中 文 信 息 学 报
第 19 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 19 No. 2

文章编号:1003 - 0077(2005)02 - 0044 - 07

利用虚拟站点定位技术的网络信息检索研究①

刘奕群,张 敏,马少平

(清华大学 智能技术与系统国家重点实验室 计算机系,北京 100084)

摘要:虚拟组织是网格体系结构中的基本组织单元,借鉴网格研究中对虚拟组织的特性分析,可以在网络信息检索研究中定义虚拟站点的概念。实验发现,虚拟站点入口页面是网络信息环境中具有较高质量的一个网页集合:实验表明,仅为全部页面数量 21 %的此类页面就涵盖了 70 %以上的超链接,对这个集合进行的内容检索也比对网页全集的检索有超过 60 %的性能提高。这提供了一种在减少索引规模前提下提高网络信息检索性能的解决方案。

关键词:计算机应用;中文信息处理;网络信息检索;非内容特征;虚拟组织

中图分类号:TP391 **文献标识码**:A

Effective Web IR Based on Virtual Site Entry Page Selection

LIU Yi-qun, ZHANG Min, MA Shao-ping

(State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing, 100084, China)

Abstract: Virtual Organization (VO) is a basic concept in grid architecture. Analysis in the link structure of Web pages showed that there exist similar organizations in internet which were called Virtual Sites. Many features of virtual organizations could be founded in virtual sites, especially some non - content features, which were further used to select entry pages of Virtual Sites. This subset of Virtual Site entry pages proved to be qualified both in content and link structure analysis. Although this entry page set contains only about 21 % pages of the whole collection, it covers more than 70 % of its links. Furthermore, information retrieval on this page set makes more than 60 % improvement with respect to that on all pages.

key words: computer application; Chinese information processing; Web information retrieval; non - content feature; virtual organization.

1 引言

网格是一个集成的计算与资源环境,虚拟组织则是指网格体系结构中能够灵活、安全、平等的共享资源的个人或机构的集合体[1,2]。网格环境中的资源管理与发现往往都依赖于虚拟组织,虚拟组织因而成为网格服务的基本单位[3]。虚拟组织由普通结点和服务器结点(VO server)组成,普通结点接受其所属的一个或多个虚拟组织中服务器结点的服务。对大规模实际网页数据的分析发现,网络页面中存在着类似虚拟组织的组织形式,即涉及一个或多个站点的不同网页由于其内容的相似性存在着较为密切的相互引用关系,这些网页通常以入口页面

① **收稿日期**:2004 - 06 - 17
基金项目:国家重点基础研究资助项目(973)(2004CB318108);自然科学基金资助项目(60223004, 60321002, 60303005)
作者简介:刘奕群(1981 —),男,山东济南人,博士研究生,主要研究方向是信息检索、机器学习.

(entry page)或者索引页面[4](hub page , index page)作为中心页面而形成组织,这种组织形式称为虚拟站点。

使用虚拟站点的概念重新理解网络页面的组织形式对网络信息检索的研究有很大的指导意义。虽然搜索引擎的发展已经为人们带来巨大的便利,但是网络信息检索技术还远远无法满足人们快捷高效获取网络信息的要求。根据 Sullivan 等的评测[5],2003 年 2 月时,Google 是世界上索引量最大的搜索引擎(索引到 33 亿 Web 页面)。但即使是按照 How Much Info 计划[6]的保守估计,当时的 Web 仅静态页面数目也将超过 200 亿。容易想象,在未来很长的时间里,随着网络技术在更多国家中得到发展,Web 数据将继续其增长趋势,这对数据收集、索引的建立更新及检索都带来几乎无法逾越的障碍。如何利用有限的存储和计算资源检索到更多高质量的页面,则成为网络信息检索面临的最大挑战之一。

我们称虚拟站点的中心页面为"虚拟站点入口页面",虚拟站点入口页面作为一类内容相似页面的逻辑中心,对人们获取这一类信息有巨大的帮助作用,利用虚拟站点入口页面获取整个网络页面集合的信息对于网络信息检索有明显的实际应用价值。利用网格研究中对虚拟组织结构形式的一些现有成果[1],并结合网络信息检索研究中对页面非内容特征的考察,我们给出了虚拟站点入口页面应有的非内容特征。根据这些特征,以内容无关的方式定位出这些中心页面成为可能。

