Computing and Informatics, Vol. 35, 2016, 1050–1078

A NOVEL COOPERATION AND COMPETITION STRATEGY AMONG MULTI-AGENT CRAWLERS

Yajun Du, Yong Xu, Min Wang

School of Mathematic and Computer Science
Xihua University
Chengdu 610039
China
e-mail: duyajun@mail.xhu.edu.cn

Abstract. Multi-Agent theory which is used for communication and collaboration among focused crawlers has been proved that it can improve the precision of returned result significantly. In this paper, we proposed a new organizational structure of multi-agent for focused crawlers, in which the agents were divided into three categories, namely F-Agent (Facilitator-Agent), As-Agent (Assistance-Agent) and C-Agent (Crawler-Agent). They worked on their own responsibilities and cooperated mutually to complete a common task of web crawling. In our proposed architecture of focused crawlers based on multi-agent system, we emphasized discussing the collaborative process among multiple agents. To control the cooperation among agents, we proposed a negotiation protocol based on the contract net protocol and achieved the collaboration model of focused crawlers based on multi-agent by JADE. At last, the comparative experiment results showed that our focused crawlers had higher precision and efficiency than other crawlers using the algorithms with breadth-first, best-first, etc.

Keywords: Multi-agent, focused crawler, collaboration, contract net protocol, JADE

1 INTRODUCTION

As web pages grow exponentially, general search engines encounter some unprecedented challenges. The results returned by the general search engine are too comprehensive to meet the personalized needs of users, and it directly gives birth to

the focused crawler [1, 2, 3]. The focused crawler only crawls the on-topic web pages, and avoids a large number of off-topic web pages. On one hand it spends less time and smaller storage space in crawling on the web, on the other hand it can be sufficient to meet the personalized needs of the users [4]. Such as Fish-search [5], Shark-search [6], Breadth-first crawler [7], Best-first crawler [8] are some classic focused crawlers. In addition to these focused crawlers, some artificial intelligence technologies are successfully adopted to the focused crawlers, they make the focused crawler more and more clever, such as HMM (Hidden Markov Model) crawler [1], Q-learning crawler [9], CCG (Concept context graph) crawler [10]. Focused crawlers mentioned above had the ability to learn, and meet the users' needs better. However, they work independently and no communication and collaboration among them exists, focused crawlers have to face two problems:

- Different web pages may have the same hyper links, if focused crawlers do not communicate with other focused crawlers, it will crawl the area that another crawler has already crawled, so the web pages are loaded down in overlap part. As shown in Figure 1 a), focused crawler A and B start crawling the web from the common URL seeds. They are responsible for processing the different web pages respectively. For example, the crawler A crawled along web pages a $\rightarrow$ c $\rightarrow$ $\cdots$, the crawler B crawled along web pages b $\rightarrow$ e $\rightarrow$ $\cdots$. However, the web page a and page b have the same hyper link that points to web page d. And so, these two focused crawlers may crawl the same area of web, and they overlap to load down web pages d, f, e, etc. If the focused crawler A and B can communicate with each other, the web page d will be crawled only by one focused crawler.

- When the task of a focused crawler is too heavy or too light, it is not able ask other focused crawlers to help, this will make focused crawlers unable to discover high-quality pages earlier, and directly lead to the low crawling efficiency and precision. As shown in Figure 1 b), focused crawler A has a large number of on-topic hyper links, but focused crawler B has less of on-topic hyper links. If they can communicate with each other, focused crawler A can tell its status to focused crawler B, and gives its own links to focused crawler B, and focused crawler B can help focused crawler A to crawl the part of web pages. They will discover more high-quality web pages and crawling efficiency and the precision will be also improved.

In order to address these problems above, a collaboration model of focused crawlers based on multi-agent system is proposed. It not only minimizes the overlap area among the focused crawlers, but also achieves the collaboration among focused crawlers to improve the crawling precision and the efficiency.

