第32卷第1期 西华大学学报(自然科学版) 2013年1月
Vol. 32, No. 1 Journal of Xihua University • Natural Science Jan. 2013

·计算机软件理论、技术与应用·

多 Agent 主题爬虫协作策略的研究与分析

杜亚军

(西华大学数学与计算机学院,四川 成都 610039)

摘 要: 在多个 Web 主题爬虫并行爬行中,如何避免重复访问网页并高效地获取与主题相关网页,成为搜索引擎主题爬行的热点研究内容之一。为完成系统爬行任务充分发挥每个爬虫自身能力,文章立足于每个爬虫相对独立爬行、共同协作、彼此竞争的思想,将爬虫的历史爬行网页作为背景知识,分析这些网页文本内容,提取网页中"概念"和概念间的语义关系,探讨不同爬虫背景知识之间的语义相似性,提出基于分层概念背景图的爬虫之间理解方法、协作和竞争策略。该策略包括4个方面的内容:主题爬虫背景知识的分层概念背景图的表示模型、基于分层概念背景图的爬虫语义理解方法、在语义理解模型下同组多个网络爬虫之间协作与竞争机制及实现、在语义理解模型下异组多个爬虫之间协作与竞争机制及实现。

关键词: 信息获取; 主题爬虫; 概念背景图; 协作与竞争

中图分类号: TP393.09 **文献标志码:** A **文章编号:** 1673-159X(2013)01-0031-08

doi: 10.3969/j.issn.1673-159X.2013.01.006

Research on Collaborating Strategy among the Multi-agent Focused Crawlers

DU Ya-jun

(*School of Machematics and Computer Science, Xihua University, Chengdu* 610039 *China*)

Abstract: In a focused cralwing system, multi-crawlers crawl parallelly Web and download Web pages. It is one of hotspot researches for a search engine how the diferent focused crawlers avoid to visit the same URLs and download efficiently Web pages related to the search topic. In order to rapidly accomplish the crawling tasks of the system for the specific topic, and embody fully every Web crawler's ability, the author considers that these history visited Web pages (URLs) of every focused crawler reflect their backgroup knowledge. On the basis of cralwing independently, collaborating togather and competing with each other for Web crawlers of the system, the paper proposes the novel understanding, cooperating and competing strategy of concept context grap. It includes four aspects as follows: constructing the mathematical model of backgroun knowledge of every Web crawler based on hierarchy concept context graph, according to the semantic characteristics-concepts of Web pages and their semantic relationships among the concepts; studying the understanding method and model among Web crawlers based on hierarchy concept context graph; studying and implementing the cooprtating, competing model among Web crawlers of the same group managing by a F-Agent; studying and implementing the cooprtating, competing model among Web crawlers of the diferent group managing by F-Agents.

Key words: information retrieval; focused crawler; concept context graph; cooperation and competition

0 背景

网络爬虫的发展经历了3个阶段:独立爬虫→集中爬虫→分布式爬虫。无论是通用搜索引擎爬虫还是主题搜索引擎爬虫,分布式体系结构的并行爬虫都是提高网页获取速度的重要因素。基于 Internet 的分布式爬虫的优点在于其多点接入总带宽较高、对 Internet 负载较小、容易实现就近高效抓取网页、可扩展性强。它已经成为学术界、商业界爬虫系统实现的优选方案[1-3]。目前国内外著名的搜索引擎 Google、AltaVista、Excite、HotBot、Lycos、百度、天网、搜狗等都利用分布管理系统,对网络爬虫进

收稿日期: 2012-09-30

基金项目: 国家自然科学基金(60872089,61271413)

作者简介: 杜亚军(1967-),男,教授,博士,硕士生导师,主要研究方向为网上信息挖掘与搜索引擎、计算机软件开发技术。

行管理,高效地获取 Web 资源。当爬行某些 URLs 时,独立爬虫和集中爬虫只需要与已爬行队列中的 URLs 进行对比分析,就很容易知道哪些网页已经下载,对这部分网页中的 URLs 不再重复进行爬行。Web 上的网页众多,各网页之间存在大量的链接关系,因此不可避免地存在着 2 种 URLs: 不同爬虫的 URLs 队列中有不同 URLs 链接到相同网页; 内容完全相同的网页具有不同的 URLs。在现有的分布式网络爬虫的研究中,由于一个网络爬虫与其他爬虫之间,待爬行 URLs 队列和已爬行 URLs 队列相对独立,各网络爬虫之间缺乏相互沟通与理解,所以这 2 种 URLs 就得不到及时处理,不可避免地会出现某些网络爬虫沿着已爬行过的网页重新爬行,然后在后端才进行网页去重处理,再提供给用户搜索服务。如此重复地下载同样的网页肯定会浪费爬行资源、增加不必要的网络负担和爬行时间; 因此,如何在前端不同的网络爬虫之间进行协调,以尽量减少重复下载网页是分布爬行面临的重要问题之一。

在多爬虫系统中,要避免不同的爬虫去分析、下载相同的页面,以最快的速度返回与搜索主题相关的网页,就需要在搜索过程中: 1) 与其他网络爬虫进行通信与协作,在充分理解自己和其他网络爬虫的爬行目的、爬行网页的范围的基础上,才能有效地避免重复其他网络爬虫的工作,将不属于自己爬行的链接交给其他应该承担该链接的主题爬虫; 2) 当有来自其他网络爬虫的请求帮助时,在自己爬行的网络资源(网络负荷、爬行剩余时间、网络存储容量) 充足的情况下,与其他爬虫竞争,尽可能地帮助其他网络爬虫完成任务和得到更多优质的链接,以使自己获得更多的系统资源奖励,提高爬行性能; 3) 在自己遇到大量的网页无法在爬行系统规定的时间内完成任务时,随时向有能力且与自己爬行主题非常相关的网络爬虫发出帮助请求,以期望能在较短的时间内,获得与主题相关的网页。基于上述 3 个方面,立足于主题搜索引擎中并行工作模式下的多个网络爬虫间语义理解、协作和竞争性研究,充分考虑爬虫的历史爬行背景知识,建立不同网络爬虫之间的语义理解模型、解决多网络爬虫的协作机制和竞争机制,尽可能地在较短的时间内更多地、高效地下载与主题相关的网页,是多 Agent 网络爬虫研究的核心问题之一。但在以往的主题搜索引擎中,对于 Agent 网络爬虫的研究集中在对单个 Agent 网络爬虫智能性的研究与模拟上,无法充分刻画或恰当地模拟 Agent 网络爬虫群体的智能行为。在主题搜索的多 Agent 网络爬虫系统中,要在较短的时间内搜索到满意的结果,就要求各个 Agent 网络爬虫既能相对独立爬行网页,又能相互协作地工作。协作能够提高多个 Agent 网络爬虫系统的整体智能,增强多 Agent 网络爬虫系统解决问题的能力; 竞争能提高网络爬虫的个体智能,增强单个 Agent 网络爬虫更多地获取有价值网页的能力。

