文章编号:1001 - 9081(2006)05 - 1164 - 03

基于兴趣挖掘的非结构化 P2P 搜索机制研究与实现

谭义红[1,3],陈治平[2],林亚平[3]
(1. 长沙学院 信息与计算科学系,湖南 长沙 410003;
2. 福建工程学院 计算机与信息科学系,福建 福州 350014;
3. 湖南大学 计算机与通信学院,湖南 长沙 410082)
(yhtan@ccsu.cn)

摘 要:在非结构化 P2P 环境下,搜索机制是影响信息检索的关键因素之一。提出一种基于兴趣索引表的搜索机制,并在此基础上实现非结构化 P2P 信息检索原型系统 Isearch。该机制首先利用向量空间模型将文件内容向量化,然后对向量空间进行聚类,得到节点的兴趣度,再计算节点之间的兴趣相似度,在本地建立兴趣索引表。在搜索时,根据兴趣索引表直接将查询请求转发到有相似兴趣的节点。实验结果表明,该机制既不影响查询结果,又能减少访问节点的数量,提高查询效率。

关键词:P2P; 搜索机制;信息检索;兴趣挖掘
中图分类号: TP31 **文献标识码**:A

Research and implementation on searching mechanism based on interest mining in unstructured P2P systems

TAN Yi-hong[1,3], CHEN Zhi-ping[2], LIN Ya-ping[3]
(1. *Department of Information and Computing Science, Changsha University, Changsha Hunan* 410003, *China;*
2. *Department of Computer and Information Science, Fujian University of Technology, Fuzhou Fujian* 350014, *China;*
3. *School of Computer and Communication, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: In the environment of unstructured P2P, routing scheme is one of the key factors affecting information retrieval. A routing scheme based on interest indexical table was proposed in this paper. Moreover, on the basis of it, a P2P full text information retrieval prototype system Isearch was implemented. Firstly, local file content of a peer was represented with vector space model. After that, vector space was clustered to obtain interest class of this peer. And then, interest similar degree was computed among peers to build interest indexical table locally. When searching, query requests were forwarded to the peers with similar interest directly according to the interest indexical table. Experimental results show that ISearch can not only make good retrieval results, but also reduce the number of query peers and make retrieval more efficient.

Key words: P2P; routing scheme; information retrieval; interest mining

0 引言

近年来,P2P(对等网络)应用非常广泛,其中最流行的应用是非结构化 P2P 环境(unstructured P2P)下的文件共享。主要是由于非结构化 P2P 网络拓扑结构、搜索机制简单容易实现。从早期的 Napster 到 Gnutella、Kazza 等,P2P 文件共享系统在不断的改进和演化,改进的主要目标是:找到更好的搜索机制,使得系统既有好的检索结果,又有很高的搜索效率。然而,现有的非结构化 P2P 系统都不能同时满足这个目标。目前使用比较多的搜索机制有 Flooding[1,2]、Random Walker[3] 和两者混合的模式[4]等。其中 Flooding 采用广播式发送和转发查询请求,查全率和查准率高且具有鲁棒性,但访问节点多,要占用大量网络带宽,且可扩充性差。Random Walker 采用随机选择一条路径发送或转发查询请求,占用网络带宽少,但查询结果差,其随意性强。这两种机制在选择搜索路径时有一个共同的缺陷,即没有采用较好的选择策略。如果在发送或转发查询请求之前,能事先判断哪些节点(peer)更有可能包含符合检索的信息资源,选择路径是有目的的,而不是随机的、盲目的,则必将提高网络的搜索效率。

统计发现,节点存储的信息资源基本上能反映节点的兴趣,同时兴趣相似的节点更有可能存储相似的信息资源。本文基于这个思想,提出了基于兴趣度的搜索机制,并在 Gnutella 的基础上实现非结构化 P2P 信息检索原型系统 Isearch。该机制的主要思想是,根据节点存储的信息资源,挖掘节点的兴趣,然后建立本地兴趣索引表。此表存储有相似兴趣的节点地址,在搜索时,作为选择路径的主要依据。因为该搜索机制所采用的策略是,每次只选择具有相似兴趣的节点进行转发查询请求,所以不但不会影响查询结果,还能提高搜索效率。

1 Isearch 系统搜索模型

Isearch 系统是在 Gnutella 的基础上,添加了辅助搜索决策功能,使得系统简单且可扩充。图 1 是 Gnutella 搜索示意图,节点 A 的查询请求首先直接发送到 A 的所有邻居节点,再

收稿日期:2005 - 11 - 02;**修订日期**:2006 - 01 - 09

作者简介:谭义红(1971 -),男,讲师,博士研究生,主要研究方向:数据挖掘、信息检索; 陈治平(1971 -),副教授,博士,主要研究方向:机器学习、信息检索; 林亚平(1955 -),教授,博士生导师,主要研究方向:计算机网络、机器学习.