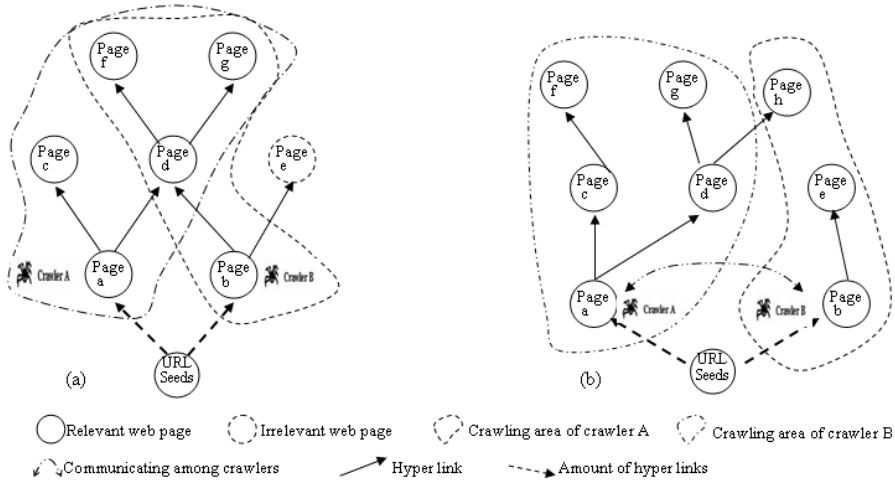


Figure 1. The overlap and cooperation among focused crawlers

2 RELATED WORKS

2.1 Contract Net Protocol (CNP)

From the late 80s, the agent theory and technique came from distributed artificial intelligence (DAI) and were widely applied to many areas. The development of DAI provided a technical basis for research on multi-agent system (MAS), the main aspects of MAS's research included: MAS theory, multi-agent collaboration and multi-agent planning, etc. [8]. Collaboration was one of the core issues of multi-agent [11], because the autonomous agent was the center of multi-agent coordinating its knowledge, desire, intention, planning, action with other agents. To achieve collaboration was the main objective for multi-agent [9, 12]. MAS's coordination was generally summarized as follows:

Structured organization. To ensure global consistency, hierarchical organizational structure used by some of the management group or intermediary agent collected information from the agent, created a plan to allocate resources and tasks to individual agent.

Contract. The technology, which was known the best on allocation of resource and tasks between the agents, was Contract Net Protocol (CNP) [13]. To save the resources and solve other issues of conflict, the mechanism of task announcement and bidding in markets was used to delegate tasks for the distributed system.

Multi-agent planning. More traditional AI researches had analyzed the coordination of multiple nodes as a planning problem. Multi-agent planning emphasized that avoiding conflict was inconsistent with the requirements of a large num-

ber of nodes of information sharing and processing. It involved many of the computation and communications.

Yun et al. [14] proposed a contract net protocol based on information intermediary service, a public message board. Degree of credibility and availability were added to the contract net, which can find contractors for the contract managers more effectively and accurately. Moreover, this reduced the amount of communications in the MAS and improved the performance of the collaboration. Gutiérrez [15] proposed a method with measuring the exchange of multi-agent, detecting the undesirable communication and classifying the agents according to sending and receiving situation. Raza et al. [16] proposed an improved contract net protocol by adding quality evaluation which is made up of Reliability, Trust and Reputation. Their proposed method narrowed the scope of the tender and reduced the time spent. Chen et al. [17] proposed a method where a threshold was set for the number of invitations for tendering. The availability of the participant was introduced into the tender evaluation process. Yang et al. [18] improved contract net protocol based on the above method, a weighted evaluation function for the quality of task completion was proposed during the tender evaluation process.

2.2 Focused Crawlers Based on MAS

Since 1990, some algorithms and their web crawlers has appeared to retrieve web pages from the internet. Fish-search [5] is a simulation of the migration activity of fish, it assimilates the crawling activity that fish finds food. When a focused crawler downloads a large number of on-topic web pages and extracts a great many related hyperlinks on them, the fish-search will produce offsprings and it continues crawling. However, when it obtains a large number of off-topic web pages, the crawler will die. The disadvantage of fish-search is to judge the relevance of web pages by binary decision. The method calculates the correlation between these hyperlinks (corresponding to web pages) and user query topics by using keyword matching or regular expression matching. Shark-search [6] is an improved fish-search algorithm; its correlation coefficient can be any real number between 0 and 1, instead of fish-search's binary decision. The method calculates the correlation of web pages not only by considering the content of the web page, but also by the anchor text and the context of their neighbors. The breadth-first crawler is the simplest crawling strategy. Najork et al. [7] shows that the breadth-first crawler is a good strategy for crawling web, as it tends to discover high-quality web pages early. Best-first crawler selects the best hyperlinks for crawling web from a frontier of hyperlink queues, according to some estimation criterion [19]. Best-first crawler is considered the most successful approach of focused crawler due to its simplicity and efficiency.

With the development of internet, the large number of web pages makes the returned result too ambiguous for the necessary information [20]. The classic focused crawler was very hard to meet the personalized needs of users. In order to address this problem, some authors introduced the idea of AI (Artificial Intelligence) into

focused crawler. Batsakis et al. [1] proposed a focused crawler named new Hidden Markov Model (HMM) crawler, it was supplied with a training set consisting of not only the content of relevant and not relevant web pages but also the paths leading to relevant pages used to train this crawler. At last, the experimental result proved that the focused crawler can make a better choice in the web pages and paths during the crawling. Zhang et al. [9] proposed a semi-supervised clustering algorithm based on Q-Learning and semi-supervised learning, which was used for choosing the on-topic pages for the focused crawler.

A multi-agent system (MAS) was a system in which a large number of agents cooperate and interact with each other in a complex and distributed environment [21]. In MAS, each agent has incomplete information or capabilities [22], and multiple agents are organized by a certain way, and they exchange information and communicate with each other on the basis of the organizational structure to achieve collaboration among agents.

Over the past few years, MAS has been developed for a variety of application domains, ranging from comparatively small systems such as personalized email filters to large and complex systems such as air traffic [23], domain knowledge [24]. To understand the causes that may lead to failure and forecast the failure mode, Liu et al. [25] proposed architecture of PHM (prognostics and health management) based on multi-agent, and focused on introducing the tasks and roles of mission planned agent, maintenance decision-making agent, resources management agent and learning agent. Yu et al. [26] proposed a cooperation framework based on proxy agent, and used a static game model to represent agents' rationality and matching mechanism to realize cooperation among agents. Hu [27] proposed an automated negotiation model of SCM (supply chain management) based on negotiation and bidding to realize the coordination between supply chains.