随着网络系统日益复杂和庞大,特别是在网页数量爆炸式增长、Internet 信息种类不断增多、人类对网络资源的利用需求日渐增强的情况下,如何提高网络爬虫的协作能力、爬行效率、性能,为用户提供更快更准确的信息是本项目研究的重点。它对网上信息获取与检索、搜索引擎、社会网络[4] 等研究起着重要的理论支撑和促进作用,在基于 Web 的应用系统的开发方面(如主题爬行、专业推荐系统、信息博物馆[5] 、网上自动问答、社会导航系统、网上舆情监测等) 具有重要的实际应用价值; 因此,本项目研究具有重要的理论意义和广阔的实际应用前景。

1 研究现状分析

网络爬虫(Web crawler) ,又名网络蜘蛛(Web spider) 。随着 Internet 的发展,为了提高爬行网页的质量,许多研究者充分考虑网页内容,使用人工智能方法来研究爬虫算法和程序实现。大量的独立爬虫和集中爬虫的研究工作已在文献[6] 进行了阐述。近年来,利用爬虫爬行的历史网页作为知识背景,指导网络爬行的后续爬行工作,成为该领域重要研究方向。其主要表现如下。2000 年 M. Dligenti 等[7] 提出 “背景图”(context graph) 的搜索策略,在该图中,与爬行目标最相关的网页作为目标网页,构成图中的核心结点,然后找出指向目标网页的网页构成图的第 1 级结点,依次形成以网页为结点,网页之间的链接为边的图,这就是背景图。它通过构建典型页面集合(某个主题有关的历史网页集合) 的 Web “背景图”(即真实的网络链接图) 来估计网页离目标页面的距离,距离较近的页面较早得到访问。2006 年 C. C. Hsu 等[8] 提出基于 “相关背景图”(relevancy context graph) 的爬行策略,在网络爬虫爬行过程中,通过相关背景图估算网页和爬行主题之间的距离和相关性,使相关性高的网页被最先爬行。与背景图不同的是,从核心结点逐级向外的链接边(核心结点←1 级结点←…←n 级结点) ,不仅具有回向链接的意义,同时,Hsu 还赋予每级 1 个与主题的相关度(1、α、α^2、…、α^{n-1}) ,形成相关背景图。受文献[7-8] 背景图和相关背景图的

启发,2009 年本研究团队提出了基于领域本体的 Web 语义爬行策略,将网络爬虫的历史爬行网页,通过形式概念分析,提取网页中的概念信息,形成概念格,将包含查询主题的概念作为核心概念,然后通过概念格,找出核心概念的父概念和子概念作为第 1 级概念,再找出第 1 级概念的父概念和子概念作为第 2 级概念,依次,……,找出第 n 级概念形成概念背景图,用来指导网络爬虫继续爬行[9]。

