieval task is selected.Experimental results show that this method can get better retrieval results with lower cost.

Key words: peer-to-peer community;information retrieval;algorithm design

摘 要: 为解决 P2P 社区的资源定位及信息检索问题,采用混合型 P2P 网络模型,将社区内的检索划分为本地检索、组内搜索和组间搜索。对于本地检索设计了新的词条权重的计算方法,解决了同构文档集内的文本检索问题。对于组内搜索和组间搜索,通过设计节点选择策略,使一部分与查询相关度高的节点执行查询任务。最后提出结果融合的方法并对特定的实验数据进行测试,实验表明设计的算法在较小的查询开销下,能取得较好的检索效果。

关键词: P2P 社区;信息检索;算法设计
DOI: 10.3778/j.issn.1002-8331.2010.03.040 **文章编号:** 1002-8331(2010)03-0134-03 **文献标识码:** A **中图分类号:** TP393.09;TP301.6

1 前言

在 P2P 环境下进行资源定位是 P2P 研究的核心问题之一[1-3]。对 P2P 资源的定位技术有非结构化 P2P、结构化 P2P 和基于兴趣局部性优化的 P2P 搜索。P2P 信息检索分为纯粹的 P2P、混合型 P2P 及基于 P2P 社区的信息检索。P2P 社区将主题相同的节点聚集在一起,为信息检索提供了良好的环境,但目前主要是按用户提交的检索行为来划分用户兴趣,没有充分挖掘节点共享内容所体现的节点兴趣,因此需要去挖掘其共享内容所反映的兴趣,以便网络中其他节点在需要时能实现高效检索。P2P 全文信息检索中用户提出的查询词尺度远小于共享信息的尺度,查询所反映的共享兴趣更有限[4-5]。因此必须设计高效的检索算法让用户在海量数据中找到所需信息。

2 P2P 社区信息检索网络模型

采用何种检索机制往往取决于 P2P 系统的网络模型[6]。以 G0.6 为代表的 P2P 系统具有易于管理、高效健壮和易扩展的优点,在实际中得到广泛应用[7-8]。其网络模型见图 1。内层节点是超节点(有较长在线时间、较高机器性能和较多资源的节点),外层节点是叶节点。每个超节点维护直接相连的超节点和直接相连的叶节点列表。每个叶节点维护与自己直接相连的超节点列表,表中记录超节点的 ID 和节点描述。超节点最多可以与社区内的 100 个节点(40 个超节点,60 个叶节点)相连。一个超节点同相连的叶节点构成一个兴趣小组(IG),超节点负责处理 IG 中的叶节点和其他超节点的查询请求,可转发查询;叶节点只能提交和执行查询,不能转发查询。

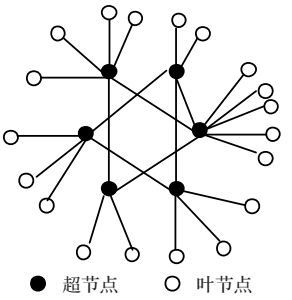

图 1 G0.6 **网络模型图**

根据图 1 网络模型,整个社区的检索分为节点内、组内和组间搜索。节点内检索是指接收到查询的节点据查询检索本地文档集; 组内检索是指在 IG 内接收到查询的超节点选择部分组内叶节点,让其执行查询任务。组间检索是指收到查询的超节点会将查询转发给其他超节点,使得查询在其他的 IG 中执行。

基金项目: 河北省科技厅基金(the Foundation of Science and Technology Department of Hebei Province No.06213556)。
作者简介: 王欣惠(1974-),女,讲师,研究方向为 P2P 技术及数据库应用。
收稿日期: 2008-05-05 **修回日期:** 2008-10-31

3 节点内信息检索算法设计

同构文档集是指节点内的文档集。在同构文档集中,与主题非常相关的词条会在文档集中的许多文档中出现,这部分词条对文档集主题非常重要。无法准确计算词条的权重,就不能获得准确的检索结果。通过放大同构文档集内不同文档间的差异可更准确地计算词条的权重。