Xiang [28] introduced the idea of multi-agent into focused crawler, proposed a multi-agent coordination model for focused crawler, and established a prototype of organizational structure of multi-agent for focused crawler. At the technical level, there are essential differences between the "agent" in the research work and the agent in the field of AI and DAI. Du [29] used Ontology and a formal concept analysis to establish an understanding model of multi-agent for focused crawler. Chen et al. [30] proposed DIAMS, and used collaborative information agents to help users access, collect and exchange information on the web. Every user can browse and query other users' information through his own personal agent, and these personal agents can communicate, exchange information and collaborate with other personal agents to supply facilities for the users. Wang [31] presented a new method to measure the understanding among web crawlers based on MAS. When calculate the similarity between the concepts of the agent crawlers, not just natural language comparison between words is considered, but also the use of ontology and its semantic relationships. Based on the similarity, a topic-specific crawler (here we referred to WYY) was designed and implemented on the JADE platform. In order to minimize the overlap between the activities of individual nodes, Chung et al. [32] proposed a topic-oriented collaborative crawler named X4 crawler, divided web into

general subject areas, and allowed specific X4 crawlers to crawl particular subject areas. Chan [33] proposed an intelligent spider to retrieve information for online auctioning, this intelligent spider is made up of URL searching agent and auction data agent, and it is used to observe customer's behavior for decision support. Jiang et al. [34] proposed a multi-agent based individual web spider system, which consists of information collection subsystem, cooperation agent subsystem and information analyzing subsystem, and adopts cooperation agent to resolve cooperation problems among agents. Especially, the system recommended individual web information to users, and was proved very effective.

3 THE ORGANIZATIONAL STRUCTURE OF MULTI-AGENT FOCUSED CRAWLERS

To achieve collaboration, agents communicate with each other on the basis of the organizational structure in MAS. According to the existence of an administrative agent, the organizational structure of MAS can be divided into centralized structure, distributed structure and hybrid structure [28, 35]. Figure 2 shows three kinds of organizational structure of MAS, hybrid structure can be broken down into two cases, as shown in C and D.

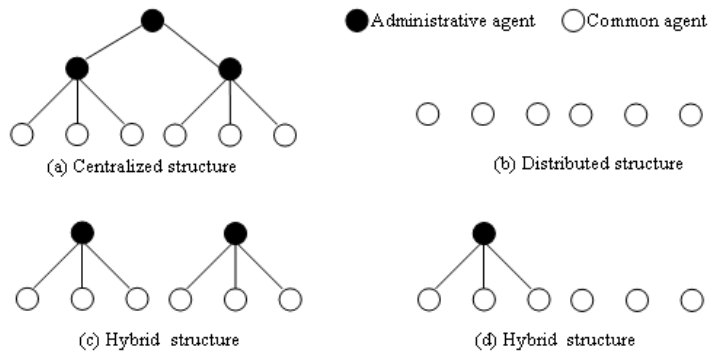


Figure 2. Three kinds of organizational structure of MAS

3.1 A New Organizational Structure of Focused Crawlers Based on Multi-Agents

The multi-agent web crawlers [28] in a focused search engine are classified into two types in our research: facilitator-agent (F-Agent) and crawler-agent (C-Agent). The F-Agents are in charge of a plan creation, task assignment and management of the communication among agents, so they are also called management agents. The C-Agents are in charge of carrying out specific tasks, and are controlled by the

corresponding F-Agents, so they are also called task agents. All agents are classified into a certain number of groups, and each group has a F-Agent and a number of C-Agents. Based on this structure, the multi-agent web crawlers can fulfill the objective of cooperation by the communication and negotiation among agents. For example, the model consisting of three groups is used to describe the structure as seen in Figure 3 a). The core task of agents in topic-specific search engines is to collect the relevant web pages to specific topics. If we search web pages with a specific topic on the internet and the specific topic can be represented by a number of keywords, then we can search the relevant web pages by the matching technology between keywords and web pages. Xiang, Du [29] had improved focused crawlers of the MAS by adding the understanding (based on concept lattice and ontology) among the agents and simulated the crawling of focused crawlers of MAS in JADE platform.