通过邻居节点转发到它的邻居节点,直至整个网络或 TTL(time-to-live)为 0。图 2 是 Isearch 搜索示意图,网络拓扑结构同 Gnutella,若网络中节点{A、B、D} 有相似的兴趣,节点{D、E} 有相似的兴趣,节点{B、C} 有相似的兴趣,则节点 A 的查询请求首先发送到节点 B 和 D,然后由 B 转发到 C,由 D 转发到 E,这样依次转发,直至 TTL 为 0。如果与 A 的查询请求相符合的信息资源存储在节点 C 和 D 中,且 TTL = 3,则 Gnutella 需执行 14 步搜索,而 Isearch 只执行 4 步搜索。

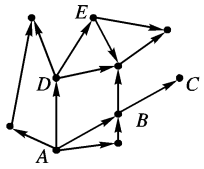

图 1　Gnutella 模型

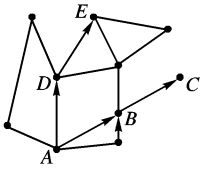

图 2　Isearch 模型

该搜索机制首先利用向量空间模型将文件内容向量化,然后对向量空间进行聚类,得到节点的兴趣度,并将兴趣度向量化,再计算节点之间的兴趣相似度,在本地建立兴趣索引表,用于存储有相似兴趣的节点地址。在搜索时,根据兴趣索引表直接将查询请求转发到有相似兴趣的节点,完成信息检索。以下对其关键技术做详细的探讨。

2　用户兴趣的挖掘及表示

2.1　文本特征提取

文本的特征表示就是对文本进行预处理,抽取代表其特征的元数据,将其以结构化的形式保存,作为文档的中间表示形式。向量空间模型(VSM)[5] 是近年来应用较多的文本特征表示的方法之一。在该模型中,节点的信息资源(文档)用 VSM 来表示,每个文档 d 表示一个范化特征向量 $V(d) = (i_1, w_1(d); \cdots; i_j, w_j(d); \cdots; i_m, w_m(d))$,其中 i_j 为特征词,$w_j(d)$ 为 i_j 在 d 中权值,一般被定义为 i_j 在 d 中出现频率 $tf_j(d)$ 的函数,即 $w_j(d) = \Psi(tf_j(d))$,$\Psi = tf_j(d) \times \log\left[\frac{N}{n_j}\right]$,其中,$N$ 为所有文档的数目,n_j 为含有特征词 i_j 的文档数。

2.2　文本聚类

目前较为广泛使用的文本聚类算法是基于平面划分的 K-means 算法[6]。若给定一文档集 $D = \{d_1, d_2, \cdots, d_n\}$,以及要生成簇的数目 k,算法则将 n 篇文档分为 k 个簇,使得簇内具有较高的相似度,簇间的相似度很低。下面是 K-means 算法的聚类过程:

第 1 步:以某种规则从集合 D 中选择 k 篇文档作为初始簇中心;

第 2 步:将 D 中所有的点分配到最近的簇;

第 3 步:重新计算每个簇的中心;

第 4 步:重复第 2 步、第 3 步直至簇中心不再发生变化。

聚类过程的时间复杂度为 $O(nkt)$,其中 n 是 D 中对文档数,k 是簇的数目,t 是迭代的次数,通常 $k \ll n$ 并且 $t \ll n$,因此能处理大规模文本数据集。

本文采用 K-means 算法作为簇生成算法,来挖掘用户的兴趣。将 VSM 中每个文档向量组合构成文档集 D,取 $k = 3$,聚类后形成 k 簇文档集。因为簇内文档有较高的相似度,具有一些共性,所以我们将每一簇看作一个兴趣,k 簇就是用户的 k 个兴趣。

2.3　兴趣的表示

在每个兴趣中,都包含许多相关的特征词,为了更好地表示兴趣的特征,本文采用特征向量来表示兴趣,$V(c) = (t_1, w_1(c); \cdots; t_j, w_j(c); \cdots; t_m, w_m(c))$,其中 m 为特征词的个数,t_j 表示兴趣 c 中所包含的特征词,$w_j(c)$ 表示 t_j 在兴趣 c 中的权值,计算公式如下:

$$w_j(c) = \frac{(\log(f_j) + 1.0) \times \log(N/n_j)}{\sqrt{\sum_{j=1}^{n}\left[(\log(f_j) + 1.0) \times \log(N/n_j)\right]^2}} \quad (1)$$