为完成某一主题的爬行任务,相互协作、共同工作,充分发挥爬虫的群体智能是网络爬虫研究的新方向。2002 年 J. Cho 等[10]第一次定义了分布式爬虫的分类方法等一系列基本概念,为分布式爬虫的理论研究、技术实现提供了重要的基础。2002 年叶允明等[11]研究了分布爬虫的体系结构,设计了系统任务高效分割的二级哈希映射算法,对系统的规模动态扩展性进行了研究。2004 年 C. Fabrizio 等[12]利用 Web 社区中网页相对稳定的特点,建立社区坐标,来引导分布爬虫对任务的分解,增强爬行的效率。2005 年 F. Liu 等[13]提出了基于网格平台的分布爬虫,通过为爬虫配置的网格信息服务中心来负责调节系统中每个爬虫的 URLs,使爬虫负荷均衡,爬虫利用 LSI 方法计算每个爬行网页的语义向量,决定是将爬行 URLs 留给自己爬行,还是提交给网格信息服务中心。2007 年 B. B. Cambazoglu 等[14]提出了基于网格的分布爬行系统 SE4SEE,从系统实现的角度,讨论了分布在不同的地理位置的网络爬虫,如何下载个性的、特定需求的网页。2008 年 A. Batzios 等[15]提出了一个生物爬虫(BioCrawler),它模拟生物翅鲨生态环境中个体与团体智能的基本思想,提出利用爬虫历史背景知识的学习策略。2009 年白鹤等[16]提出一种基于数据抽取器的分布式爬虫架构,解决同一爬虫系统内多主题自适应兼容的问题和基于目标导向、负载均衡的 URL 分配问题。2010 年许笑等[1]提出了基于顾问服务的分布式爬虫系统模型和分布式爬虫 Web 划分的概念,研究了分布环境下 Agent 协同算法框架、Web 划分单元选取的方法、Web 划分策略。他们将分布式爬虫系统的实现框架划分为 3 个层次: 1) 逻辑层,将爬行任务切分成多份,交给不同的 Agent 执行; 2) Agent 协同层,不同 Agent 之间相互通信和协同工作; 3) 物理层,使用已有硬件平台和网络资源在物理层上构建分布式爬虫系统。在技术实现上出现了许多代表性的分布爬行系统,如 SE4SEE[13]、Apoidea[16]、北京大学的天网搜索引擎的分布式爬行系统[5]、上海交通大学的 IglooG 分布式爬行系统[11]。文献[1]明确指出这些分布爬虫还缺乏爬虫之间的协作机制,在现有的分布爬虫系统中各个不同网络爬虫之间没有为完成某一目标任务的相互协商、相互帮助的办法。从上世纪 80 年代末开始,协作成为多 Agent 研究的核心问题之一。为了合理分配资源和任务调度,以保证全局一致,多 Agent 之间的协作已经有一些方法: 从组织结构化方面,采用分层组织结的方法; 在协作过程方面,合同式协商协议广泛用于 Agent 之间的任务和资源分配[17]; 在全局协调方面,多 Agent 规划强调避免不一致与冲突情况,要求结点共享和处理大量的信息。随着 MAS 系统理论研究的深入,基于 JADE (Java agent development environment) 的平台被广泛用于 MAS 系统模拟环境的建立和实际应用平台的开发,如半导体制造过程模拟系统[18]、公共卫生检测系统[19]。近年来,在网络信息获取与检索中,MAS 技术也得到了一定的应用。1998 年 H. C. Chen 等[20]研究了智能化、个性化的 Internet 搜索代理方法; 2002 年 W. Kim 等[21]提出基于语义分类的元搜索代理方法; 2004 年 J. P. Lage 等[22]讨论了能自动快速产生网络代理收集器的方法; 2006 年 R. Z. Wang 等[23]提出了一个用于处理 Web 检索事物的多 Agent 系统模型 UAC; 2006 年杨烁颖等[24]在对 MAS 理论调研的基础上,提出一个基于 MAS 的搜索引擎的模型,在基于 Agent 的网络爬虫方面,2007 年本研究团队 D. Xiang 等[25]将多 Agent 的思想引入主题爬虫,提出了多智能 Agent 网络爬虫任务协作模型,在爬行系统的召回率和精度不变的情况下,初步建立了节省网络资源、减少网络负荷、缩短获取网页时间的多 Agent 网络爬虫结构雏形: 将主题爬虫系统中每个网络爬虫当成一个 Agent,根据功能和结构的不同,将 Agent 网络爬虫划分成 2 大类,即 Facilitator - Agent(F - Agent) 和 Crawler - Agent(C - Agent),在此结构中所有 Agent 被分成若干个组,每个组均有一个 F - Agent 和一系列的 C - Agent。图 1 是 3 个组的多 Agent 系统协作模型的组织结构。假设要在网上搜索某一个特定的主题,而这个主题又可由一系列的主题索引词来表示,系统中每个组的 F - Agent 会接收到一个关键词,作为该组任务的引导词,也就是说,同一组中的所有 C - Agent 均有一个相同的引导词,但 F - Agent 为不同的 C - Agent 分配不同的初始 URLs 集合,达到爬行相同引导词的不同主题网页的目的,从而导致了最终

任务执行情况的不同。2009 年 Y. Y. Wang 等[26]借用形式概念分析和本体等初步研究了多网络爬虫爬行过程相互理解的基础。2011 年 Y. Xu 等[27]为了减少系统的频繁通信,对图 1 的结构进行了改进,在 F - Agent 和 C - Agent 之间增加了一个助理 Agent,充当 C - Agents 之间信息交换和通信的作用。2012 年 A. A. Fatemeh 等[28]分析了 Web 上网页的点击流,分别讨论了单个 Agent 的结构和基于用户点击流的多 Agent 网络爬虫的系统结构。

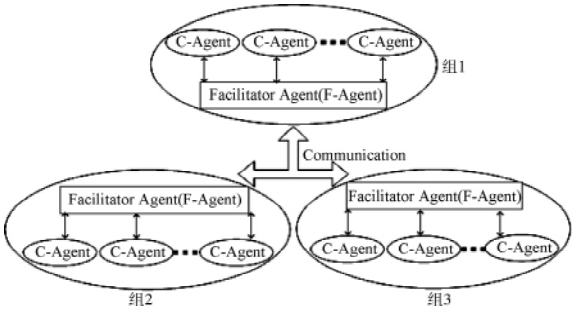

图 1 多 Agent 爬行系统协作模型的组织结构示意图

2 多 Agent 爬虫协作模型的研究内容

借鉴图 1 多 Agent 爬虫系统的结构,建立一个多网络爬虫能相互协作工作的网上爬行系统有许多问题值得研究。充分利用网络爬虫的背景知识,在各爬虫之间相互语义理解的基础上,相互协作与竞争、共同完成爬行任务是项目重点研究的内容,可分为以下 4 个问题。

2.1 主题爬虫背景知识的分层概念背景图表示模型

每一个网络爬虫爬行初期,系统为其分配初始 URLs 集合,随着爬行过程的推移,网络爬虫取回新的网页。初始 URLs 集合对应的网页和不断新取回的网页构成的网页集合,不同程度地反映了爬虫爬行的历史背景知识。值得研究的问题有:1)借助初始 URLs 集合对应的网页,提取反映每个网络爬虫历史背景知识的概念(一个概念包括内涵和外延[29],内涵就是概念的特征词,外延是指对应的 URLs),进而研究概念间的相同、相近,以及 Part of、ISA 等语义关系,建立每个网络爬虫的历史背景的初始分层概念背景图;2)网络爬虫在爬行的过程中,随着新网页的取回,网络爬虫的背景知识更为丰富,研究这些新增网页中概念提取的方法和新提取的概念对分层概念背景图动态更新的方法;3)F - Agent 管理着组内的爬行 C - Agent,组内 C - Agent 的背景知识的合成就是 F - Agent 的背景知识,以 C - Agent 的概念背景图为基础,研究 F - Agent 概念背景图的产生、合并、动态更新的方法。