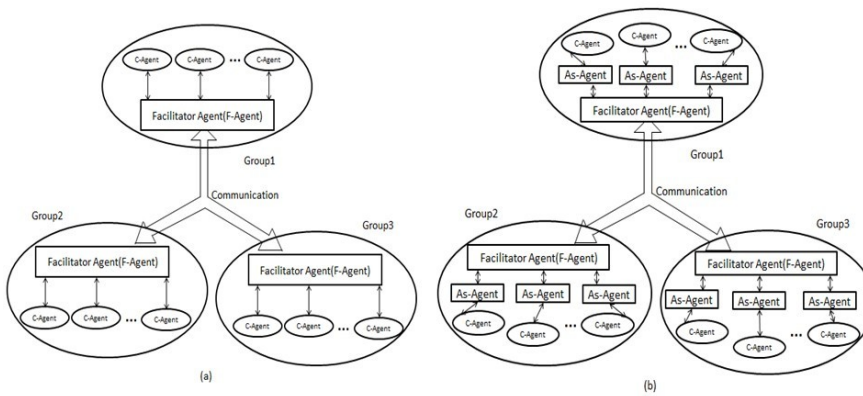


Figure 3. The organizational structure of multi-agent focused crawlers

When all C-Agents have crawled for a period of time, the C-Agents are going to evaluate the quality of these web pages and calculate their own ability. At the same time, maybe some C-Agents are fetching web pages, other C-Agents must wait these C-Agents until they finish downloading documents and extracting links, then they are going to coordinate with other C-Agents. If they do not coordinate in time, it leads to wasting a large amount of time in waiting. The tasks of C-Agents are very heavy, C-Agents are not only responsible for crawling web pages (including page's content, extracting web links and filtering and ranking the web links), but also continually communicating with F-Agents, exchanging information, and regularly constructing or updating the concept lattice. And so, this organizational structure is unreasonable. In order to address this problem, we divide the agents in MAS into three categories (Figure 3 b)), namely F-Agent (Facilitator-Agent), As-Agent (Assistance-Agent) and C-Agent (Crawler-Agent). We will describe their roles and functions in MAS, respectively.

F-Agent: F-Agents take on management and service functions for accepting the task and dividing into sub-tasks. The main roles are to communicate with other team's F-Agents and the same team's As-Agents to receive the user query, to provide the initial URLs for the C-Agents, to manage and guide the crawling of C-Agents.

As-Agent: As-Agents are seen as C-Agent's assistant, their responsibilities are to communicate with the same team's F-Agent and C-Agents, to evaluate the quality of web pages regularly. As-Agents are a communication bridge between the same team's F-Agents and C-Agents, C-Agents focus on downloading web pages, but cannot directly communicate with the outside world. If C-Agents want to contact with the outside world, they must go through an As-Agent, which is responsible for exchanging and collaborating with other F-Agent and As-Agents in the same group.

C-Agent: C-Agents play a role of a focused crawler. Their responsibilities are crawling on the web, downloading the web pages, extracting URLs from downloaded web pages, removing the label of the web pages and save them as files.

The structure of multi-agent for focused crawlers mentioned above makes the C-Agent mainly focus on web crawling, reduce the communication with the outside world and assign the task of communication and collaboration to its assistant As-Agent. Every As-Agent has different knowledge background, if As-Agent wants to correctly calculate its own ability, it must use Knowledge Base, in which storing WordNet and some concept lattice constructed by F-Agent and As-Agent. Our experimental results show that the reasonable division of C-Agent's functions can improve the performance of the system significantly.

4 COLLABORATION MODEL OF FOCUSED CRAWLERS BASED ON MAS

After describing the new structure of multi-agent for focused crawlers and the function of each agent in detail above, this section focuses on the basic processes of focused crawler based on MAS and how to collaborate with each other.

4.1 Collaboration Between Agents

Contract net protocol is a high-level exchange protocol put forward by the R. G. Smith initially used in solving distributed problems. In a distributed problem, communication and collaboration between nodes are based on contract net protocol [36]. Contract net protocol is the best coordination mechanism in multi-agent which is widely used.

Contract net protocol mainly divides the collaborative process of tasks and resource allocation into four stages [16, 17, 18]:

1. Task announcement/call for proposals (cfp),
2. Bidding,
3. Contracting,
4. Termination.

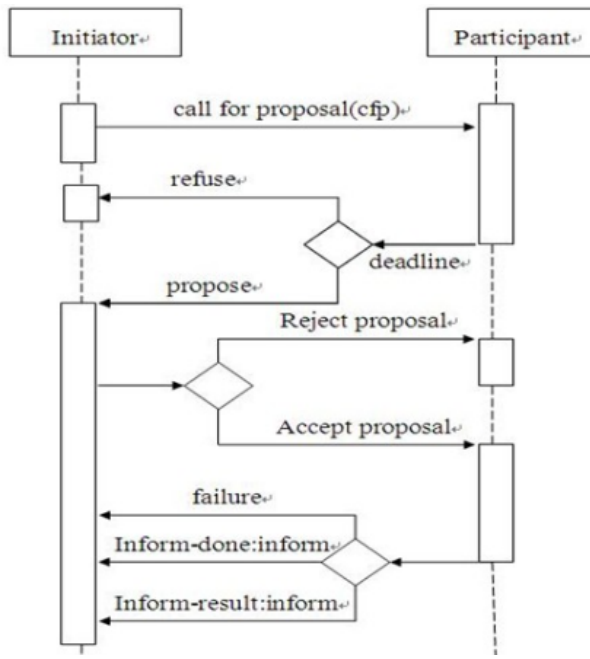


Figure 4. FIPA contract net protocol sequence diagram

Figure 4 shows the sequence of above steps based on the FIPA (Foundation for intelligent physical agents) contract net protocol [16]. When agents cannot independently solve the problem in a limited time, they need to ask other agents for help to complete their task together, the agents named initiator agents need to send call for proposals message to some possible contractors named participant agents, each participant agent reviews the received cfp and bids on the most feasible contracts before a declared deadline. Initiator agents will contract with the participant agents that have the biggest competitive ability in all agents participated in bidding, and in this contract, the initiator agents become the managers and the participant agents become the contractors. When the contractors complete the tasks specified in the contracts; even if the tasks are not completed within the prescribed times; or they inform the intermediate results of the tasks, the contracts terminate.