本文按照如下方式组织:第二部分讨论相关研究工作,详细阐明虚拟组织在网格体系结构中的地位以及虚拟组织的内部组成形式;第三部分指出网络页面中类似虚拟组织的结构样式,并提出虚拟站点的概念;第四部分结合虚拟组织研究,给出虚拟站点入口页面的非内容特性分析,并提出以非内容特征为基础的入口页面判定标准;第五部分介绍虚拟站点入口页面非内容特性及检索性能实验;最后总结并给出主要结论。

2 相关研究工作概述

2.1 网格体系结构中的虚拟组织概念

虚拟组织概念的提出比网格研究有更长的历史,1993 年 2 月 8 日,美国《商业周刊》就提出了虚拟组织的概念,当时的虚拟组织,指的是两个以上的独立的实体,为迅速向市场提供产品和服务,在一定时间内结成的动态联盟。众多大型企业包括福特汽车、西北航空公司等利用虚拟组织,从上个世纪末开始为用户提供一体化的完美服务。网格的出现,一定程度上被认为是在虚拟组织内部以及多个动态的虚拟组织之间共享资源,协同解决问题的需要。而网格所要解决的问题,事实上也就是实现对等的资源共享和解决动态的,分布式的虚拟组织所遇到的问题。这说明虚拟组织不仅是网格体系结构组成的基本单位,更是网格服务的核心环节之一。

在大规模网格基础设施尚未成形的今天,能否将已有的数据规模超过 170TB[6]的 Web 数据环境用于网格基础研究是值得考虑的问题,总的来说,现有的 Web 技术既不适应资源种类的多种多样,也不能提供建立虚拟组织所需要的资源共享的灵活性和可控制性。但如果我们能够在 Web 拓扑结构中找到与虚拟组织结构类似的对应,那至少我们可以对于虚拟组织的管理、定位、信息传递机制有一个大规模的实验参考环境。也能够将现有的一些理论和算法通过实际应用加以改进,而且这些应用对于 Web 本身,可能也是大有裨益的。

2.2 虚拟组织的组成形式

虚拟组织中的每个参与结点都提供一个或多个服务,每个虚拟组织内部提供的服务类型和属性则具有相似性。每个结点掌握的信息包括大量的虚拟组织内部信息与少量的其他虚拟

组织信息。

虚拟组织的结点分为两类:普通结点和服务器结点,服务器结点在虚拟组织形成时就得到确定。不同于普通结点的是,服务器结点要求能够长时间稳定工作,普通结点知道一台或多台所属虚拟组织的服务器结点位置,并接受服务器结点的调度管理。服务器结点承担了比普通结点更多的工作,它不仅负责管理虚拟组织内部的普通结点,还与其他虚拟组织内的服务器结点或普通结点保持联系,能够接受这些结点的服务申请。

3 从虚拟组织到虚拟站点

站点是 Web 页面的基本组织形式,一般来说,站点是一个物理意义上的网页集合,这部分网页集合拥有类似的 URL 域名,存储在同一台或同一组服务器上,并且有较强的相互链接关系。站点的概念对于组织网络页面是必须的,但不可否认的是,在网页内容上讲,站点内部的所有页面并没有一致性,这就给向读者提供信息带来了诸多不便。人们不得不自行或者借助信息检索工具访问站点内的一系列页面,才能找到关于某方面需要的信息,尤其是对于大型综合性的站点更是如此。

为解决这个问题,人们提出了子站点的概念。如http://sports.sina.com.cn 就是新浪网关于体育方面的子站点,读者查找体育方面的信息时,网络信息检索工具只要返回这一个子站点的主页,而用户也只要关注这个子站点内部的页面就可以了。这大大地方便了信息的查找和使用。

然而对于网络信息检索的需要来说,即使是子站点,也是一个过大的粒度。归根到底,人们对于信息查询的需要是无限的,而站点和子站点的数目,则是有限的。可以建立"足球"、"篮球"等一系列的子站点,但又如何保证人们的下一个查询需求能够符合已有的站点/子站点的主题呢?一旦与主题不符,网络信息检索工具返回给用户的,大都就是动辄成百上千的没有组织和逻辑联系的页面了。