已知文档集 $C=d_1,d_2,\cdots,d_m$ 及其索引词条集合 $TS=t_1,t_2,\cdots,t_n$,各词条在各文档的出现情况及文档间的相关关系;求解对于文档 $d_j\in C$ 以及词条 $t_i\in TS$,词条 t_i 在 d_j 中的权重 ω_{ij}。即求 d_j 对应的权重向量 $\boldsymbol{d}_j=(\omega_{1j},\omega_{2j},\cdots,\omega_{nj})$。

利用推理网络查询 q 与文档 d_j 的相关度可看作是条件概率 $P(q|d_j)$,则

$$P(q|d_j)=\sum_{k=1}^{n}P(q|t_k)P(t_k|d_j)P(d_j) \tag{1}$$

若 q 中仅包含一个词条 t_i,可将 $P(t_i|d_j)$ 视为 t_i 在 d_j 中的权重 ω_{ij},则

$$\omega_{ij}=P(t_i|d_j)=\sum_{k=1}^{n}P(t_i|t_k)P(t_k|d_j)P(d_j) \tag{2}$$

式中,$P(d_j)=\frac{1}{|C|}$。设 P 为出现 t_i 的文档与 d_j 相关的概率,则

$$P(RDF(t_k,d_j)=r)=C_l^T p^r(1-p)^{l-r} \tag{3}$$

其中,l 为文档集 C 中包含 t_i 的文档数目,即 $l=df(t_k)$。利用 P 的估计值 $\hat{P}$ 来估算 $P=(t_k|d_j)$:

$$\hat{P}(t_k|d_j)=\hat{P}=\frac{RDF(t_k,d_j)}{df(t_k)} \tag{4}$$

对于 $P(t_i|t_k)$,采用估算方法:

$$P(t_i|t_k)=\frac{df(t_i,t_k)}{df(t_i)} \tag{5}$$

由式(2)、(4)和(5)得:

$$P(t_i|d_j)=\sum_{k=1}^{n}RDF(t_k,d_j)\cdot\frac{df(t_i,t_k)}{df(t_i)df(t_k)}\cdot\frac{1}{|C|} \tag{6}$$

其中 $df(t_i)$ 为文档集 C 中出现 t_i 的文档的数目;$df(t_i,t_k)$ 为同时包含 t_i、t_k 的文档的数目。$RDF(t_k,d_j)$ 是与文档 d_j 相关,同时又包含词条 t_k 的文档的数目。$\frac{df(t_i,t_k)}{df(t_i)df(t_k)}$ 是互信息(MI)的估算方法,则

$$\omega_{ij}=P(t_i|t_j)=\frac{1}{|C|}\sum_{k=1}^{n}MI(t_i,t_k)RDF(t_k,d_j) \tag{7}$$

其中 $\frac{1}{|C|}$ 为常数,$MI(t_i,t_k)$ 是词条 t_i 和词条 t_k 间互信息。

利用 $MI\times RDF$ 权重计算方法,计算查询 q 和文档 d_j 的相关度为:

$$rel(q,d_j)=\sum_{t_i\in q\cap TS}\omega_{ij} \tag{8}$$

式(8)是查询向量与文本向量的内积。其中查询向量的词条权重为 1;只要 t_i 出现在其他属于 C 的文档中,在文本向量 $\boldsymbol{d}_j$ 中可有 $\omega_{ij}\neq 0$。因此提出适用于 VSM 的平滑技术:

$$rel(q,d_j)=\prod_{t_i\in q\cap TS}\omega_{ij} \tag{9}$$

式(9)体现了概率模型的思想。$MI\times RDF$ 权重计算利用文档 d_j 的相关文档($RDF(t_k,d_j)$)作中介,同时考虑词条间的相互关系($MI(t_i,t_k)$),能更准确地计算词条 t_i 在文档 d_j 中的权重。权重计算机适合同构文档集,同时也是一种通用的平滑技术。