2.2 基于分层概念背景图的爬虫语义理解方法

以反映 C - Agent 和 F - Agent 背景知识的分层概念背景图为基础,研究网络爬虫之间的理解模型,包括同组网络爬虫 C - Agents 之间的理解、不同组 F - Agents 之间的理解和不同组 C - Agents 之间的理解。从概念语义角度讲,特定主题爬行领域中刻画概念的关键词间也存在相同、相近,以及 Part of、ISA 语义关系,它们反映了 2 个概念内涵之间的相似度大小。2 个概念外延(URLs 集合)之间有相同以及外延中 URLs 的链接关系,反映了 2 个不同概念外延之间的相似度大小。值得研究的问题如下。1)根据这 2 种关系(特定领域中词的语义关系、URLs 相同或 URLs 链接关系),研究组内不同 C - Agent 网络爬虫分层概念背景图中 2 个概念外延和内涵语义相似度;根据一个 C - Agent 网络爬虫分层概念背景图中的一个概念与另一个 C - Agent 分层概念背景图中对应层概念之间的语义最大匹配,进而研究 2 个分层概念背景图之间外延相似度和内涵相似度。2)由于爬虫分层概念背景图反映爬虫的背景知识,不同分层概念背景图的所有概念外延、内涵相似度越大,说明 2 个爬虫之间爬行的主题相似度越大,在主题相似度越大和下载相同网页的可能性越小的情况下,说明 2 个 C - Agent 网络爬虫的理解能力越大;所以根据不同概念背景图的外延、内涵相似度、重复下载网页率,研究建立网络爬虫之间的语义理解策略,进而建立适合主题爬虫 C - Agent 之间的相互理解和沟通程度的度量方法。3)借鉴 C - Agent 之间理解度度量方法,研究 F - Agent 之间和异组 C - Agent 之间理解度的度量方法。

2.3 在语义理解模型下同组多个爬虫之间协作与竞争机制及实现

借鉴 Smith 提出的合同网协议中协商解决问题的思想:招标→投标→中标→合同签定[30-31]。当 C - Agent 待爬行队列的链接超过了 F - Agent 为其限定值时,该 C - Agent 需要与同组其他 C - Agent 协作,组内的 C - Agent 爬虫为获取这个任务,就会产生竞争。在多网络爬虫系统中,为了主题爬虫能高效地获取网页、减少通信开销,增强网络爬虫的智能性,将基于分层概念背景图的理解模型用于网络爬虫爬行过程的协作,提出基于语义理解的同组协作和竞争模型(招标→拍卖与竞标→中标→合同签定与监督)。有以下研究问题。1)研究同组协作与竞争模型、方法及实现。模型示例图 2 可以描述为:C1F1 请求同组其余 C - Agent 协作;根据理解模

型 F1 选中 C2F1 和 C3F1 作为投标 C – Agent 爬虫，F1 进行 URLs 拍卖，C2F1 和 C3F1 参与竞标; 根据理解模型 F1 决策 C2F1 中标; 在 F1 监督下，胜出者 C2F1 与 C1F1 签定合同，F1 监督 C2F1 任务完成情况。2) 根据组内网络爬虫间的语义理解模型，研究任务投标对象(其他网络爬虫) 的选择方法。3) 研究收到招标书的网络爬虫参与竞标且与其他 C – A–gent 竞争获取 URLs 以及 F – Agent 向竞标者拍卖 URLs 的机制与模型。4) 根据语义理解模型，研究组内 F – Agent 决策 C – Agent 中标的模型和对中标网络爬虫的合同签定与管理方法。5) 设计与探索在协作与竞争模型中，系统多个 Agent 均衡工作和通信代价、信息批量交互处理的模型，即在任务招标、发送招标书、参与投标、URLs 拍卖、合同签定、监督合同执行情况等几个方面需要信息的交换，为了减少系统的负载和通信量，当爬虫信息量积累到一定程度而没到系统规定的时间间隔、每隔一定的时间间隔而信息量没积累到一定量时，爬虫之间的信息批量交互处理的模型。6) 研究以上问题的算法和程序实现方法。

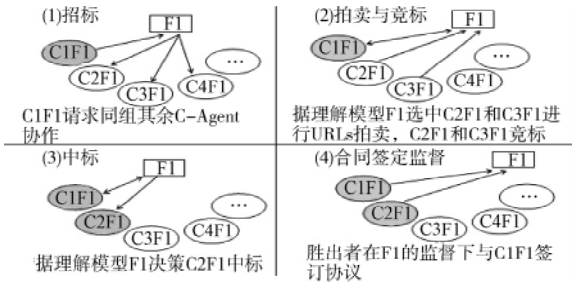

图 2　同组 C – Agents 合作协作过程(招标→拍卖竞标→中标→合同签定与监督)

2.4　在语义理解模型下异组多个爬虫之间协作与竞争机制及实现

当 C – Agent 遇到那些并非自己的目标链接，而可能是其他组 C – Agent 的目标链接时，它会通过本组的 F – Agent 告知另一组的 F – Agent，通知其所属的 C – Agent 承担这些链接。值得研究的问题如下。1) 研究异组 C – Agent 之间的协作与竞争模型、方法。模型示例图 3 可描述为: C1F1 请求在 F2 组的 C – Agent 协作; 根据理解模型，F1 选中 F2，C1F1 通过本组 F1 向 F2 发布请求协作信息，F2 选中 C1F2、C2F2 参与竞标，并向它们拍卖 URLs; C1F1、F1、F2 一起决定中标者 C1F2，在它们的监督下 C1F1 和 C1F2 签定合同。2) 根据异组 F – Agent 间的语义理解模型，研究任务投标对象(其他网络爬虫) 的选择方法。3) 研究收到招标书的异组网络爬虫参与竞标的竞争模型; 根据语义理解模型，研究异组 F – A–gent 对 URLs 拍卖、决定中标 C – Agent 的模型; 研究在双方 F – Agents 的监督下，对中标网络爬虫的合同签定与管理的方法。4) 与同组协商不同的是，异组协商需要跨越 2 个 F – Agent。研究异组 Agent 均衡工作、协商通信和信息批量交互处理模型，共包括 3 层通信: C – Agent→F – Agent 层、F – Agent→F – Agent 层和 F – Agent→C – Agent 层。5) 研究以上问题的算法和程序实现方法。