In classical contract net protocol, the initiator agents usually sent cfps in a broadcast way. With the system scaling up, the number of participant agents and initiator

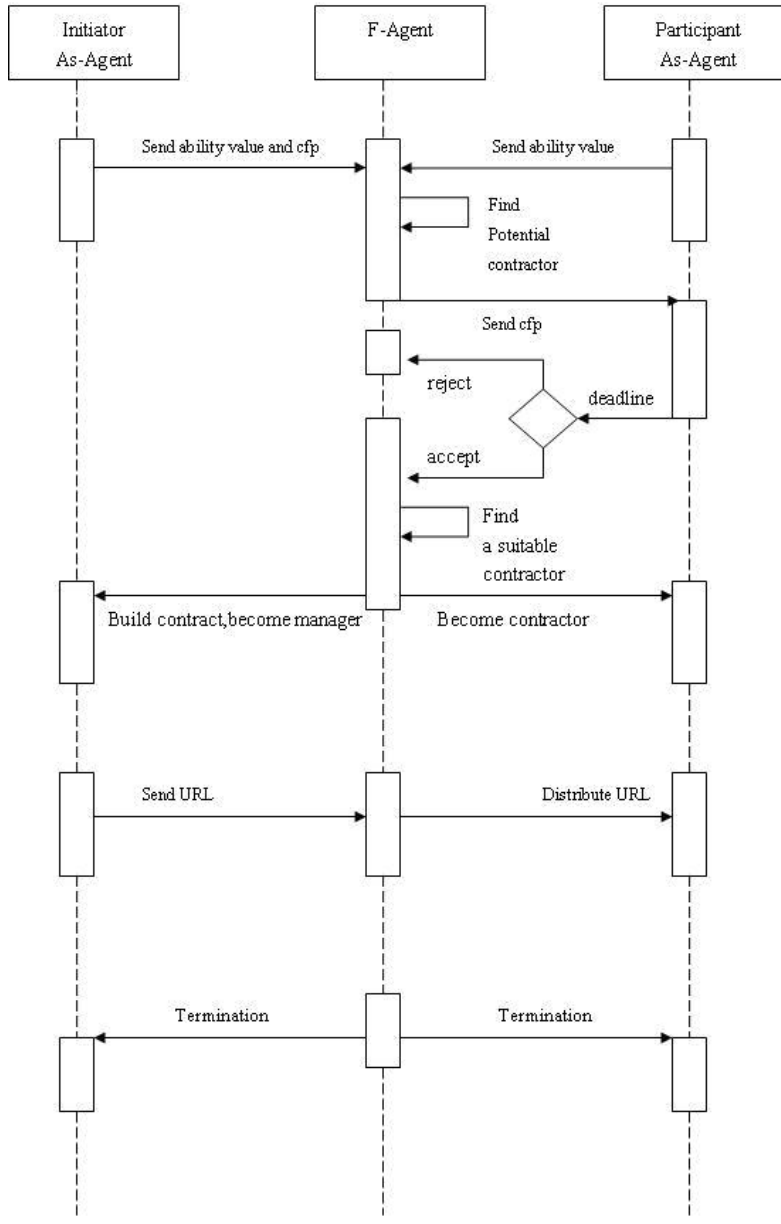


Figure 5. Collaboration of focused Crawlers based on MAS

agents would increase greatly. Communication load of the system would probably undergo a drastic change and cause communication congestion. In order to address this problem, Yun et al. [14] proposed a contract net protocol based on information intermediary service (CNPIIS). In CNPIIS, a common message blackboard (CMB) is introduced to provide public information services for the negotiations and collaborations among agents, and CMB is used to record the dynamically changing capability of each agent with time, this reduces the amount of communications in the MAS and improves the performance of the collaboration.

In this paper, an improved CNPIIS is used in MAS for collaboration and coordination between focused crawler agents. Instead of CMB in CNPIIS, we use the F-Agent to record the capability of each As-Agent, and the F-Agent can find some potential contractors for the initiator As-Agent. At last, the F-Agent will find a suitable contractor for the initiator As-Agent. We set up an ability threshold of calling for proposals for the initiator of a contract, Only the As-Agent's ability is greater than the given threshold, the As-Agent can call for proposals. In the process of competing, F-Agent will select the agent that has the biggest competitive ability from all agents participated in competing, and let this agent be a contractor. We define competitive ability as the reciprocal of agent's ability value. So competitive ability is defined in Section 4.2.