根据虚拟组织的定义方式,我们给出虚拟站点的定义:

定义:虚拟站点是 Web 环境中,与某个主题相关的,能够被某个中心页面所链接到的一系列页面的集合。虚拟站点中能够连接到其他页面的中心页面称为虚拟站点的入口页面,其他页面称为普通页面。

构成虚拟站点必须满足两个条件,首先,组成虚拟站点的页面都要与某个主题相关,否则虚拟站点没有组织的必要;其次,虚拟站点必须有一个能够链接到站点内其他页面的中心页面,否则虚拟站点没有组织的可能。

例如美国药物滥用治疗研究所(NIDA)关于大麻滥用方面信息的一系列页面,包括历年滥用大麻情况的统计数据,防止大麻滥用的历次国内会议简报等等,都可以被页面http://www.nida.nih.gov/drugpages/marijuana.html 所连接到,因而以这个入口页面为中心,形成了一个专门提供大麻滥用信息的虚拟站点。如果这些页面涉及的内容没有共同点,或者不存在一个页面能够链接到它们,都无法构成这个虚拟站点。

虚拟站点也是有层次结构的,设计不同主题的虚拟站点之间可能有相互包含的关系,例如涉及足球方面信息的虚拟站点就有可能是涉及体育方面信息虚拟站点的子集。

虚拟站点的定义比一般意义上的站点更加广泛,一般意义上的站点,以这个站点的主页为中心肯定也能够构成一个虚拟站点,但虚拟站点的定义也涵盖了除一般站点之外更多的网页组织在内。如上述关于大麻滥用信息的虚拟站点,就肯定不属于一般站点的范畴。

4 虚拟站点入口页面的判定

4.1 虚拟站点入口页面非内容特性分析

前文已经阐述,入口页面对于虚拟站点的形成至关重要,对于网络信息检索而言,虚拟站点的入口页面同样是举足轻重的,因为它是整个站点信息的入口,在检索结果中提供给用户入口页面,就相当于是用户方便的阅读到虚拟站点中的全部信息。因此讨论虚拟站点入口页面的非内容特性,进而试图利用这些非内容特性找到入口页面是非常有意义的。

根据文献[3]的分析,虚拟组织中的服务器结点能够访问虚拟组织中的各个普通结点,并协调管理服务范围和服务频度。同时,服务器节点又可能与其它虚拟组织中的服务器结点或普通节点有联系。对应到虚拟站点中的链接关系,入口页面应当具有如下两个非内容特性:

特性 1:入口页面能够直接连接到本虚拟站点内部的其他页面。

可以称这个特性为导航特性,即虚拟站点入口页面的读者可以很容易的访问到站点内的其他页面。正如服务器结点可以组织协调各个普通结点一样,入口页面也负责引导对其他页面的阅读。

特性 2:入口页面被大量的虚拟站点外部页面所引用。

这个特性称为接口特性,即对虚拟站点内部网页的大部分访问都是通过对入口页面的访问实现的。这并不是说其他页面完全与站点外页面没有联系,而是说入口页面应该是最重要的访问接口,正如虚拟组织中的一般结点是可以与组织外的结点有联系的,只不过大部分服务请求还是通过本虚拟组织的服务器结点分发的。

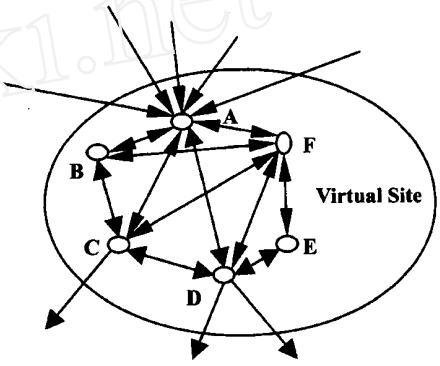

图 1 虚拟站点入口页面非内容特性示意
(页面 F 拥有导航特性但不具有接口特性,
页面 A 是入口页面)

虚拟站点的入口页面应该同时具有这两个特性,单独具有某一特性,成为入口页面的可能性都会大大降低。

图 1 中的页面 F 拥有到本站点所有页面的链接,但由于它没有外部引用的链接,因此并不是一个虚拟站点的入口页面,实际网络环境中,它而更倾向于是站点中的一个"站点地图"页(sitemap page)。页面 A 同时具有两个特性,应该是一个入口页面。

