

文章编号:1001-9081(2010)01-0114-04

基于 P2P 的个性化 Web 信息检索

付崇国¹, 汤志忠²

(1. 大连东软信息学院 网络中心, 辽宁 大连 116023; 2. 清华大学 计算机科学与技术系, 北京 100084)

(fuchongguo@neusoft.edu.cn)

摘要:为了克服 Web 搜索引擎在可扩展性、协作性和个性化等方面存在的不足,提出了一种基于 Peer-to-Peer 的全分布、协作式、自组织的个性化 Web 信息检索,定义了以查询主题为中心进行主题聚类、数据组织和查询路由的用户协作共享策略,设计了协作生成用户兴趣列表向量、对相似语义查询进行主题聚类和更新、基于查询集建立倒排索引以及基于查询主题进行语义路由等算法和机制,以提供人性化、协作式、个性化的搜索。模拟实验表明,原型系统可以加快查询速度,减轻网络负荷,提高搜索的准确率。

关键词:Web 信息检索;对等网络;个性化;主题;协作过滤

中图分类号:TP393 **文献标志码:**A

Peer-to-peer based personalized Web information retrieval

FU Chong-guo¹, TANG Zhi-zhong²

(1. Network Center, Neusoft Institute of Information, Dalian Liaoning 116023, China;

2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: To overcome the shortcomings of the Web search engines on scalability, collaboration, and personalization, a personalized P2P based Web information retrieval was proposed based on wide distribution, collaboration and self-organization. The strategy of users' collaboration and sharing was defined. That is, user's query topics were used to cluster the queries, to store data and to route queries. Towards the goal of providing more humanized and personalized retrieval by utilizing users' collaboration, some algorithms and mechanisms were designed in respect to building user's favorite list vector collaboratively, clustering the queries to update the user's interest topic by the semantic similarity, structuring the inverted index based on per unit of keyword group, and forwarding the query among peers according to the similarity of topic. The experimental results show that the prototype system can speed up the searching process, reduce the network load and improve the accuracy of the search.

Key words: Web information retrieval; Peer-to-Peer network (P2P); personalization; topic; collaborative filtering

0 引言

P2P^[1] 技术将高度分布的大量普通节点平等互助地组织在一起,在分布性、扩展性、容错性及成本等方面优势明显。特别的,因 Web 数据分布式的特性,P2P 结构有着适合处理 Web 搜索的天然优势,可以克服传统集中式搜索引擎 (Centralized Search Engine, CSE) 的不足。由此,研究人员提出了基于 P2P 的 Web 搜索技术,在 P2P 网络中构建 Web 搜索引擎,提供 Web 搜索服务。目前,此领域的研究刚刚起步并正在成为热点,现有的原型系统尚不多见,主要有 GALANX^[2]、eSearch^[3]、Coopeer^[4]、ODISSEA^[5]、PeerBridge^[6] 等。在它们的拓扑结构中,呈结构化的居多,需要构建超级节点甚至一般节点的全局标示信息,搜索效率高,可扩展性好;呈非机构化者较少,以 Coopeer 为代表,纯粹由普通节点构建,易构性、鲁棒性、维护性好。目前,这些原型系统离实用还有较大距离,其主要原因有如 Li 等人的分析结论^[7]:简单实现的基于 P2P 的 Web 搜索是不可行的,因为 P2P 搜索系统能够提供的资源不足以满足 Web 搜索的巨大负载要求。上述系统都从追求或达到现有集中式引擎的规模、水平出发进行

全局性搜索引擎设计,其中 GALANX 等的结构化设计需要面对边缘网络普通节点资源、动态自治分布式环境及海量数据挖掘处理之间的多维矛盾,解决数据集存放及选择、查询路由、索引切分和网页收集排序等关键性问题,并需要大规模的部署和开销以应付 Web 搜索的巨大规模,这些方面目前难度仍然很大;而在 Coopeer 等的分布性设计中,所有节点完全对等,彼此之间只在一个层面上进行大量的无结构、无组织查询,这对节点有限的资源和性能提出较大挑战,容易出现局部特征,不利于主动扩展和全局收敛。