Our improved CNPIIS is divided the negotiation strategy of tasks and resource allocation into five stages (Figure 5):

1. Call for proposals (cfp)
2. Bidding,
3. Contracting,
4. Distribute URL,
5. Termination.

(1) Call for proposals The pseudo-code description of the cfp can be described as follows:

Algorithm: Call for proposals

(01) $As\text{-}Agent_i$ needs to calculate its own ability
(02) OnTick()
(03) Begin
(04) ExamAgent()
(05) Begin
(06) ActionTfidf(fileBeginNum, fileEndNum);
(07) ConstructConceptLattic(URLs, Keywords);
(08) End
(09) Ability = CalculateAbility();
(10) If (ability >= W)
(11) Send ability and cfp to F-Agent;
(12) Else

(13) Send ability to F-Agent;
(14) End

When all C-Agents have crawled for a period of time, As-Agents will use a formal context that is made up of URLs and keywords to build concept lattice [29], which will be used to calculate the ability values of As-Agents. This function of ability will be introduced in Section 4.2. According to its ability, As-Agent will make the decision whether to call for proposals, if its ability value is greater than the given threshold W, the ability value and a cfp message will be sent to F-Agent; otherwise only the ability value is sent to F-Agent. In cfp message (Figure 6 (up)), the initiator As-Agent sends n links and their scores to F-Agent.

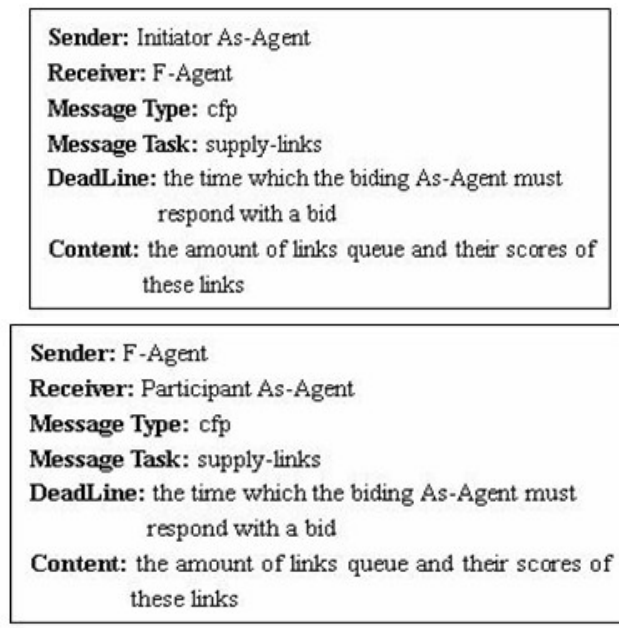


Figure 6. The cfp of As-Agent and F-Agent

(2) Bidding F-Agent will select m potential contractors for the initiator As-Agent, and these potential contractors will decide whether to bid. The bidding algorithm can be described as follows:

Algorithm: Bidding

(15) ArrayList.addLast($As\text{-}Agent_i$);
(16) ArrayList.sort();
(17) Manager = ArrayList.getFirst();
(18) If (manager.ability >= W)
(19) Begin

```
(20)      for i = 1:m
(21)      Begin
(22)          potential_i = ArrayList.getLast();
(23)                  send cfp to potential_i;
(24)      End
(25)  End
(26)  For i = 1:m
(27)  Begin
(28)      If (potential_i.URL.Score <= cfp.URL.Score)
(29)      Begin
(30)          Send Bidding to F-Agent;
(31)      End
(32)  End
```

When the F-Agent receives the ability values and the cfp message, it will select m potential contractors for the initiator As-Agent from the participant As-Agents with lower ability, and then it will send a cfp message to these m potential contractors, the cfp message is shown in Figure 9. Those m potential contractors receive the cfp message, they will decide whether to accept the cfp with the deadline, if the highest score of the link from the link queue of the potential by contractor is lower than all the scores of the links in the cfp message, the potential contractor will accept the cfp, and send a bidding message to F-Agent; otherwise the potential contractor will reject the cfp. Figure 7 shows the bidding message.

Sender: Participant As-Agent
Receiver: F-Agent
Message Type: bidding
Message Task: receive-links
DeadLine: the time which the F-Agent must respond with a contact or a review of the bid
Content: the ability of the bidding As-Agent and the links supplied by initiator As-Agent

Figure 7. Bidding

(3) Contracting The pseudo-code description of the contracting can be described as follows:

Algorithm: Contracting

```
(33)  ArrayList.addLast(Bidding As-Agent_i);
(34)  ArrayList.sort();
(35)  contractor = ArrayList.getLast();
```

(36) send contracting to manager;
(37) send contracting to contractor;

At this stage, the F-Agent will find a suitable contractor for the initiator As-Agent from the potential contractors who have participant in bidding. It will read the bidding message from the potential contractors, and choose the As-Agent with the lowest ability value as a contractor of the initiator As-Agent, and a contract will be established between the initiator and the participants. Meanwhile, the initiator will become the contract manager, and this participant becomes the contractor. At last the F-Agent will send the contracting message to the manager and the contractor, the contracting message is shown in Figure 8 (up) and Figure 8 (down).