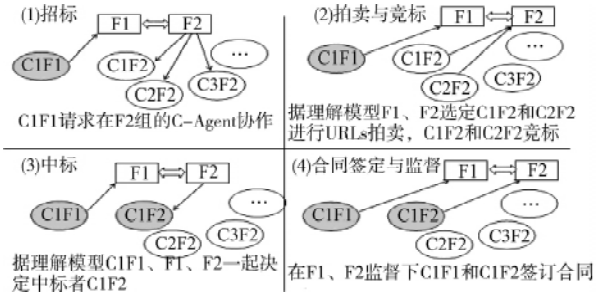

图 3　异组 C – Agents 合作协作过程(招标→拍卖与竞标→中标→合同签定与监督)

3　多 Agent 爬虫协作模型的研究方法

3.1　Agent 网络爬虫的背景知识的分层概念背景图表示模型研究

1) 研究初始 URLs 种子集对 C – Agent 爬虫背景知识的刻画能力和同组不同 C – Agent 爬虫的负责爬行的子空间，让其相对独立，尽可能避免各 C – Agent 爬行区域重复，以使 C – Agent 初始 URLs 种子集能反映各 C – Agent 爬虫的背景知识。

2) 研究 C – Agent 爬虫初始 URLs 种子集对应网页的特征词提取方法，利用形式概念分析理论[32]，提取背景知识中的所有概念，形成背景知识概念集合。

3) 借助 Wordnet 等字典库，提取背景知识概念集合中的有关概念内涵的特征词之间的语义关系，重点梳理各个网页特征词之间的同义、近义以及词与词之间 ISA、Part of 关系。

4) 利用 3) 中词之间的语义关系，通过信息论的方法，对 C – Agent 背景知识中同义概念、近义概念以及概念与概念之间的同义、近义、ISA、Part of 等语义关系进行形式化、数学化的描述，建立 C – Agent 爬虫背景知识中概念的语义关系模型。

5) 将背景知识概念集合划分成核心概念(C – Agent 爬虫的所有主题词和该爬虫主题词的所有同义词的概念，定义为核心概念) 、非核心概念 2 类。

6) 根据 C – Agent 爬虫背景知识中概念的语义关系模型，确定非核心概念和核心概念的语义相似度，充分分析语义相似度大小在整个系统的分布情

况,确立概念语义相似度不同层次,为每个 C – Agent 网络爬虫构建分层概念背景图。在该图中核心概念位于最里层,从里向外,相邻两层概念间的连接线表示概念间的语义相似度大小和具体的语义关系。与背景图和相关背景图中不同的是:图中的结点已经是概念,而不是网页;连接边是语义关系和语义相似度,而不是网页之间的链接。与概念背景图的不同之处在于概念的分层性和相邻两层结点的连接线不再是父和子的关系。

7) 在 C – Agent 网络爬虫的分层概念背景图基础上,研究同组 C – Agent 网络爬虫的分层概念背景图合并成 F – Agent 爬虫分层概念背景图的产生方法。

8) 研究随着 C – Agent 网络爬虫爬行,背景知识中网页增加时,经过 2)、3)、4) 处理,每个 C – Agent 爬虫分层概念背景图中新增概念、更新概念产生的方法,以及这些概念对 C – Agent 爬虫分层概念背景图中已有概念之间的同义、近义、ISA、Part of 关系的语义影响,从而建立 C – Agent 网络爬虫分层概念背景图中概念及概念间的语义关系更新的理论、方法和算法。

9) 通过研究部分 C – Agent 网络爬虫的分层概念背景图的更新,从而研究引起组内 F – Agent 网络爬虫分层概念背景图更新的理论、方法和算法。

10) 研究这一过程的各部分的实现算法,编写程序代码。

3.2 基于分层概念背景图的爬虫语义理解方法研究

1) 以 C – Agent 网络爬虫分层概念背景图为基础,研究不同 C – Agent 网络爬虫分层概念背景图中概念的语义相似度,其中,概念由外延(网页对应的 URLs) 和内涵(网页的特征词) 2 部分组成,重点探讨:(a) 概念外延相似度,主要考虑 2 个概念外延 URLs 相似比例(2 个概念外延相同的情况主要用于理解模型,这里不再考虑)、不同 URLs 所对应网页指向相同链接的比例等因素;(b) 概念内涵相似度,借助 WordNet 等词库和互信息理论提出 2 个概念不同特征词之间的相同、相近、Part of、ISA 等语义关系,计算这些特征词的相似度大小。

2) 研究适合特定主题爬行领域中概念间的相同、相近,以及 Part of、ISA 关系所反映出的概念内涵与外延相似度大小的权重。

3) 综合 1) 和 2),借鉴 A. Formica 概念相似度理论和方法[29] 建立 2 个概念语义相似度计算模型。

4) 根据 C – Agent 网络爬虫分层概念背景图中概念语义相似权重大小,研究一个 C – Agent 网络爬虫的分层背景图与另一个 C – Agent 网络爬虫分层概念背景图对应层概念相似度最大匹配模型。

5) 在 4) 的基础上,研究 2 个 C – Agent 网络爬虫分层概念背景图对应层相似度计算方法、算法,进而建立 2 个 C – Agent 网络爬虫分层概念背景图相似度计算模型、方法。

6) 根据 2 个 C – Agent 网络爬虫分层概念背景图相似度,研究 C – Agent 之间的理解度度量方法,重点考虑 2 个方面的因素:C – Agents 网络爬虫分层概念背景图的相似度和 2 个 C – Agent 网络爬虫爬回的网页对应的 URLs 的重复率。一方面,C – Agents 网络爬虫分层背景图相似度越大,表明 2 个 C – Agent 爬行的主题比较接近。在这种情况下,考虑 C – Agent 已获取 URLs 重复率,重复率越低,说明 2 个 C – Agents 网络爬虫爬行的主题相近且下载网页重复数量越少,2 个 C – Agent 彼此就越理解;重复率越大,说明 2 个 C – Agents 网络爬虫爬行的主题相近且下载重复网页多,2 个 C – Agent 网络爬虫彼此就越不理解。另一方面,2 个 C – Agents 网络爬虫分层概念背景图相似度越低,2 个 C – Agent 之间爬行主题差别就越大,它们之间就不需要相互理解(他们之间就不发生协作),进而建立 2 个 C – Agent 网络爬虫间语义理解模型。