在前人研究工作的基础上,鉴于目前全局性搜索引擎设计中存在的关键问题和研究现状,本文以先局部性搜索进而自组织全局扩展的研究思路,借鉴 Coopeer 等的分布式设计思想,提出了一种新颖的基于 P2P 的非结构化、主题型、轻量级 Web 信息搜索。其主要的设计策略是:在全分布式、各自独立的 peer 节点搜索引擎之上,增加一层虚拟的主题社区网络,依靠用户主动参与及共享其结构化资源,通过自组织机制,使搜索域从节点间到局部域再向全局域扩展。1) 分布式节点自组织形成“P2P 主题社区”,以轻量级引擎提高局域搜索效率,满足用户搜索需求;2) “P2P 主题社区”以分类资源

收稿日期:2009-07-07;修回日期:2009-08-14。基金项目:国家自然科学基金资助项目(60773149);国家 973 计划项目(2007CB310900);国家 863 计划项目(2008AA01Z108)。

作者简介:付崇国(1967-),男,辽宁大连人,副教授,主要研究方向:分布式计算、计算机网络;汤志忠(1946-),男,江苏海门人,教授,博士生导师,主要研究方向:计算机系统结构、并行编译。

的主题基于语义路由进行相似查找和关联,在局域间进行分布式自组织扩展,分类资源的类别数量和查询搜索域随着关联社区的增多而不断扩展;3)依靠全局范围内广大网民对资源的分类、共享和自组织,最终将使资源共享扩展为全局化,查询搜索从局部域收敛至全局域,Web搜索巨大的规模和负载要求得以分块化解,而不再完全只施压于搜索系统上。

系统的主要特点是:1)个性化,用户间根据查询语义的内容相似性形成主题兴趣组,通过提出的基于向量空间模型(Vector Space Model, VSM)的主题聚类、基于主题的语义路由机制,将查询范围定位在一小部分节点内,实施个性化的主题搜索;2)协作性,用户间共享网页及其评价信息,通过提出的基于加权矢量和的协作过滤机制,根据用户主题兴趣组成员的共同推荐评价计算评分,对网页进行排序;3)人性化,用户间共享搜索经验,扩充查询集,通过提出的基于查询集的数据集表示机制来索引和组织文档,进行快速匹配;4)轻量级,针对普通节点的有限资源特点进行优化设计。算法复用,以简单成熟的向量相似性算法做全局计算;效率优先,通过 RETE 快速匹配减少文档检索计算量;降低开销,传输 URL 信息,离线聚类归并操作,空闲时自配置、自组织。本文研究的意义在于从实用化角度探索 P2P 环境下 Web 搜索策略,充分利用网络中大量分散的普通节点的计算能力及其分类资源,进行协同化共享性搜索,渐进式实现全局搜索目标。

1 系统结构与流程

引擎基于非结构化的 P2P 网络构建,包含协作过滤、主题更新、索引查询、路由转发 4 个模块。协作过滤模块根据推荐节点的共同评价及相关分析获得带权网页推荐列表,主题更新模块对相似查询进行聚类,生成兴趣主题并保持更新;索引查询模块对每个兴趣主题基于查询集建立倒排索引、归并数据集和检索文档;路由转发模块以主题邻居索引进行主题内容或相似用户的查询转发。具体系统结构框架如图 1 所示。

