基于 DOCAI 的多 AGENT 系统在信息发现中的应用

邹若凡* 丁伟 蔡敏
（东南大学计算机科学与工程系，南京 210096）

【摘要】 本文集成分布式对象计算技术和人工智能的研究成果，基于 DOCAI 的体系结构并利用多 Agent 技术构建了一个网络环境中的信息搜集模型。其主要特点是具有灵活的结构，特别适用于分布式搜索引擎和多数据源环境。

【关键词】搜索引擎 信息发现 计算机网络 DOCAI 多 Agent 系统 分布式信息搜集

1．引言：

随着 Internet 的快速发展，网络信息形成了一个分布于全球的信息空间。与传统的媒体信息相比，目前 Internet 上的网络信息更加明显的特征是：分布广，无结构性，变化快和多样化等。 由于传统的搜索引擎采用集中式信息检索技术，越来越不适应这些网络信息发展出现的新特点。因此，在多数据源分布式搜索引擎系统的构筑当中，我们需要解决信息分布的新特性带来的对高效的信息搜集系统的需求。而传统的 WEB 服务中信息访问是基于 CLIENT/SERVER 模型的，需要下载大量信息到本地形成本地数据源供分析、查询使用，需要多个搜集子系统合作。此外，由于信息变化极快，且结构多样化，固定的搜集系统已经无法胜任搜集任务，需要在使用中动态的对系统加以改进或组合。在传统的搜索引擎搜集系统中，由于采用的是静态的集中式实现方式，不仅效率不高，而且可升级性和可组合性均较差。这在实践中已经越来越显露其缺点了。因此有必要以新的研究角度来考虑信息搜集问题，并利用相应的技术思路应用到系统的构筑当中去。

本文从一个全新的角度，利用面向对象的思想方法和分布式人工智能技术来建立一个多 Agent 系统来完成搜索引擎的搜集系统的功能，从而解决实际搜集系统中遇到的、用传统技术无法解决的问题。

2．基于 DOC 技术的分布式人工智能的基本模型

为了利用目前的面向对象技术和分布式人工智能技术，OMG 组织提出了对象管理体系结构（Object Manager Architecture），构建了一个基于 DOC 技术的分布式人工智能的基本模型。

因此，为了解决传统的信息发现技术的一些缺陷，如无法实现分布式程序的可重用性和对于异构数据源的透明性等，我们应用该基于分布式对象计算的 DOCAI 基本模型，其体系结构如图 1 所示。

在该模型中，其底层的支撑环境可以是基于不同网络体系结构和网络协议的异构网络。通过增加一个基于对象请求代理的中间件层，来屏蔽网络协议的具体实现细节，并为分布式程序的设计提供异构环境的透明性。同时，对象请求代理中间件还为上层应用提供了一个分布式的、面向对象的和基于请求代理的分布式应用的开发环境。