7) 重复 3)、4)、5) 的思想,研究 F – Agent 分层背景图的相似度计算模型。

8) 按 6) 的思想,研究 F – Agent 之间理解度的度量方法。

9) 随着背景知识的动态更新,研究 Agent 爬虫间分层概念背景图的语义相似度动态更新、C – Agent 之间和 F – Agent 之间理解能力动态更新的理论、方法和算法。

10) 编写这一过程各部分算法实现的程序代码。

3.3 在理解模型下同组多个爬虫之间协作与竞争机制及实现研究

1) 研究合理分配给 C – Agent 爬虫的爬行任务的分配模型。重点探讨:(a) 网络信息获取系统接受到用户查询请求,系统根据不同用户区分查询主题,研究查询主题对应的初始 URLs 种子集的范围、优选方案、子主题的划分;(b) 根据系统总体爬行时间要求、CPU 资源、存储资源、网络带宽、网络吞吐能力等,确定整个系统 F – Agent 的数量和各组 C – Agents 的数量;(c) 结合(a) 和(b),研究各 F – Agent 和 C – Agent 承担子主题数量和分配给各 C – Agent

的初始 URLs 种子集的任务分配模型; (d) 研究(a) (b) (c) 的实现算法、程序。

2) 研究 C - Agent 任务预警机制,重点讨论: (a) F - Agent 对组内 C - Agent 的爬行能力(已下载网页与查询主题的相关性、离初始爬行任务的距离、爬行所占用系统的资源、爬行范围内可预见的剩余 URLs 数量) 实时动态跟踪和评价,建立任务预警模型; (b) 各 C - Agent 在爬行过程中定期对自己的爬行能力进行评价,研究 C - Agents 负载过重时主动向 F - Agent 请求同组 C - Agent 协作的模型; (c) 研究(a) 和(b) 模型实现的算法、程序。

3) 研究同组 C - Agent→F - Agent、F - Agent→C - Agent 通信方式,建立 F - Agent 和 C - Agent 信息批量交互处理模型。C - Agent 协作模型中,招标书和投标书发放、合同签定会造成系统通信量。F - Agent 是信息交互中心,C - Agent 一有信息请求,F - Agent 立即向其他 C - Agent 发布信息,会造成系统负担过重(网络资源占用频繁) 。在该模型的信息交互时考虑 2 种因素: F - Agent 积累的信息量和固定的时间间隔。当 F - Agent 标书和合同积累到一定信息量后,再与各 C - Agent 交换信息会减少系统 Agent 之间的对话次数,但在这种情况下 F - Agent 长时间没有积累到一定量的标书和合同,就会造成网络资源的浪费。每隔一定时间内 F - Agent 信息积累过少,长时间网络资源得不到使用,会浪费更多的网络资源; 每隔一定时间内 F - Agent 信息积累过多,会造成瞬间网络负载过大。充分结合信息量和固定的时间间隔,根据系统总体爬行时间要求、CPU 资源、存储资源、网络带宽、网络吞吐能力等建立系统通信模型。

4) 研究 C - Agent 招标书格式(包括在待爬行队列中需要抛出的 URLs) 和招标书自动产生方法、向同组 F - Agent 请求协作的算法和程序实现。

5) 研究 F - Agent 根据招标 C - Agent 请求和同组 C - Agents 之间的语义理解模型,计算招标 C - Agent 对同组其他 C - Agents 网络爬虫的理解度,然后结合本组的 F - Agent 理解度、对组内 C - Agents 爬行能力的评测和系统资源使用状况,确定招标书发放的对象(其他 C - Agents 网络爬虫) ,选定对招标 C - Agent 网络爬虫理解度较大的、爬行能力强的 C - Agents 作为招标对象,建立招标模型。

6) 当组内 C - Agents 收到同组 F - Agent 的标书后,研究 C - Agent 根据自己的爬行能力、对招标 C - Agent 之间的理解度决定是否参与竞标的决策方法、算法和程序实现。

7) 为完成系统总的爬行目标,研究和探讨系统负载平衡。根据系统负载(网络带宽、系统存储能力、CPU 速度等) 情况,探讨 F - Agent 对同组中 C - Agents 抛出的 URLs 价值的评估方法,研究 F - Agent 对不同招标书中 URLs 拍卖的先后顺序。对 Houssein 拍卖模型[33] 研究,分析该拍卖模型的优缺点。根据该研究项目提出的“拍卖价值不同的 URLs 时,F - Agent 总是会将价值最好的 URLs 留到最后”“在不同 C - Agent 竞标时,加大对招标 C - Agent 越理解的 C - Agent 获胜的机会”的思想,研究具体实现的方法、策略和算法。研究在竞标范围内的 C - Agent 为获得系统更多的奖励(C - Agent 获得更多系统资源) ,力求获得拍卖 URLs 而采用的竞标办法和算法。研究在竞争过程中对 Agent 能力值的奖惩办法。

8) 研究 F - Agent 与发标 C - Agent 一起对其他 C - Agent 竞标书评估的方法和算法。研究中标管理模型、方法、算法和程序实现。主要考虑竞标 C - Agent 对发标 C - Agent 理解度(根据同组 C - Agents 网络爬虫之间的语义理解模型,计算竞标 C - Agent 对发标 C - Agent 理解度) 、剩余的爬行资源、剩余爬行时间、对系统资源的占用情况、系统总爬行时间。

9) 研究在 F - Agents 监督下,发标与中标 C - Agent 爬虫双方合同签定方法、算法和程序实现。

10) 研究 F - Agent 对合同执行情况监督方法模型、方法、算法和程序实现。

3.4　在理解模型下异组多个爬虫之间协作与竞争机制的实现研究

1) 当 C - Agent 遇到那些并非本组的目标链接时,由于要与异组 F - Agent 和 C - Agent 通信产生协作请求,所以使用的招标书与同组 C - Agent 标书的具体内容自然不一样,研究这种情况下 C - Agent 的招标书格式和招标书自动产生方法以及 C - Agent 向同组 F - Agent 发出异组协作请求的模型、方法和算法。