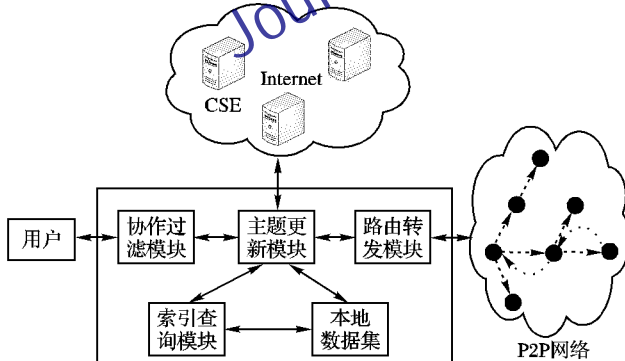


图1 系统体系结构框架

用户可以以单一或混合模式进行外部 Web 引擎搜索或 peers 节点间查询,返回的推荐信息经计算推荐分值后以 URL 列表形式向用户排序推荐,包含网页地址、推荐分值、摘要、推荐者信息等,经用户筛选浏览的兴趣网页被归入兴趣列表。用户可离线进行主题浏览,对相应的兴趣列表进行增、删、调序等操作,并可接受系统的辅助浏览推荐。用户也可以按单个或多个兴趣主题查找相似用户,进行两点或多点讨论组的交流和共享。

节点运行查询时,基于其维护的主题邻居索引进行语义路由,在有限范围内转发查询;节点收到查询后,通过基于查

询集的倒排索引快速匹配检索本地数据集,对查询和文档进行向量相似性分析,返回超过阈值的文档信息,同时查找本地主题邻居索引,选择更近距离的节点转发查询。节点收到查询结果集后,基于 VSM 加权矢量和进行协作过滤计分,按分值排序推荐,下载用户的兴趣网页到本地数据集,并进行主题聚类、扩充查询集、更新倒排索引和主题邻居索引等操作。

2 核心设计

2.1 基于加权矢量和的协作过滤排序

传统协作过滤经常使用基于 Kendall、Pearson 等的相关相似性分析算法,但它一般应用于集中式环境,并会随相关元之间的相关关系波动产生一定的稀疏性。在动态性、自治性的 P2P 环境下,它的稀疏性明显增大,并且会因增加列表等长操作、积差计算等导致普通节点整体计算复杂度偏高。为此,采用向量相似性分析方法,提出了一种基于 VSM 的加权矢量和的用户兴趣列表向量生成算法,结合 VSM 内容过滤与协作推荐来降低稀疏性,利用 VSM 的简单有效并复用其网页向量计算结果以降低节点的计算负荷。

定义 1 兴趣列表 FL 是用户某一查询主题下带有用户兴趣度权重的兴趣网页 URL 的集合,可表示为: $FL = \{\dots, \langle u_i, w_i \rangle, \dots, \langle u_n, w_n \rangle\}$, 其中 u_i 表示网页 i 对应的 URL, w_i 表示用户对 u_i 的兴趣度权重, n 为兴趣列表长度。

定义 2 网页向量 V_i 是网页 i 在 N 维关键字向量空间中的向量表示,即: $V_i = \langle v_{i1}, \dots, v_{iN} \rangle$ 。

定义 3 兴趣列表向量 V_{FLP} 是节点 P 的某一兴趣列表 FL 对应的所有兴趣网页在 N 维关键字向量空间中的加权矢量和的归一化表示,其计算公式为:

$$V_{FLP} = \frac{1}{\rho} \sum_{i=1}^n w_i \cdot V_i \quad (1)$$

$$\rho = \sum_{i=1}^n \sum_{j=1}^N w_i \cdot v_{ij} \quad (2)$$

式中: n 为兴趣列表中全部兴趣网页的数量,归一化处理后 V_{FLP} 各分量的权重和等于 1。用户对 i 的兴趣度权重 w_i 可以简单地与它在用户兴趣列表中的位次序对应,排位越前权重越大,兴趣度越大。设 i 为 u_i 在列表中的位次,则 $w_i = n - i + 1$,排在第一位的网页权重值 $w_{\max} = n$ 。

一个节点运行查询,当结果集返回到本机后,对结果集中的每个网页 URL 计算其推荐分值并按序排列给用户,用户感兴趣者则被列入兴趣列表,同时按其 URL 取回对应的文档并归入本地数据集中。推荐分值计算采用了社区选举的思想,其大小取决于该网页全部推荐者及其相似性的综合情况,计算公式为:

$$R(u_i) = \sum_{P_j \in C(u_i)} [k^{\text{sim}(V_{FLP}, V_{FLP_j})} \times r_{FLP_j}(u_i)] \quad (3)$$