---

* 邹若凡，硕士研究生，研究方向为信息发现。

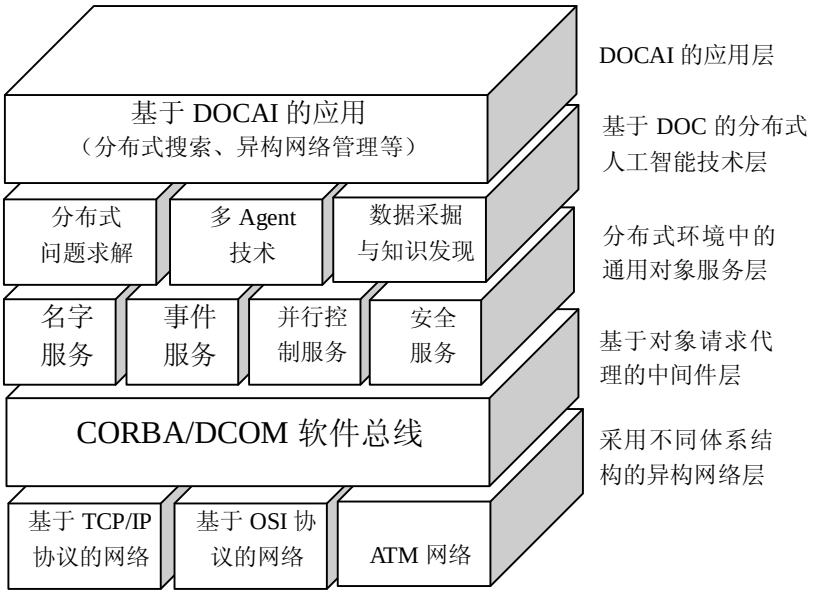

图 1　DOCAI 模型的体系结构图

在分布式的环境中，命名服务，事件服务以及安全服务等在构建应用程序时必不可少的，从而，构成一个通用对象服务层。在该层中，利用中间件层提供的组件可重用性和面向对象的支持能力，使得上层的应用程序可以使用这些服务。

在基于对象请求代理的中间件层和通用对象服务层之上为基于 DOC 技术的分布式人工智能（DOCAI）层。在该层中，对分布式问题求解、多 Agent 技术和数据挖掘和信息发现等具有普遍性的应用问题提供一些标准的支持。此外，我们可以开发一些 DOCAI 在不同领域中的应用，如分布式问题求解的分布式搜索和基于多 Agent 技术的异构网络信息搜集等。

3．信息搜集概述

3.1 目前搜索引擎及信息搜集系统现状

信息检索工作的实质是利用一切可能的方法和手段，在海量的文档信息和其使用者之间建立最有效的连接，使得使用者可以快速查找到所感兴趣的信息。

搜索引擎是网络信息发现技术的一个主要实现手段。同其他信息查询工具一样，搜索引擎把纷繁复杂、形式各异的海量信息按照一定的形式组织起来，通过在数据集上查找与用户需求属性相关的元素，来提供一个有效的途径使得用户可以根据关键词从海量信息库中找到相应信息。

搜索引擎工作的主要思路是根据系统需求激活搜集系统，搜集系统根据相应的索引算法完成对远端数据库上的信息的搜集、索引和定位，并将索引信息返回给搜索引擎，索引信息经过组织存入信息索引库中。查询 Agent 接受用户的查询请求，在数据库中进行相关页面的搜寻，将所搜寻到的页面连接与摘要按匹配程度排序反馈给用户。

其中搜集系统完成的功能是：搜集 WWW 或者其他数据源服务器上的信息以及其他相关目录信息，利用分析模块分析其特征，作为进行索引的本地数据源。根据信息或相关目录信息来确定进一步搜集的路径信息，来继续搜集。

显而易见，搜集系统需要将 Internet 数以亿记的各种结构存放的文档信息下载到本地，然后进行索引，并将处理过的页面丢弃。在搜索引擎运行期间，需要不断更新本地数

据源库来保证结果的正确性。这种搜集必须需要高效的搜集系统来完成。

3.2 搜集系统目前存在的问题：

由于现存的信息服务系统，特别是 WWW 服务，是基于典型的 Client/Server 结构的，所以基于 C/S 结构的检索与搜寻也必然是 C/S 结构的。在搜寻过程中，需要收集大量远端服务器上的信息，在索引之后形成本地数据源。而传统搜集 Agent 由于其实现的集中性，无法适应分布式搜索的要求，低效的、没有协作的搜集系统将会导致本地数据源更新速度的降低以及搜集空间的重复或遗漏。

此外，由于目前 Internet 在体系结构、网络协议、操作系统以及数据源的高度异构性，种类众多。对于一个多数据源搜索引擎来说，一方面各种数据源的搜集和分析应该是动态变化，因为在固定的一段时间内所需要的系统功能的组合是固定的；另外一方面，随着网络信息的动态变化，对新的数据源的支持也应该是动态加入系统的。因此，系统需要动态加载、卸载或更换其功能和模块。而传统的实现方式在适应性、灵活性等方面均存在严重的缺点，一旦出现新的数据源，系统需要重新完成功能的添加，集成和编译，这对于需要连续工作的搜集系统来说是相当不利的。因此我们的目标应该是构筑一个可以灵活组合其功能并支持功能的动态添加而不做大量修改的系统。