4 组内信息检索算法设计

组内检索的核心问题是如何有效地选择与查询相关度高的叶节点。与查询相关度高的节点是指拥有较多与查询相关的文档(量),同时节点内的文档与查询的相关度较高(质)的节点。因此应综合考虑量和质两个因素。选择执行查询的叶节点公式为:

$$rel(q,p_j)=\alpha\cdot\exp(rel(q,C_j))+\beta\cdot\lg(|C_j|_d) \tag{10}$$

$rel(q,p_j)$ 是考虑质和量后得到的查询 q 与节点 p_j 的相关度,$rel(q,C_j)$ 反映质,$\lg(|C_j|_d)$ 反映量。α 和 β 为可调系数,其取值决定于 $rel(q,C_j)$ 计算的准确性和文档集主题。

系统中的每个节点都构建并维护自己的节点资源描述(PRD),PRD 包含了节点内文档集中所有词条。对于词条 t_n,利用语言模型($p(t_n|M_{dk})$)可以计算出 t_n 在虚拟文档 C_j 中的权重 ω_n 为:

$$\omega_n=\frac{\sum_{d_k\in C_j}p(t_n|M_{d_k})}{|C_j|_d} \tag{11}$$

$|C_j|_d$ 是节点 p_j 内文档集 C_j 的大小,可将 C_j 看作一个虚拟文档,PRD 是虚拟文档的索引信息。利用 K-L 散度来计算查询 q 与虚拟文档 C_j 的相关度($rel(q,C_j)$):

$$rel(q,C_i)=KL(q,C_i)=\sum_{t_n\in q}p(t_n|q)\left|\log\frac{p(t_n|q)}{p(t_n|C_j)}\right| \tag{12}$$

显然,$rel(q,C_i)$ 越大,p_j 与 q 就越相关。

对于一个拥有 s 篇文档的文档集 C,在 C 中存在于 q 相关的文档的概率为:

$$P\{R(q,d_l)|d_l\in C\}=\left[1-\left(1-\frac{m}{M}\right)^s\right] \tag{13}$$

$R(q,d_l)$ 表示文档 d_l 与查询 q 相关。显然,s 越大,C 包含的文章数越多,C 中存在与查询相关的文档的概率也就越大。

p_i 据全部所在 IG 的叶节点与查询的相关度 $rel(q,p_j)$,选择部分与 q 的相关度较高的叶节点,然后令这些叶节点执行检索任务,并返回查询结果。具体实现时将叶节点按与 q 相关度从大到小排序,然后选择前 δ% 的节点作为真正执行查询任务的节点。

所有与 p_i 在同一 IG 的叶节点都需计算 $rel(q,p_j)$。由式(10)知,$rel(q,p_j)$ 所需计算量小,额外计算开销可忽略不计。由于 p_i 管理的叶节点数较少($\leqslant 60$),排序操作计算量很小,同时发送查询与返回结果所占网络带宽也较少。节点选择的优点是将检索任务限定在少数与查询密切相关的节点,节省了 IG 内节点的计算力资源,同时提高检索结果的查准率。带来的损失

是检索结果的查全率可能会有所降低。但对于普通用户,更关注少量与查询非常相关的文档,考核检索结果的查全率比较困难且意义不大。

5 组间信息检索算法设计

组间检索的任务是如何选择执行查询任务的超节点。超节点层的超节点间是典型的均匀随机网络,利用优化的 Random Walk 搜索算法并结合叶节点选择,提出组间检索策略:

(1)社区内最先接收查询 q 的超节点 p_i 为 q 设定 HTL 值,执行一次查询 HTL 就减 1,若 HTL=0,则 q 不再被转发和执行。然后,p_i 从自己的超节点邻居列表中随机选择超节点转发查询;

(2)接收到转发查询的超节点(p_j)计算自己与查询 q 的相关度公式为:

$$rel(q,p_j)=\frac{\alpha\cdot\exp(rel(q,C_j))+\beta.\lg(|C_j|_d)+\gamma.n}{l+1} \tag{14}$$

C_j 为超节点 p_j 内的文档集;n 为与 p_j 相连的叶节点数目,n 反映了节点 p_j 所在 IG 的规模;l 为 p_j 中任务队列的长度,反映了 p_j 的繁忙程度;α,β,γ 是 3 个可调参数。p_j 将计算得到的 $rel(q,p_j)$ 发送回 p_i;