式(3)中 $R(u_i)$ 为网页 i 的推荐分值, $C(u_i)$ 是所有向当前节点 P 推荐网页 i 的远程节点集, $\text{sim}(V_{FLP}, V_{FLP_j})$ 是远程节点 P_j 与当前节点 P 二者间兴趣列表向量的余弦相似度, k 为松弛因子,是一个大于 1 的常量。

$$r_{FLP_j}(u_i) = \frac{w_i}{w_{\max}} = \frac{n - i + 1}{n} \quad (4)$$

式(4)中 $r_{FLP_j}(u_i)$ 是网页 i 在 P_j 兴趣列表中权重大小的相对度量,可以用它在用户兴趣列表中的位次序对应表示。

2.2 查询主题的生成与更新

基于单次查询来存储和组织数据的搜索结构具有空间耗费高和不易扩展的局限。对此提出了一个以用户查询主题为中心进行主题聚类、数据组织和查询路由的策略。

定义 4 查询主题即是用户的兴趣主题,可用六元组表示为:(主题 ID,查询集,兴趣列表,数据集,倒排索引,主题邻居索引)。

定义 5 查询集是一个节点某个查询主题下以每次查询所使用(或解析)的查询关键字集合为元素构成的集合的广义并。

一个查询主题通过一个主题 ID 标识,与一个兴趣列表对应,并通过兴趣列表向量进行运算。兴趣列表中的每个兴趣网页存入本地数据集,基于查询集对数据集建立文档倒排索引,基于远程节点对兴趣网页的推荐分值构建主题邻居索引。

两个相似查询经聚类后生成新的主题。新的查询首先与本地查询集匹配,然后展开搜索。查询结束后,生成的兴趣列表向量与原有兴趣列表向量进行 VSM 余弦相似度计算,当达到阈值后则进行聚类,兴趣列表归并重排后生成新兴趣列表及其向量,根据新兴趣列表下载兴趣文档并更新数据集、计算远程节点的推荐分值并更新主题邻居索引,将查询词与查询集进行归并,更新倒排索引。

网络中的每个节点作为一个主题搜索引擎,根据用户兴趣从 Web 中搜索特定主题相关的信息。主题更新在 Internet 浏览或主题搜索后自动进行,相关的主题元组进行相应调整和更新,具体操作与主题生成的元组操作相同。

基于主题的数据结构还为用户提供有效的辅助浏览、导航和推荐,配合用户自主更新兴趣主题及内容,通过基于主题的相似语义路由支持推荐网络的自配置和自组织。

2.3 基于查询集的数据表示和检索

传统 Web 搜索系统通常基于网页重要性抽取关键字建立倒排索引,但因为机器对自然语言的处理远不及人的语义理解和表达能力,人们的搜索时常被误导。由此,参考分布式搜索系统进行结果缓存的思想^[8],提出了一种基于查询集的倒排索引结构。针对一个查询主题可以有多个相似查询和重复查询,各次查询使用的关键字归入该主题的查询集中。

利用人对搜索目标精细的查询语义描述集和对返回查询结果的肯定(兴趣网页),可以更好地建立查询集与数据集之间的语义关联和数据结构。如表 1 所示,一个查询主题在倒排索引中被表示为同一个主题 ID,对本主题数据集中的每篇文档进行索引。当查询主题更新、查询集和数据集的内容发生变化时,该主题的索引内容同时进行更新。

表 1 基于查询集的倒排索引

Query Set	Doc	Topic ID
computer network course	d1	TID2
computer network course	d2	TID2
computer network storage	d3	TID3
computer network storage	d4	TID3
computer network storage	d5	TID3
parallel computing	d5	TID1
parallel computing	d6	TID1
pervasive computing	d2	TID4
pervasive computing	d7	TID4