4. 一个基于 DOCAI 技术的多 Agent 搜集系统

为了解决以上传统实现方式中难以解决的问题，我们利用 DOCAI 技术实现一个多 Agent 系统来完成搜集功能，并对 DOCAI 技术存在的问题进行了一些探索。对于搜集系统来说，我们把功能自然地分解为控制、搜集、分析和配置等几个部分，把每个功能作为一个自主的子 Agent 系统，各子 Agent 系统之间利用下层的分布式对象服务来支持名字的透明性。从而得到搜集 Agent、分析 Agent、配置 Agent 以及控制 Agent。

- 搜集 Agent 负责信息探询和搜集工作。
- 分析 Agent 对搜集的信息进行分析和理解，形成本地数据源。
- 控制 Agent 负责对搜索问题进行划分成多个搜索集合，并允许分布到不同的主机或进程中去。此外，它还负责子 Agent 系统之间信息交换的控制和解释。
- 配置 Agent 允许管理者动态提供系统的配置参数，来搜集符合条件的信息。

整个系统运行时刻体系结构如图 2 所示。

DOCAI 支持系统层主要利用多 Agent，分布式问题分解以及信息发现中的一些公共技术以及公用信息，完成诸如信息的格式表达，搜索路径空间以及功能的分解等服务。

分布式对象服务主要包括了构筑分布式 Agent 之间传递消息的事件服务以及命名服务。从而提供了一个透明对象引用的环境。从而可以按名访问搜集系统的各子 Agent 系统，并且在需要时，加载、卸载或更换子 Agent。

Interface & NTP 主要是下层通信协议，Agent 最终还是要通过通信协议和信息服务器之间进行交互。这是搜集系统和 Internet 上的信息服务器之间进行通信的接口。