而在图 2 中,页面 A 和 E 都被很多外部页面所引用,但 A 页面没有链接到除了 E 页面之外的其他站点内部页面,因此它只拥有接口特性而不拥有导航特性,从而无法成为此虚拟站点的入口页面。

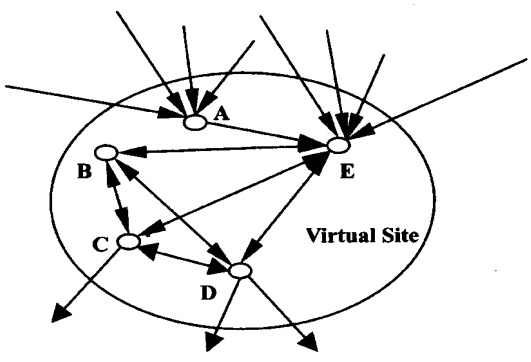

图 2 虚拟站点入口页面非内容特性示意
(页面 A 具有接口特性,但不具有导航特性,
不是入口页面)

4.2 虚拟站点入口页面判定标准

将虚拟站点入口页面的两个非内容特征转

化成为可以量化的标准,就必须引入一些衡量指标。这就需要使用网络信息检索中链接结构分析的一些方法[4,10,11],同时需要引入一些链接结构之外的非内容特征进行考察。这是因为,尽管以 PageRank[7]和 HITS[8]为代表的链接分析的方法在实际应用中取得了不小的成功,但即使是 Google 公司的 Henzinger 等人也在文献[9]中指出,对于判断网络页面质量这样的复杂的任务来讲,仅仅使用链接结构分析是远远不够的。利用虚拟站点入口页面的定义,可以总结出一些有助于进行页面判定的特性(如下文中提到的链接间隔特性)。

评价页面是否满足特性①,可以使用现成的判别标准"页面入度"(in - degree)[10]。入度是反映页面被多少个其他页面所引用的度量。

由图 3 可以看出,网络环境中的绝大部分网页页面入度较小,而特性 2 决定了虚拟站点入口页面的入度较大(根据实际统计,超过 50 %的虚拟站点入口页面入度大于 10)。这说明页面入度可以用于入口页面的判定。

为评价页面满足特性 1 的程度,则需要引入一个新的衡量标准:链接间隔。网页中可能存在一定数目的链接,相邻的两个链接之间用单词数衡量的距离成为链接间隔。页面中所有链接间隔的平均数成为平均链接间隔。根据特性 1,页面中应该具有较多的链接,且这些链接倾向于以索引的形式存在。这就决定了是否具有较小的链接间隔,可以作为衡量页面是否满足特性 1 的标准。根据对手工标注的虚拟站点入口页面的统计,只有 1.74 %的页面链接间隔超过 300 词,而实际网络环境中则有超过 25 %的页面链接间隔超过 300 词,因此把链接间隔作为衡量虚拟站点入口页面的标准是恰当的。

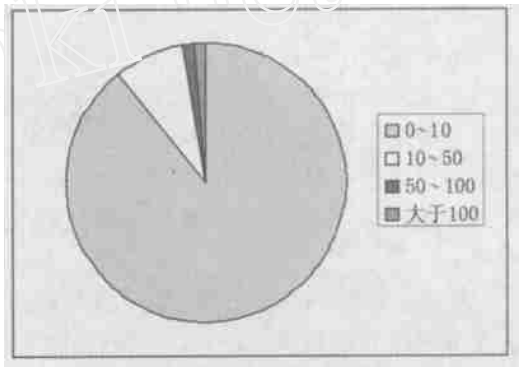

图 3　实际网络环境 1 中页面入度的分布情况

5　实验与结果分析

本节将重点讨论根据非内容特性判定虚拟站点入口页面的实验效果,以及在判定得到的入口页面集合上进行非内容特性统计与检索特性分析的结果。在讨论这些之前,还将介绍本文的实验环境。

5.1　实验环境和方法简述

本文所采用的实验数据均来源于 .GOV 语料库中的页面。利用上面提到的非内容特性,对 .GOV 语料库的所有页面进行筛选,就可以得到虚拟站点入口页面的试验结果集合。