进行查询检索时,在匹配到的数据集内对各文档与查询

项进行 VSM 余弦相似性计算,对超过阈值的文档返回其 URL 等信息。若匹配到多个查询集,则按多个主题项分别如上返回各自结果。如查询“computer network”,则会分别按照 TID2 和 TID3 在 d1、d2 和 d3、d4、d5 中检索并返回结果。

检索效率和准确性是轻量型搜索的关键。基于查询集的倒排索引结构设计,实现了查询与查询结果集之间内容更加精准的语义表示以及范围更加集中的快速定位,在时间复杂性和空间复杂性上有了更好的优化。快速定位通过改进的 Rete^[9] 多模式多对象匹配算法实现,将新的查询与已有各查询集模式匹配,根据匹配上的查询集的对应索引定位其数据集,获取查询结果。查询集中各关键字则以哈夫曼树的结构在模式中进行组织,加快匹配速度。

2.4 基于查询主题的语义路由

查询路由是 P2P 搜索的核心环节之一,其算法的效率和开销直接决定了搜索系统的性能。Gnutella 等采用完全分布式的洪泛机制,带宽和 CPU 资源消耗大,扩展性和收敛性难以保证。CAN、Tapestry 等基于 DHT 来构建,需要全局路由和定位保存文档,维护开销大。

针对搜索范围过大和全局管理复杂的问题,提出了一种基于查询主题的语义路由算法,进行定向广度优先搜索。每个节点维护一个查询主题邻居索引,提供了在本地用户某一查询主题上有相似内容的节点,并通过节点间自适应路由和主动配置机制,最终形成一个实际的主题兴趣组。当用户搜索某一兴趣主题相关内容时,则可通过索引进行路由,快速定位目标节点范围,完成个性化搜索。节点在通过阈值控制调整的有限范围内转发查询,以应对节点频繁上下线的异动情况,保证有效查询。接收节点在查询、返回有效文档信息的同时,通过本地主题邻居索引选择相似距离更近的节点继续转发查询。

表 2 主题邻居索引

Topic ID	Query Set	Node ID	Rate
TID1	parallel computing	NID156	9.16
TID1	parallel computing	NID486	8.96
TID2	computer network course	NID429	5.22
TID3	computer network storage	NID486	9.65
TID3	computer network storage	NID175	9.13
TID3	computer network storage	NID233	8.45
TID4	pervasive computing	NID619	7.08
TID4	pervasive computing	NID429	5.57

如表 2 所示,在查询主题邻居索引中,Topic ID 项代表了本地用户的某一查询主题,Node ID 项标明了哪些远程节点存有被本地用户纳入到其兴趣列表中的推荐内容,Rate 项分值越高则意味着此节点提供给本地用户的感兴趣的文档越多,即二者的相似性越大,从而其路由转发的优先度就越高。Rate 项分值以增量方式更新,在用户每次完成相似查询或重复查询后进行,其计算公式为:

$$\Delta R_{P_i} = k^{sim(V_{FLP}, V_{FLP_j})} \quad (5)$$

式(5)中, ΔR_{P_i} 为 Rate 的增量值,每推荐一篇本地用户喜欢的文档则进行一次计算。 k 为大于 1 的常数, V_{FLP} 、 V_{FLP_j} 分别是本地 peer 和远程 peer 的兴趣列表向量, $sim(V_{FLP}, V_{FLP_j})$ 为二者的余弦相似度。

针对全分布式搜索的扩展性和收敛性问题,基于查询主题的语义路由机制设计了一个节点间通过主题集查找具有某些相同主题的节点的算法。

定义6 用户主题集是以某个用户的每个查询主题为元素构成的集合。

定义7 用户局部主题集是用户主题集的一个非空子集。

当节点间传递一个用户局部主题集的查询时,查询节点和转发节点执行如下基本操作:

- 1) 对用户局部主题集与本地用户主题集求交;
- 2) 若交集为空,则选取相近的本地用户局部主题集,从中选取 *Rate* 值较大的若干查询主题邻居节点转发查询,转5);
- 3) 若交集不为空,则取出交集内各主题的查询主题邻居节点集求交;
- 4) 向交集内的所有节点转发查询;
- 5) 重复上述过程,直至目标用户找到或可选节点为空止。