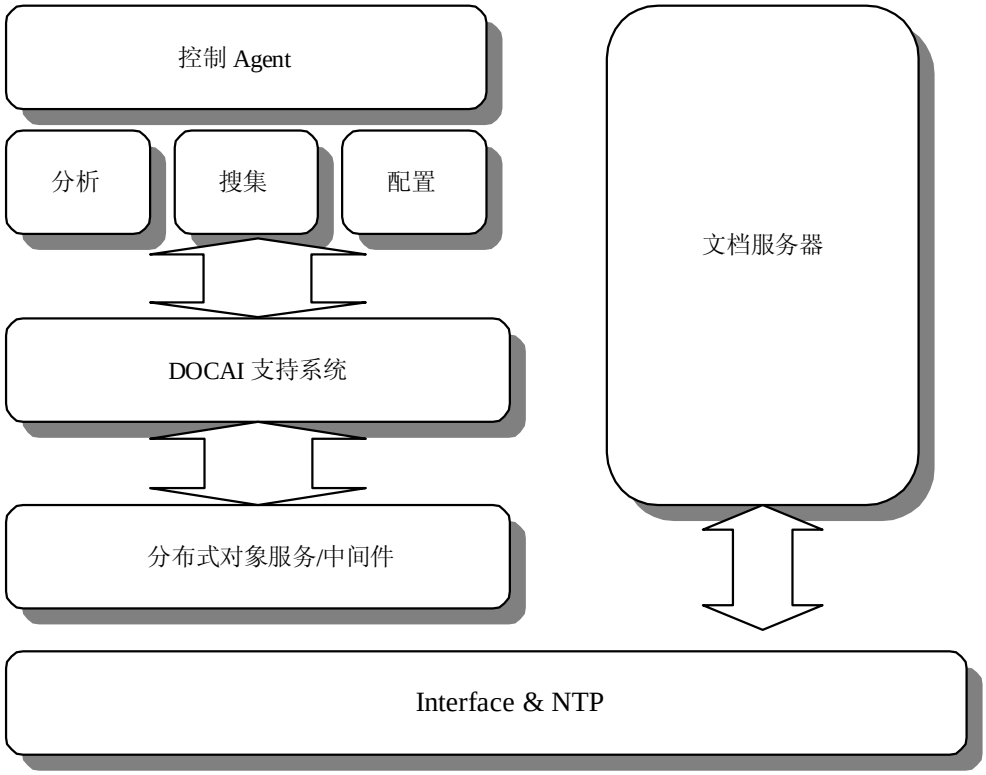

图 2．一个信息发现中搜集系统的结构图

5．基于 DOCAI 的多 Agent 搜集系统的特点及关键技术

与传统的信息搜集技术相比，基于 DOCAI 的多 Agent 搜集系统有明显的优势：

- 通过将搜索问题本地分解，并根据一定的策略分布到不同主机或进程中的 Agent 去，使得各 Agent 可以充分发挥各自系统的能力，有利于负载均衡。
- Agent 直接交互，提高了响应速度。
- 各 Agent 的运行是并行的，提高搜索速度，并增加了系统的稳定性，当个别 Agent 发生故障，不会导致整个系统的崩溃。
- 各 Agent 可以自己实现相关的任务实现、差错控制等功能，使得系统设计变的简单化。
- 由于使用了面向对象的组件技术，从而可以实现软件的即插即用，降低了系统功能动态加载、卸载或更换的开销。

在目前基于 DOCAI 模型的多 Agent 信息搜集系统的实现涉及到计算机网络、分布式计算以及人工智能等多个领域。为了更好的完成信息搜集的任务，实现智能化服务和最优化的任务空间分解的目标，我们需要解决的以下的一些关键技术：

- 系统功能如何分布到多 Agent 上去。优化过后的分解应该有利于 Agent 之间交互的数据信息和控制信息的减少，降低 Agent 之间的耦合程度，提高系统的抗毁性。
- Agent 之间的语言交互应该有利于对多数据源的搜集。对不同的信息格式搜集的 Agent 之间的交互应具有灵活性和可扩展性，并可以利用并且方便下层的 IDL 语言表达。从而使得交互语言具有一定的抽象性和环境无关性。

- Agent 的控制机制的表达。良好的控制机制要求有明确的、无二义性的、独立性较高的控制语言。使得系统不会产生紊乱，而导致信息搜集工作的失败。同时还要保持各独立 Agent 的高度自治性。

6．结束语

本文基于分布式对象计算技术的思想和体系结构，利用一种的基于分布式对象计算技术的分布式人工智能的模型。最后通过一个信息发现中搜集系统例子，给出应用该模型如何解决信息发现中分布式搜索问题。

随着计算机网络和分布式应用的发展，DOCAI 模型在信息发现领域中的应用将越来越广泛，特别是对于多数据源的分布式混合语言的搜索引擎，更加需要利用 DOCAI 的种种技术优点。同时，随着网络信息的进一步丰富，信息发现也将给 DOCAI 模型提出更多的要求，促进 DOCAI 技术的发展。因而，我们应该密切结合 DOCAI 技术和信息发现的技术，来解决信息发现中不断出现的新问题。

参考文献

1. Victor R. Lossor, An Overview of DAI: Viewing Distributed AI as Distributed Search.
2. 张钹，网络时代的人工智能,中国计算机世界技术专题,1997
3. 陶伟，宿利，基于异质网络的软件 Agent 系统,中国计算机世界技术专题，1997
4. Object Management Group, The Common Object Request Broker: Architecture and Specification, Revision 2.0, July 1995　Object Management Group, CORBAServices: Common Object Service Specification, Revised Edition, November, 1996
5. Cai Min，Zou Ruofan， The research and application on the Distributed AI based on the DOC technology，ICYCS’99

An Application of DOCAI-based Multi-Agent System In Information Retrieval

Zou Ruofan, Ding Wei, Cai Min

(Dept. of Computer Science and Engineering, Southeast University, 210096, Nanjing)

【Abstract】 This article integrated the fruit of the technology of Distributed Object Computing and Artificial Intelligence. It constructs a model of Information Retrieval in a network environment, which is based on the architecture of DOCAI and the technology of Multi-agent. This model has a flexible structure. It is especially convenient for the distributed search engine and multiple data source environment.

【Keywords】 search engine , Information Discovery , Computer Network , DOCAI , Multi-Agent system , Distributed Information Retrieval