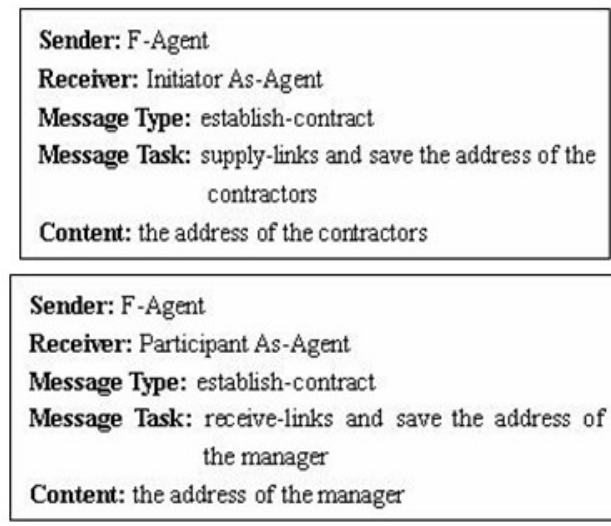


Figure 8. The contracting message of the manager and contractor

(4) Distribute URL When the contract manager receives the contracting message from a F-Agent. The manager will give its own crawling URL to it, and then the F-Agent allocates the URL to the corresponding contractor, afterward the contractor crawls on the web according to the URL.

(5) Termination At this stage, a F-Agent will send termination message to the manager and contractors, the contract relationship between the manager and the contractors will be terminated. All As-Agents will calculate their abilities for the next collaboration.

4.2 Function of Ability

A reasonable ability function not only can correctly evaluate the ability of a C-Agent, but also can decrease the unnecessary interaction between the multiple C-Agents. It plays an important role in improving the precision and the crawling efficiency of our focused crawler.

In this paper, our ability function takes not only the semantic importance of downloaded pages into account, but also the score of hyperlinks from downloaded pages. The ability function is as follows:

$$E(A_i) = \alpha * S(A_i) + \beta * M(A_i) \tag{1}$$

where $S(A_i)$ reflects the score of the hyperlinks from A_i (C-Agent i), $S(A_i) = \frac{1}{n}\Sigma_{j=1}^{n} P_{u_j}(A_i)$; it is the average of URL prediction score; n is the sum of all hyperlinks that are on-topic.

$$P_{u_j}(A_i) = a_{u_j} \times \mathrm{Sim}(C_{u_j}^{p}(A_i)) \tag{2}$$

where a_{μ_j} is the weight of the anchor text of hyperlink j, calculated by:

$$a_{u_j} = \begin{cases} 1, & u_j \text{ contains } q, \\ cs(a_j, q), & \text{else} \end{cases} \tag{3}$$

where u_j is the anchor text of hyperlink j, q is the user query word, and $cs(a_j, q)$ calculated by [14]:

$$cs(a_j, q) = \begin{cases} 1 & a_j = q, \text{ or } a_j \text{ and } q \text{ have synonymous relation in WordNet,} \\ \delta_1 & a_j \text{ and } q \text{ have } ISA \text{ relation in WordNet,} \\ \delta_2 & a_j \text{ and } q \text{ have PartOf relation in WordNet,} \\ 0.001 & \text{others,} \end{cases} \tag{4}$$

where δ_1, δ_2 $(0 < \delta_1 < \delta_2 < 1)$ can be given by user.

$\mathrm{Sim}(C_{u_j}^{p}(A_i))$ is the similarity of the parent pages of the hyperlinks, we use cosine similarly to calculate this similarity.

$$\mathrm{Sim}(C_{u_j}^{p}) = \frac{\vec{d_i} \cdot \vec{q}}{|\vec{d_i}| \cdot |\vec{q}|} \tag{5}$$

where $\vec{d_i} = (w_1, w_2, \cdots, w_n)$, we calculate ω_k of item by using TF-IDF for the document with statistic method, and then extract the top n terms as the vector $\vec{d_i}$.

$\vec{q} = (q_1, q_2, \cdots, q_n)$, where $\vec{q_i} = cs(S_{w_i}, q), (1 \le i \le n), S_{w_i}$ is the word corresponding to the i^{th} component of $\vec{d_i}$, q indicates the query word.

$M(A_i)$ reflects the importance of downloaded pages. $M(A_i)$ is the concept similarity in lattices constructed by C-Agent and F-Agent. We use the definition given in [29]. α, β are weights assigned to the above formula (2); $0 \le \alpha \le 1, 0 \le \beta \le 1$ and $\alpha + \beta = 1$. They can be determined by the user to enrich the flexibility of this method.

4.3 The Related Compositions of Calculating Ability

In Section 3.2, we have showed that every As-Agent has different knowledge background if an As-Agent wants to correctly calculate its own ability; it must use the Knowledge Base, in which some knowledge and semantic relations are extracted from WordNet and some concept lattice (as shown in Figure 9) constructed by F-Agent and As-Agent. In this section, we will introduce the related compositions of calculating ability in detail. As shown in Figure 10, there are three related compositions of calculating ability, which respectively are DocPool, Knowledge Base and Calculate Ability.