算法实际上是采用了“小世界”的思想^[10],即通过平均6人次的熟人传递就可以把社会中任意两个人联系起来。“小世界”说明了两个现象:①短链效应普遍存在;②人们可以找到短链。也就是当网络呈现某种拓扑结构时,仅利用局部信息就可以有效地找到短链。这使算法的扩展性和收敛性有了一个好的理论基础。另外,算法仅将查询设计在相似用户之间进行,也使搜索具有良好的效率和性能。

上述算法同样可以实现本地节点通过远程查询获取新的主题。当节点查询在主题邻居索引中尚还没有的新主题时,则在本地用户主题集中取出若干与新主题相近的主题,与新主题一起组成用户局部主题集,向选取的若干 *Rate* 值较大的查询主题邻居节点转发,过程同上。

本地节点通过向邻居节点传递本地用户局部主题集查询,可以进行相似用户检索,构成相似用户群,开辟两点或多点讨论组,进行共享和交流,进而对推荐网络进行自配置和自我组织。

3 实验结果与分析

3.1 实验环境

在本文的系统结构设计基础上,实现了一个原型系统并进行模拟实验。实验是在一台配置为 Intel Pentium 2.8 GHz 的处理器、1 GB RAM 和 Windows 2003 Server 操作系统的 PC 机上模拟完成。使用软件模块模拟实际网络,在单机上模拟运行 500 个 peer,物理拓扑网络采用随机生成。实验用的检索数据通过使用网页爬虫从新浪网上下载,生成测试页面集合。为使节点数据具有一定的稠密性,有效进行模拟实验,只选择了其网站分类目录中的球类主题,共下载整理了 5000 个有效网页,随机分布到 P2P 网络中的 500 个 peer 上。对于网页集合中的每个页面,我们对其分词并取 Top-10 单词组,以供生成查询词条。

3.2 实验设计

实验分为两组进行。在与结构化、非结构化系统结构的对比测试中,选择 Chord 和 Gnutella 两种典型结构以及带宽消耗、响应时间和检索效果 3 个指标。另外,为减少计算量,

Chord 模拟中采用对文档和关键字编号而不是实际计算的方式模拟文档和关键字的唯一标识符,这对实验结果没有影响。

在与集中式、分布式搜索引擎的比较实验中,集中式引擎选择 google,分布式搜索引擎模型(non-TopicBased)则以本文系统模型(TopicBased)去除基于主题的逻辑结构进行模拟,相应地调整为协作过滤、索引查询、路由转发三个组成模块。启动实验时,每个 peer 随机选择部分网页构成主题兴趣列表,每个网页的查询集由其 Top-10 中单词组合生成。采用本系统引擎中的混合查询模式操作,即同时在 peer 间和 google 上进行查询。测试指标选择搜索引擎检索结果的平均权重。

3.3 实验结果与分析

本文提出的体系结构设计思路有别于基于有结构和无结构的现有两类搜索模型,以克服有结构体系难以进行语义查询和无结构体系泛泛范围过大等的不足。图2反映出三者之间的可扩展性和性能对比测试。其中第1组指标反映了三者分别对 200 个搜索所返回结果的响应时间,第2组指标反映了三者对应传递查询消息的数量,第3组指标反映在同一检查时刻三者分别累计收到的命中结果数。从中可以看出三种模型的查询速度、查询带宽消耗和查询准确性的比较情况。可见,原型系统不仅减少了查询带宽消耗,加快了响应时间,而且也有效提高了查询准确性。