在虚拟站点入口页面的结果集合上,分别进行了页面链接结构的分析和旨在考察信息检索性能的试验。检索实验沿用了 TREC2003② 网络信息检索任务的查询主题及标准答案,此任务一共提供了 50 个查询主题和对应主题的 516 个标准答案。任务的目的就是查找与主题相对应的关键资源页面,查询主题来源于真实网络搜索引擎的用户查询,包含的内容领域涉及

① 基于对语料库 .GOV 的统计,.GOV 是由超过 120 万个网页、近 20G 数据组成的实验性网络语料库,它是由 CSIRO(澳大利亚联邦工业研究组织)在 2002 年初基于真实网络环境抓取到的。可以认为它是一个真实网络环境的采样,详细信息参见 http://es.csiro.au/TRECWeb/govinfo.html

② TREC:文本信息检索会议,即 Text REtrieval Conference,是信息检索领域国际上权威的性能评测专题会议。会议由 NIST(美国国家标准与技术局)和 DARPA(美国国防高级研究计划局)赞助并组织,其举办目的是为了促进文本信息检索技术的研究与发展。详细情况参见 http://trec.nist.gov。

社会政治、经济生活的方方面面,因此具有较高的权威性。

5.2 虚拟站点入口页面的链接结构分析实验

在入口页面定位实验中,选取页面入度 > 10 并且链接间隔 < 300 作为判定虚拟站点入口页面的标准,此外,根据试验效果,还将 URL 特征也作为定位的一个辅助标准。

实验得到的入口页面集合页面占 .GOV 语料库的 21.48 %,对这个集合和整个 .GOV 语料库进行链接关系分析的实验结果如表 1。

表 1 虚拟站点入口页面集合链接分析实验

链接类型	比例
入口页面集合内部链接	37 %
其他页面集合内部链接	9 %
入口页面集合指向其他页面集合	11 %
其他页面集合指向入口页面集合	43 %

实验结果说明,与入口页面集合相关的链接占全部链接总数量的 90 %以上,这与特性 2 的描述,即入口页面是虚拟站点访问的主要接口非常一致。进一步的实验结果显示,入口页面与入口页面直接链接到的页面占全部页面的 72 %以上。这又说明了网络页面中的大部分可以归结到虚拟站点结构中去,即一个页面要么是入口页面,要么是被入口页面直接链接到的虚拟站点的组成页面。

5.3 基于虚拟站点入口页面集合的检索实验结果

虚拟站点概念的提出,一定程度上就是方便网络信息检索的需要。把虚拟站点入口页面作为网络信息检索的结果,既可以避免现有检索中动辄返回上万结果页面的情况,又不会漏掉对用户有用的信息(用户可以通过访问入口页面的链接,得到绝大部分有用信息)。因此虚拟站点入口页面定位的最终结果评价,还要落实到信息检索的效能提高上。实验结果说明,基于入口页面集合的检索效果比全部页面检索的效果有明显的提高,如表 2 所示。

表 2 不同页面集合上的检索效果比较

评价方式	全部页面集合	虚拟站点入口页面集合	TREC2003 最优结果
Precision @10	0.0720	0.1240	0.1240
R - Precision	0.1145	0.1670	0.1636

实验比较了 TREC2003 网络信息检索任务在两个页面集合上的性能,可以看出虚拟站点入口页面集合的检索效果明显好于页面全集。为了方便比较,两组实验都只采用了信息检索中常用的 BM2500 权重计算公式和此公式默认的实验参数。评价方式采用的是 TREC 网络信息检索任务通用的前十位结果平均精度(Precision @10)和 R - 精度(R - precision)。在 Precision @10 评价上,关键资源页面检索比较全部页面集合检索有 72.22 %的提高,而在 R - precision 评价上性能提高的比例是 45.85 %。检索性能的差异可以作如下解释:虚拟站点入口页面集合中用少量的页面集中了大量的高质量信息,在这样的集合里进行检索的难度要远小于在页面全集上进行检索。从另一个角度,也可以认为虚拟站点入口页面定位的过程去除了 Web 信息环境中的大量冗余信息,在一个信息有效性高的页面集合上进行检索的效果自然会好。