其中 f_j 为特征项 t_j 在兴趣 c 的文档中的频率,N 为兴趣 c 中文档集的文档总数,n_j 为兴趣 c 的文档集中包含 t_j 的文档数。

在每一个兴趣中,计算每个特征词在兴趣 c 中的权值($w_j(c)$),然后按权值的大小依次排序,选定前 30 个特征词来表示该兴趣,即 $V(c) = (t_1, w_1(c); \cdots; t_{30}, w_{30}(c))$。

3　兴趣索引表的建立

该模型利用兴趣索引表来存储节点之间的兴趣相似关系,兴趣索引表的结构为 < Interest, Address >,其中 Interest 项表示兴趣,address 表示与该兴趣相似的邻居节点 IP 地址。节点加入到网络的同时,将自己的兴趣特征向量发送到邻居节点,该节点接受到这个信息后,与自己的兴趣 c_j 分别进行相似度计算,计算公式见(2)。

$$sim(c_{ai}, c_{bj}) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\left(\sum_{k=1}^{n} w_{ik}^2\right) \times \left(\sum_{k=1}^{n} w_{jk}^2\right)}} \quad (2)$$

其中 c_{ai} 表示节点 a(刚加入网络的节点) 中第 i 个兴趣,c_{bj} 表示节点 b(a 的邻居节点) 中第 j 个兴趣,n 为特征词的个数,w_{ik} 为特征词 t_k 在 c_{ai} 中的权值,w_{jk} 为特征词 t_k 在 c_{bj} 中的权值。

$sim(c_{ai}, c_{bj})$ 的值越大,说明 c_{ai}, c_{bj} 的相似程度越高,如果该值超过指定阈值 r,则认为节点 a 中第 i 个兴趣与 b 中第 j 个兴趣相似,此时在节点 a、b 的兴趣索引表中分别添加 < c_i, b 的地址 >、< c_j, a 的地址 >,以此方式建立和更新本地兴趣索引表。

4　搜索机制

兴趣索引表建立后,就可以用来作为搜索决策的依据。我们把查询请求看作一个向量,记 $V(q) = (t_1, w_1(q); \cdots; t_j, w_j(q); \cdots; t_m, w_m(q))$,其中 q 为查询请求,t_j 表示 q 中查询词,m 为 q 中查询词的个数,$w_j(q)$ 表示 t_j 在 q 的权值,在此设定 $w_j(q) = 1$,即 q 中所有查询词的权值为 1。

下面通过实例说明搜索机制的主要步骤:节点 A、B、C、D 的兴趣索引表如图 3 所示。现有节点 A 提交一个查询请求 q,A

A(IP:202.121.10.2)

Interest	address
C1	202.121.10.5
C1	202.121.10.6
C2	202.121.11.10
…	…

B(IP:202.121.10.5)

Interest	address
C1	202.121.10.2
C1	202.121.12.3
C2	202.121.11.11
…	…

C(IP:202.121.12.3)

Interest	address
C1	202.121.10.9
C2	202.121.10.5
C3	202.121.13.4

D(IP:202.121.10.6)

Interest	address
C1	202.121.14.3
C2	202.121.13.3
C2	202.121.10.5
…	…

图 3　节点的兴趣索引表

检索是否有符合 q 的信息资源,如果有则返回检索结果,同时将 q 向量与每个兴趣向量按公式(2)进行相似度计算,得出与 q 最相似的兴趣。假设 q 与兴趣 $C1$ 最相似,则查找兴趣索引表,得知与 $C1$ 相似的邻居节点有 B 和 D,于是将 q 发送到 B 和 D,B 和 D 接受查询请求后,完成与 A 类似的操作,依次转发查询请求,直至 TTL 为 0,结束搜索。

5 实验

本模拟实验是在一台 PC 机上完成的,PC 机的配置为 CPU P4 2.0GHz,内存 lGB,操作系统 Windows XP,模拟程序由 Java 编写,网络拓扑结构 Gnutella 由 PLOD[7] 算法产生。网络拓扑为无向图,其中有 1 000 个节点,节点的平均出度为 3.6,因此我们的实验比较接近现实中的 P2P 网络。Isearch 的文档集采用 SMART[8] 系统中所使用的测试文档集 CACM,CACM 包含 3 204 个计算机类文档摘要和 64 个查询,文档集的分布采用均匀分布,即将文档集均匀地分布在网络中的各个节点上。