(3)p_i 在接收到上述邻居超节点返回的相关度的值后,将邻居超节点按与 q 的相关度从大到小排序,然后再选择前 δ_2% 的节点,请求这些节点执行检索任务。执行检索任务的超节点可以将 HTL 减 1 后,进一步转发 q。

每个执行检索任务的超节点最多只选 δ_1%×δ_2%的邻居超节点做查询执行节点。可有效地降低对计算力、网络资源的占用,在减少返回查询结果数量的同时,提高检索的精度。

6 结果融合及实验评估

6.1 结果融合

对特定的 P2P 系统,所有节点都采用相同的文本检索引擎。由于采用了权重和文档查询相关度计算方法,不同节点内文档与查询的相关度具有可比性。采用 CombMNZ 方法进行结果融合[9-10]:

$$GScore(d_j)=n\times\sum_{i=1}^{n}LScore_i(d_j) \tag{15}$$

式中,对于文档 d_j 共有 n 个节点认为其与查询相关,$LScore_i(d_j)$ 为在 d_j 节点 p_i 上与查询的相关度,$GScore(d_j)$ 为最终求得的 d_j 与查询的相关度。基本思想是对于一篇文档,如果许多节点都认为它与查询相关,则该文档很可能与查询相关。

6.2 实验环境组建

实验采用 TREC5 系统,节点根据其主题形成 4 个 P2P 社区。表 1 列出了各社区的主题和社区规模(社区内节点的数目)。在模拟系统的每个社区内,分别执行与该社区主题相关的查询。表 2 列出了实验所涉及的主题以及与主题相关的查询信息。

由于受实验数据集和设备性能的限制,模拟系统的社区规模都不大(76≤|C|≤187),一个社区内只包含 1~4 个 IG。实验中,主要考察节点内检索机制与组内搜索机制的性能。由于社区内 IG 的数目少,组间搜索采用泛洪方式发送查询。评估指标

表 1 各主题对应 P2P 社区规模列表

主题	社区规模
Science & Technology	124
Education	130
Environment	168
Sports	189

表 2 实验中涉及主题及其对应查询

主题	与主题相关的查询
Science & Technology	Internet
	Computer
	Automation
	Information technology
Education	Vocational education
	Secondary education
	HIGher education
Environment	Greenhouse effect
	Water Pollution
	Poaching
	Greenpeace
Sports	NBA
	FIFA

定义为:

$$n\text{-}precision(q)=\frac{n_q}{n} \tag{16}$$

其中,q 为用户查询;n_q 为按与 q 的相关度排序的前 n 个查询结果中与查询相关文档的数目。对于大规模 P2P 系统不能准确地计算查全率,用户往往只关注部分与查询相关度较高的返回结果。实验中,n 的取值为 n=5,10,20,40,60,80,100,150,200。

6.3 实验结果

为了验证该文算法,设计 4 组实验,实验中节点内的文本检索都采用了 $MI\times RDF$ 词条权重计算方法,叶节点选择 α=0.4,β=0.60。实验详细情况见表 3。

第一组实验说明对于很多主题,通过采用节点选择机制,可以在很小的查询开销下得到较好的检索结果。第三组实验对于主题 Education,两种机制的检索性能非常接近,其原因为:α 和 β 参数的取值与社区主题有关,实验中将 α 和 β 都设定为固定值不太合理,如果据社区主题有针对性地调整 α 和 β,应能获得更好的检索结果;另外仅考虑查询与节点资源描述相似度的节点选择机制检索性能已相当好,提高较困难。

总之,通过采用优化技术,分布式搜索引擎以及 P2P 系统中的文本搜索引擎的检索效果可以超过集中式搜索引擎的检索效果。搜索引擎的性能往往与测试用的数据集以及考核指标有关。

7 结论

采用混合型 P2P 网络模型,将社区内检索划分为本地检索、组内搜索和组间搜索。针对三种检索分别设计了提高检索效率的算法。实验结果表明,设计的算法在较小的查询开销下,取得了较好的检索效果。如果节点内有多个主题的文档,可以使节点按照其拥有的主题同时加入多个社区,而在每个社区内部可使用所设计的算法。

(下转 158 页)