2) 研究根据 F - Agents 之间的语义理解模型,计算其他组 F - Agent 与招标 F - Agent 的理解度; 研究根据其他 F - Agent 组内 C - Agent 共同的爬行能力、对系统资源的占用情况、负载能力等有针对性地将招标书发放到异组 F - Agent 的模型和算法。

3) 异组 F - Agent 接到招标书后,借鉴同组 C - Agents 之间的语义理解模型,计算组内 C - Agents 与招标 C - Agent 网络爬虫之间的语义理解度; 探讨异组 F - Agent 根据这种理解度和组内 C - Agents

网络爬虫的爬行能力,有针对性地选择招标书发放到同组 C - Agent 的模型和算法。

4) 当 C - Agent 收到招标书后,研究 C - Agent 根据自己的爬行能力、对招标 C - Agent 之间的理解度和对系统资源消耗情况决定是否参与竞标的决策方法、竞标书的格式和竞标书的自动产生模型、方法、算法。

5) 为完成系统总的爬行目标,研究系统资源(网络带宽、系统存储能力、CPU 速度等)的优化配置模型,研究系统负载平衡;确立异组 F - Agent 对收到的标书中的 URLs 进行拍卖的先后顺序,探讨异组 F - Agent 对 URLs 拍卖的方法、具体实现的数学模型、实现算法(拟借鉴同组 F - Agent 拍卖模型和方法)。研究在竞标范围内的 C - Agent 为获得系统更多的奖励(C - Agent 获得更多系统资源),力求获得拍卖 URLs 而采用的竞标办法、策略和算法,以及在竞争过程对 Agent 能力值的奖惩办法。

6) 研究中标 F - Agent 和发标 F - Agent 一起对各竞标书评估(主要考虑参与竞标 C - Agent 同招标 C - Agent 之间的语义理解能力,竞标 C - Agent 剩余的爬行资源、剩余爬行时间、对系统资源的占用情况,系统总的爬行时间)的方法和算法,建立中标 C - Agent 选择模型。

7) 研究招标 F - Agent 和中标 F - Agent、C - Agent 之间合同签定机制和方法。建立中标 F - Agent 在合同签定过程的监督模型、方法和算法。

8) 中标 F - Agent 监督中标 C - Agent 对合同的执行情况的监督与评价模型、方法和算法。

9) 研究同组 C - Agent→F - Agent、F - Agent→C - Agent 以及异组 F - Agent 之间的 3 层通信模型。由于 C - Agent→F - Agent、F - Agent→C - Agent 之间的竞标、合同发放发生在组内,因此同组批量交互处理通信模型适用于解决异组通信模型中的这类问题。由于异组 F - Agent 之间信息交换,主要产生于招标 F - Agent 在发现组内 C - Agent 有属于异组主题的 URL 协作请求时,因此其异组 F - Agent 通信模型,以招标 F - Agent 为信息交换中心,除了考虑系统信息交换量和固定的间隔时间 2 个因素外,同时还要充分考虑投标 F - Agent 是否重复爬行发标书中 URLs 因素以减少系统 F - Agent 与 F - Agent 通信次数。在这 3 个因素下,根据系统总体爬行时间要求、CPU 资源、存储资源、网络带宽、网络吞吐能力等研究资源优化调配方案和信息批量交互处理模型。

10) 研究上述算法程序的实现。

参 考 文 献

[1] 许笑,张伟哲,张宏莉,等. 广域网分布式 Web 爬虫[J]. 软件学报,2010,21(5):1067 - 1082.

[2] Cho J, Hector G. M. Effective Page Refresh Policies for Web Crawlers[J]. ACM Transactions on Database Systems, 2003,28(4):390 - 426.

[3] 彭涛,孟宇,左万利,等. 主题爬行中的隧道穿越技术[J]. 计算机研究与发展,2010,47(4):628 - 637.

[4] Przemyslaw K, Tomasz K. Label - Dependent Node Classification in the Network[J]. NeuroComputing, 2012, 75(1): 199 - 209.

[5] 李晓明,闫宏飞,王继民. 搜索引擎原理、技术与系统[M]. 北京:科技出版社,2004:23 - 35.

[6] 杜亚军. 网络爬行虫智能化研究分析[J]. 西华大学学报:自然科学版,2010,29(3):217 - 222.

[7] Diligenti M, Coetzee F M, Lawrence S. Focused Crawling Using Context Graphs[C]//The Proceedings of the International Coference on Very Large Database(VLDB). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc,2000: 527 - 534.

[8] Hsu C C, Wu F. Topic-Specific Crawling on the Web with the Measurements of the Relevancy Context Graph[J]. Information System, 2006,31:232 - 246.

[9] Du Y J, Dong Z B. Focused Web Crawling Strategy Based on Concept Context Graph[J]. Journal of Computational Information Systems, 2009, 5(3): 1097 - 1106.

[10] Cho J, Garcia-Molina H. Parallel Crawlers[C]// Proc. of the 11th Int1 Conf. on World Wide Web. New York: ACM Press, 2002: 124 - 135.

[11] 叶允明,于水,马范援,等. 分布式 Web Crawler 的研究:结构、算法和策略[J]. 电子学报,2002,30(12A): 2008 - 2011.

[12] Fabrizio C, Paolo F. Distributed Community Crawling[J]. Thirteenth International World Wide Web Conference Proceedings, 2004: 1094 - 1095.

[13] Liu F, Ma F Y, Ye Y M, et al. IglooG: A Distributed Web Crawler Based on Grid Service[C]//The Proceedings of Asia-Pacific Web Conference. Shanghai: IEEE, 2005: 207 - 216.

[14] Cambazoglu B B, Karaca E, Kucukyilmaz T, et al. Architecture of a Grid Enabled Web Search Engine[J]. Information Processing & Management, 2007, 43(3): 609 - 623.

[15] Batzios A, Dimou C, Symeonidis A L, et al. BioCrawler: An Intelligent Crawler for the Semantic Web[J]. Expert Systems with Applications, 2008, 35(1 - 2): 524 - 530.

[16] 白鹤,汤迪斌,王劲林. 分布式多主题网络爬虫系统的研究与实现[J]. 计算机工程,2009,35(9): 13 - 16, 19.