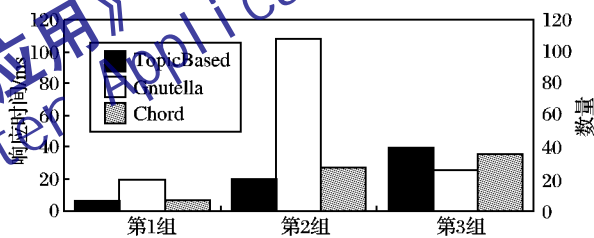


图2 3种系统结构性指标比较

根据搜索引擎用户行为研究结果,90%的用户只点击最初的网页搜索结果。因此,在3种引擎搜索比较实验结果中,本文只对用户兴趣列表取前30条进行统计计算。先对兴趣列表中的网页逆序赋值,第1条为30,第30条递减为1,然后依其搜索获取方式分别归于 google、non-TopicBased、TopicBased 三组并求平均。由图3可见,在搜索次数较少时,google 的搜索结果权重大于 TopicBased;随着搜索次数的加大,TopicBased 的权重逐渐增加,并好于 google。这是因为相似用户社区在形成和扩大,用户间协同共享过滤在逐渐发挥作用,较显著地改善了搜索质量,更加符合个性化要求。而 non-TopicBased 基于纯分布式的单次查询行为进行组织,随着搜索次数的增加,用户维护的兴趣列表增加,彼此间重复性加大,在用户相似主题生成和协作过滤上产生一定程度的抖动,使得系统性能受到影响。

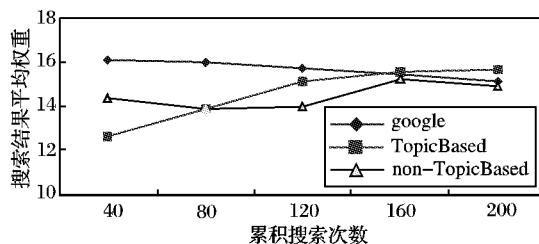


图3 三种模型搜索结果权重比较

```
axis = eyepos * face;
axis = -axis;
axis.Normalize();
}
```

使用 Billboard 技术时,另一个关键是使用透明纹理技术将纹理中的背景色去掉。下面是使用透明纹理技术绘制粒子的部分代码:

```
// 打开混合
glEnable(GL_BLEND);
// 设置混合类型
glBlendFunc(GL_ONE, GL_ONE);
// 设置粒子的纹理
glEnable(GL_TEXTURE_2D);
glBindTexture(GL_TEXTURE_2D, m_Texture);
// 绘制粒子
Rendering();
```

3 化学实验现象模拟的实现

本文算法在普通微机:CPU 为 Pentium4 2.66 GHz,内存 1 GB,Windows 平台,采用 VC++ 和 OpenGL 进行了实验,获得了很好的效果。

图 2 展示采用粒子系统对化学实验中酒精灯火焰和气泡现象模拟的效果图。

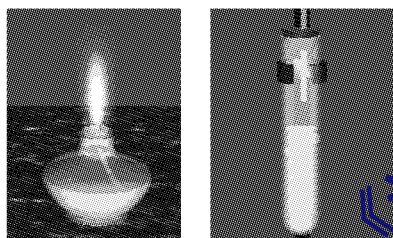


图 2 基于粒子系统的化学实验现象模拟效果

实验表明:该粒子系统所生成的火焰和气泡等化学实验现象不但形态逼真,具有真实的视觉效果,而且每秒至少可生成 30 帧画面,完全能够达到实时绘制的要求。

4 结语

实验现象的模拟是化学实验仿真系统的一个重要组成部分。由于实验现象具有不规则性、动态性和随机性等特点,难以用传统的建模方法进行模拟。本文针对实验现象的特点,提出了基于粒子系统的实验现象的动态模拟方法,并进行了参数化的设计,建立了一个统一的参数化模型,实现了火焰、气泡等实验现象的三维动态模拟。实验表明,该方法有效地提高了实验现象模拟的真实感和绘制效率。

参考文献:

- [1] REEVS W T. Particle system: A technique for modeling a class of fuzzy objects[J]. ACM Transactions on Graphics, 1983, 2(2): 359-376.
- [2] 张芹, 吴慧中, 谢隽毅, 等. 基于粒子系统的火焰模型及其生成方法研究[J]. 计算机辅助设计与图形学学报, 2001, 13(1): 78-82.
- [3] 童若锋, 陈凌钧, 汪国昭. 烟雾的快速模拟[J]. 软件学报, 1999, 10(6): 647-651.
- [4] 赵春霞, 张艳, 战守义. 基于粒子系统方法的三维火焰模拟[J]. 计算机工程与应用, 2004, 40(28): 73-75.
- [5] 赵春霞, 张慧, 郑国勤. 基于粒子系统的喷泉模拟[J]. 计算机应用研究, 2006, 23(1): 244-249.
- [6] 丁强, 陈青林, 左福强. 基于 LOD 的火焰粒子生成技术[J]. 计算机应用, 2005, 25(s1): 257-259.
- [7] 丁纪云, 陈利平, 李思昆. 基于 OpenGL 的烟花动态模拟方法的研究与实现[J]. 计算机工程, 2002, 28(4): 240-275.
- [8] 张芹, 吴慧中, 张建. 基于粒子系统的建模方法研究[J]. 计算机科学, 2003, 30(8): 144-146.

(上接第 117 页)

4 结语

本文在非结构化的 P2P 环境中构建一种个性化、人性化、协作性的轻量级 Web 信息检索,以用户查询主题为中心进行数据组织和结构布局,基于加权矢量和算法进行协作过滤排序,基于 VSM 进行主题聚类更新,基于查询集组织数据和检索,基于主题进行语义路由。通过对原型系统的模拟实验,表明了系统的总体设计能有效保证普通节点的搜索性能,减轻网络负载,提高检索速度和准确性。

参考文献:

- [1] LUA E K, CROWCROFT J, PIAS M, *et al.* A survey and comparison of peer-to-peer overlay network schemes[J]. IEEE Communications Surveys and Tutorials, 2005, 7(2): 72-93.
- [2] WANG Y, GALANIS L, de WITT D J. Galanx: An efficient peer-to-peer search engine system [EB/OL]. [2009-05-06]. <http://pages.cs.wisc.edu/~yuanwang/papers/galanx.pdf>.
- [3] TANG C, XU Z, DWARKADAS S. Peer-to-peer information retrieval using self-organizing semantic overlay networks[C]// Proceedings of the 2003 Conference on Applications, Technologies, Architectures and Protocols for Computer Communications. New York: ACM, 2003: 175-186.
- [4] ZHOU JIN, LI KAI, TANG LI. Coopcer: A peer-to-peer Web search engine towards collaboration, humanization and personalization[EB/OL]. [2009-05-08]. <http://security.riit.tsinghua.edu.cn/share/coopcer.pdf>.
- [5] SUEL T, MATHUR C, WU J W, *et al.* ODISSEA: A peer-to-peer architecture for scalable Web search and information retrieval[EB/OL]. [2009-05-06]. <http://cis.poly.edu/westlab/papers/tr-cis-2003-01.pdf>.
- [6] 傅向华, 明仲. 基于 P2P 的个性化 Web 搜索系统的设计与实现[J]. 计算机工程与应用, 2007, 43(7): 111-113.
- [7] LI JINYANG, LOO BOON THAT, HELLERSTEIN J M, *et al.* On the feasibility of peer-to-peer Web indexing and search[EB/OL]. [2009-05-06]. http://pdos.csail.mit.edu/~rtm/papers/search_feasibility.ps.
- [8] BHATTACHARJEE B, CHAWATHE S, GOPALAKRISHNAN V, *et al.* Efficient peer-to-peer searches using result-caching[EB/OL]. [2009-05-06]. <http://www.cs.umd.edu/~bujor/papers/iptps2003.pdf>.
- [9] FORGY C L. Rete: A fast algorithm for the many pattern/multiple object pattern match problem[J]. Artificial Intelligence, 1982, 19(1): 17-37.
- [10] MILGRAM S. The small world problem[J]. Psychology Today, 1967, 1(1): 60-67.