我们从 Isearch 的检索结果和搜索效率两个方面,分析系统的有效性:

(1) 测试 Isearch 系统的检索结果。检索结果包括查准率(precision)和查全率(recall)。由于检索匹配的机制是一样的,所以查准率肯定是一样的。对于查全率的测试,我们通过设置不同的 TTL,分别统计返回的文档,计算查全率,得出结果如图 4 所示,从图 4 中可以看出,如果设置相同的 TTL,则 Isearch 与 Gnutella 的查全率非常接近,同时,随着 TTL 的值越大,查全率也越高。主要是由于 Isearch 系统在转发查询请求时,只是没有对与兴趣无关的节点进行转发,而这些节点基本上不会保存符合检索的信息资源,所以不会影响查全率。

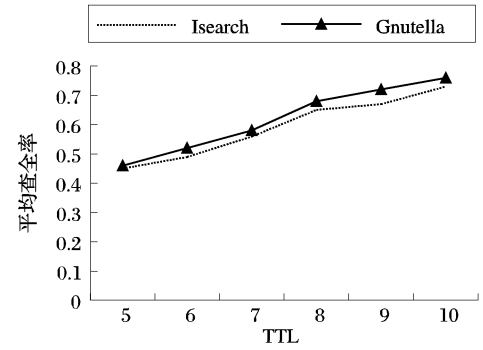

图 4 查全率

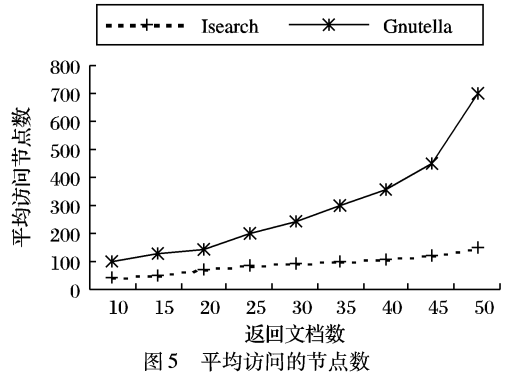

图 5 平均访问的节点数

(2) 测试 Isearch 系统的搜索效率。在搜索过程中,如果减少了访问节点数,但又不影响搜索结果,则会减少网络带宽的占用,提高搜索效率。所以我们用访问节点数量的多少来衡量网络的搜索效率。通过多次提交查询请求,统计访问节点的数量,得到实验结果如图 5 所示。图 5 中显示在返回的文档数量相同的情况下,Isearch 的节点访问量远远低于 Gnutella,随着返回文档数的增加,这种变化越来越明显。主要是由于 Isearch 系统在转发查询请求时,是根据兴趣的相似度,有针对性选择节点进行转发,而不是像 Gnutella 盲目的转发。

6 结语

本文提出了一种基于兴趣挖掘的非结构化 P2P 环境下信息检索搜索机制,主要目标是使信息检索既有好的检索结果,又有很高的搜索效率。由于系统在转发查询请求时,根据兴趣索引表,能事先判断哪些节点更有可能存储符合要求的信息资源,所以减少了访问节点数量,提高查询效率,同时又不会影响查询结果。实验结果表明该系统是有效的。但本文在兴趣的挖掘、兴趣索引表的建立方法上还是相对静态的,如果能根据检索返回结果等因素动态调整和优化,将会进一步提高搜索效率,这是我们下一步要研究的目标。

参考文献:

[1] RIPEANU M . Peer - to - Peer architecture case study : Gnutella network[R]. Technical Report, TR-2001-26, University of Chicago, 2001.

[2] Gnutella website[EB/OL]. http://gnutella.wego.com.

[3] LV Q, CAO P, COHEN E. Search and Replication in Unstructured Peer-to-Peer Networks[A]. International Conference on Supercomputing(ICS'02)[C]. ACM, 2002.

[4] GKANTSIDIS C, MIHAIL M, SABERI A. Random walks in peer-to-peer networks[A]. IEEE INFOCOM[C]. Hong Kong, 2004.

[5] 王继成,潘金贵,张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展. 2000,37(5):513 - 520.

[6] DUDA RO, HART PE. Pattern Classification and Scene Analysis [M]. New York: John Wiley and Sons. 1973.

[7] PALMER CR, STEFFAN JG. Generating network topologies that obey power laws[A]. Proceedings of GLOBECOMM[C], 2000.

[8] BUCKLEY C. Implementation of the SMART information retrieval system[R]. Technical Report, TR35-686, Cornell University, 1985.