为了验证方法的有效性,我们还把这两组结果与 TREC2003 的最优结果[12]进行了比较,实验证明,虚拟站点入口页面集合上的检索效果与 TREC2003 网络信息检索任务的最优结果性能相当,在 R - precision 评价上还优于这个结果。

6 结论与未来工作

网络数据的爆炸性增长与低质量信息的泛滥给网络信息检索技术的发展带来了巨大的挑战,因此将具有大规模信息处理能力的网格技术应用于网络信息检索成为自然而然的想法。本文试图找出网格体系结构与现有的网络信息组织形式之间的联系,对这种联系的考察一方

面可以推动对 Web 页面链接关系的全新认识,从而推动网络信息检索的发展;另一方面也能够在网格基础设施尚未完全建立的情况下,提供一个网格技术的大规模实验环境。

虚拟站点概念的提出,是受到虚拟组织概念与网络信息组织特点的启发,但如下的实验结论,又反映了虚拟站点的定义方式是合理的:

1)虚拟站点入口及其直接链接到的页面占网络页面的 70 %以上,这说明大部分网络页面可以纳入虚拟站点的结构形式中;

2)基于虚拟站点入口页面的检索性能比较网络页面全集的检索有大幅度的提高,这说明虚拟站点入口页面作为虚拟站点内容的代表是可靠有效的。

本文重点考察了虚拟站点概念对于网络信息检索的影响,它对于在索引量一定的条件下提高搜索引擎的信息覆盖率至关重要;同时也为在信息覆盖率一定的情况下减少搜索引擎维护索引的成本提供了一个解决途径。但是,如何在实际应用中,将对虚拟站点入口页面的索引和对普通页面的索引相结合,以保证搜索引擎同时具有较高的查全率与查准率,是需要进一步考察的方向。

参 考 文 献:

[1] I Foster, C Kesselman, S Tuecke. The anatomy of the grid: Enabling scalable virtual organizations [J]. International Journal of Supercomputer Applications, 2001, 15(3):200 - 222.

[2] I Foster, C Kesselman, J M Nick et al. The physiology of the grid: An open grid services architecture for distributed systems integration [A]. Proceedings of Open Grid Service Infrastructure WG, Global Grid Forum [C]. Toronto, Canada, 2002.

[3] Shang Erfan, Du Zhihui. Efficient Grid Service Location Mechanism Based on Virtual Organization and the SMALL - WORLD Theory [J], Journal of Computer Research and Development, 2003, 40(12): 1743 - 1748.

[4] N Craswell, D Hawking and S Robertson. Effective Site Finding using Link Anchor Information [A]. D Kraft, W Croft, D Harper et al. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval [C]. New York, USA: ACM Press, 2001. 250 - 257.

[5] D Sullivan, Search Engine Sizes [R]. Search engine watch website; September 2, 2003; Online at: http://searchenginewatch.com/reports/article.php/2156481

[6] L Peter and H Varian, How Much Information[R], 2003. From http://www.sims.berkeley.edu/how-much-info-2003 on April 2th, 2004.

[7] S. Brin and L. Page. The anatomy of a large - scale hypertextual Web search engine [A]. In Proceedings of the 7th International World Wide Web Conference [C], Brisbane, Australia. Elsevier Science, April 1998, 107 - 117

[8] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment [A]. In Proceedings of the Ninth Annual ACM - SIAM Symposium on Discrete Algorithms [C], San Francisco, California, 25 - 27 January 1998. 668 - 677.

[9] Monika R. Henzinger, Rajeev Motwani and Craig Silverstein, Challenges in Web Search Engines [A], in proceedings of the International Joint Conference on Artificial Intelligence [C], 2003. 1573 - 1579.

[10] N Craswell and D Hawking. Query - independent evidence in home page finding [J]. ACM Transactions on Information Systems, 2003, 21(3): 286 - 313.

[11] R Baeza - Yates, B Ribeiro - Neto. Chapter 13: Searching the web, in Modern Information Retrieval [M], New York, USA: ACM Press, 1999. 367 - 397.

[12] D Hawking and N Craswell. Overview of the TREC 2003 web track [A], November 2003. E. M. Voorhees. Proceedings of the Twelfth Text Retrieval Conference (TREC 2003) [C]. Gaithersburg, Maryland, 2003. 78 - 93.