[17] Smith R G. The Contract Net Protocol: High Level Communication and Control in a Distributed Problem Solver[J]. IEEE Transactions on Computers, 1980, C - 29(12): 1104 - 1113.

[18] Lin J, Long Q Q. Development of a Multi-Agent-Based Distributed Simulation Platform for Semi-conductor Manufacturing[J]. Expert Systems with Applications, 2011, 38(5): 5231 - 5239.

[19] Su C J, Wu C Y. JADE Implemented Mobile Multi-Agent Based, Distributed Information Platform for Pervasive Health Care Monitoring[J]. Applied Soft Computing, 2011, 11(1): 315 - 325.

(下转第 50 页)

成管理的支持。工作流管理系统是一种将信息流和工作流有效结合的通用工作流应用开发系统。

项目群管理只是提出了要求,仅仅指出了“做什么”的问题。“如何做”,还需要实施方法、技术和工具的支持。工作流技术作为实现过程管理与过程控制的一项关键技术,为项目群的实施过程提供了一个从模型定义、执行到管理并监控的完整框架;同时,工作流管理系统通过一套集成化、可互操作的软件工具为这个框架提供了全过程的支持。工作流技术在项目管理系统的应用代表了当前项目管理的一个研究和发展方向。

4 结语

本文将集成管理思想与工作流管理技术相结合应用到工程项目群流程管理实践中,强调项目群子项目多流程协同管理,资源协同配置,促进项目群协同管理环境下项目之间的整合性和效率。在系统论述工程项目群管理的流程、组织、信息的集成特性的基础上,建立了工程项目群集成管理框架模型,并提出了实现工程项目群集成管理“1 + 1 > 2”的方法与途径:基于工作流技术,定义项目群工作流模型及资源模型,实现流程集成;构建项目群工作流管理系统,为信息集成提供平台,促进组织集成,保障流程的集成,最终实现工程项目群的集成管理。

参 考 文 献

[1] 程铁信,霍吉栋,刘源张. 项目管理发展评述[J]. 管理评论,2004,16(2):58,59 –62.

[2] 杜兰英,石永东,杨春方. 基于项目群视角的战略管理层次观[J]. 科技进步与对策,2007,24(12):109 –111.

[3] 尹贻林,刘艳辉. 基于项目群治理框架的大型建设项目集成管理模式研究[J]. 软科学,2009,23(8):20 –25.

[4] 李海凌,刘克剑. 建设工程项目群管理模型构建研究[J]. 西华大学学报:自然科学版,2011,30(2):88 –91.

[5] 傅道春. 建筑业企业项目群管理模式研究[D]. 上海:同济大学,2006.

[6] 黄慧. 建设项目过程集成管理理论与方法研究[D]. 武汉:武汉理工大学,2006.

[7] 叶萍. 基于信息技术下的组织变革[J]. 企业经济,2005(8):21 –23.

[8] 何清华,卢勇,何伟华. 基于 Internet 的大型工程项目信息系统[J]. 同济大学学报:自然科学版,2002,30(2):238 –242.

[9] Workflow Management Coalition(WfMC). Workflow management coalition specification: technology and glossary [R]. Brussels: WfMC –TC –1011, 1996.

[10] 张劲文. 大型交通建设项目管理集成研究[D]. 长沙:中南大学,2005.

(编校:叶超)

(上接第 38 页)

[20] Chen H C, Chung Y M, Ramsey M S, et al. An Intelligent Personal Spider(Agent) for Dynamic Internet/Intranet Searching[J]. Decision Support Systems, 1998, 23(1): 41 –58.

[21] Kim W, Kerschberg L, Scime A. Learning for Automatic Personalization in a Semantic Taxonomy-Based Meta-Search Agent[J]. Electronic Commerce Research and Applications, 2002, 1(2): 150 –173.

[22] Lage J P, Silva A S D, Golgher P B, et al. Automatic Generation of Agents for Collecting Hidden Web Pages for Data Extraction[J]. Data & Knowledge Engineering, 2004, 49(2): 177 –196.

[23] Wang R Z, Xu X L, Huang H P. Intelligent Agent & Application in Web[M]. Beijing: University of Posts and Telecommunications Press, 2006: 95 –101.

[24] 杨烁颖,白万民. 一个基于 MAS 的搜索引擎模型[J]. 计算机技术与发展,2006,16(12): 195 –198.

[25] Xiang D, Du Y J. Coordination and Communication among Topic Specific Search Agents[J]. Proceedings of the Third International Conference on Natural Computation, 2007, 4: 703 –707.

[26] Wang Y Y, Du Y J, Chen S M. The Understanding between Two Agent Crawlers Based on Domain Ontology[C]//The Proceeding of Computational Intelligence and Natural Computing. Wuhan: IEEE, 2009: 47 –50.

[27] Xu Y, Du Y J, Zhang P Y. Improving the Performance of Focused Crawlers Based on Multi-Agent System[J]. Journal of Computational Information Systems, 2011, 7(12): 4375 –4382.

[28] Fatemeh A A, Ali S. An Architecture for a Focused Trend Parallel Web Crawler with the Application of Clickstrearn Analysis[J]. Information Sciences, 2012, 184(1): 266 –281.

[29] Formica A. Concept Similarity in Fuzzy Formal Concept Analysis for Semantic Web[J]. International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems, 2010, 18(2): 153 –167.

[30] Smith R G. The Contract Net Protocol: High Level Communication and Control in a Distributed Problem Solver[J]. IEEE Transactions on Computers, 1980, C –29(12): 1104 –1113.

[31] Jégou D, Kim D W, Baptiste P, et al. A Contract Net based Intelligent Agent System for Solving the Reactive Hoist Scheduling Problem[J]. Expert Systems with Applications, 2006, 30(2): 156 –167.

[32] Wille R. Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts[M]. Ordered Sets, Dordrecht, Reidel: [s. n.], 1982: 445 –470.

[33] Houssein B, Brahim C, Peter K. Multi-Item Auctions for Automatic Negotiation[J]. Information and Software Technology, 2002, 44(5): 291 –301.

(编校:饶莉)