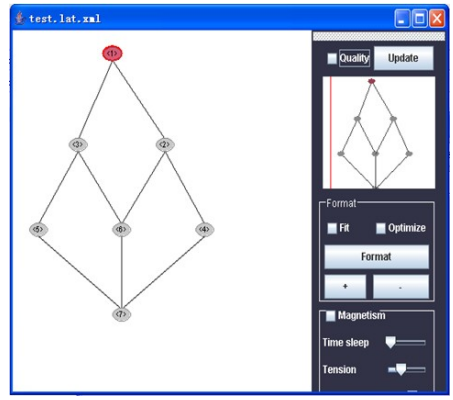


Figure 9. Formal concept lattice

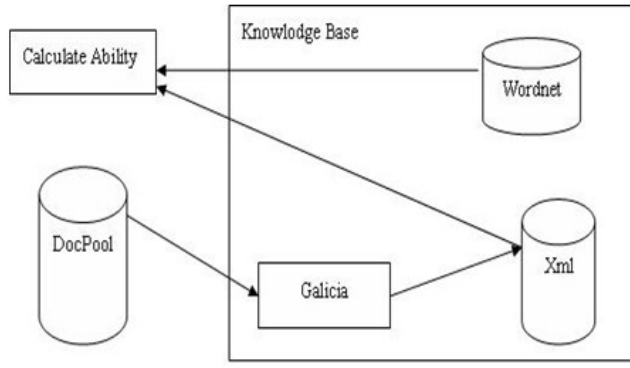


Figure 10. The related compositions of calculating ability

DocPool:

1. Web pages and files with HTML tags removed: When C-Agents crawl on the web, they will fetch out related hyperlinks and download the web pages and remove HTML tags of these web pages, and then save the web pages and files with HTML tags removed in DocPool.
2. TF files, IDF files and TF-IDF files: After crawling on the web for some time period T, all As-Agents will calculate the TF-IDF of the above files with HTML tags removed, and store the calculating results as TF files, IDF files and TF-IDF files in DocPool.

Knowledge Base:

1. Galicia [37]:
   (a) Bivariate the table file: establish the bivariate table which can be mapped to the object according to the relationships of the URLs and the keywords, then generate the file with the postfix *.bin.xml.
   (b) The concept lattice files: build a concept lattice object according to the given *.bin.xml file, and automatically start a thread to produce the file with the same name *.lat.xml.
2. Xml: *.bin.xml and *.lat.xml produced in Galicia will be stored in Xml, and they will be used to calculate the ability of As-Agent and F-Agent.
3. Wordnet: Wordnet is a lexical database of English, it is used to compute the similarity between two terms.

Calculate Ability:

We will use the scores of the links of C-Agents and knowledge base to calculate the ability of As-Agents, the function of ability has been introduced in Section 4.2.

5 SYSTEM ARCHITECTURE

Our proposed MAS crawler system is made up of three compositions: user interface (as shown in Figure 11), MAS Crawler and DocPool. The user interface is used to connect a user and the system, and it is used to receive user query terms and other parameters input by users. MAS Crawler is the core composition of this system, in which focused crawlers can communicate and collaborate with each other. DocPool has been introduced in Section 4.3, which is used to store web pages and files with HTML tags removed from focused crawlers (C-Agents and F-Agents).

Figure 12 shows the process diagram of our MAS crawler system, and we will introduce the working process in detail.

1. This system firstly starts F-Agent, F-Agent receives user query terms and other parameters input by users, and uses win socket technology to access the Google servers to get a series of URLs as the seed URLs.

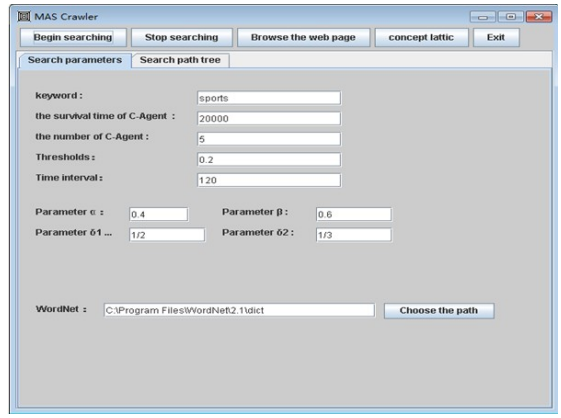


Figure 11. User interface of MAS crawler

2. Each new As-Agent sends a registration message to the F-Agent, and then this As-Agent will be registered in the F-Agent.
3. F-Agent provides an initial URL for the new registered As-Agent, the As-Agent will start a C-Agent and allocate its URL to the C-Agent to crawl on the web.
4. After crawling on the web for some time T, all As-Agents measure their own ability values to determine whether their ability values are greater than the given threshold value W. If there are some As-Agents' ability values greater than the given threshold value W, these As-Agents will send a call for proposals (cfp) to the F-Agent. Otherwise, they will continue to crawl on the web.
5. When F-Agent receives a cfp message (we assume that this cfp message has come from As-Agent1), it will find an As-Agent with the minimum ability value (we assumed it is As-Agent3), and allows As-Agent1 and As-Agent3 to establish contracts, at the same time, the F-Agent will send a contract message to As-Agent1 and As-Agent3.
6. Contract managers provide URLs for their contractors, to let them continue crawling on the web.
7. After crawling on the web for some time T, the F-Agent will judge whether the running time is over the dead limit, so the F-Agent will send STOP message to all As-Agents. And these As-Agents will send STOP message to all C-Agents, at this time, the system stops. Otherwise, C-Agents will ask As-Agent for a URL that has not been crawled, and continue working on the web.

6 EXPERIMENTATION

In this section, we used the JADE (Java agent development framework), an open source development framework, to achieve the proposed focused crawler based on

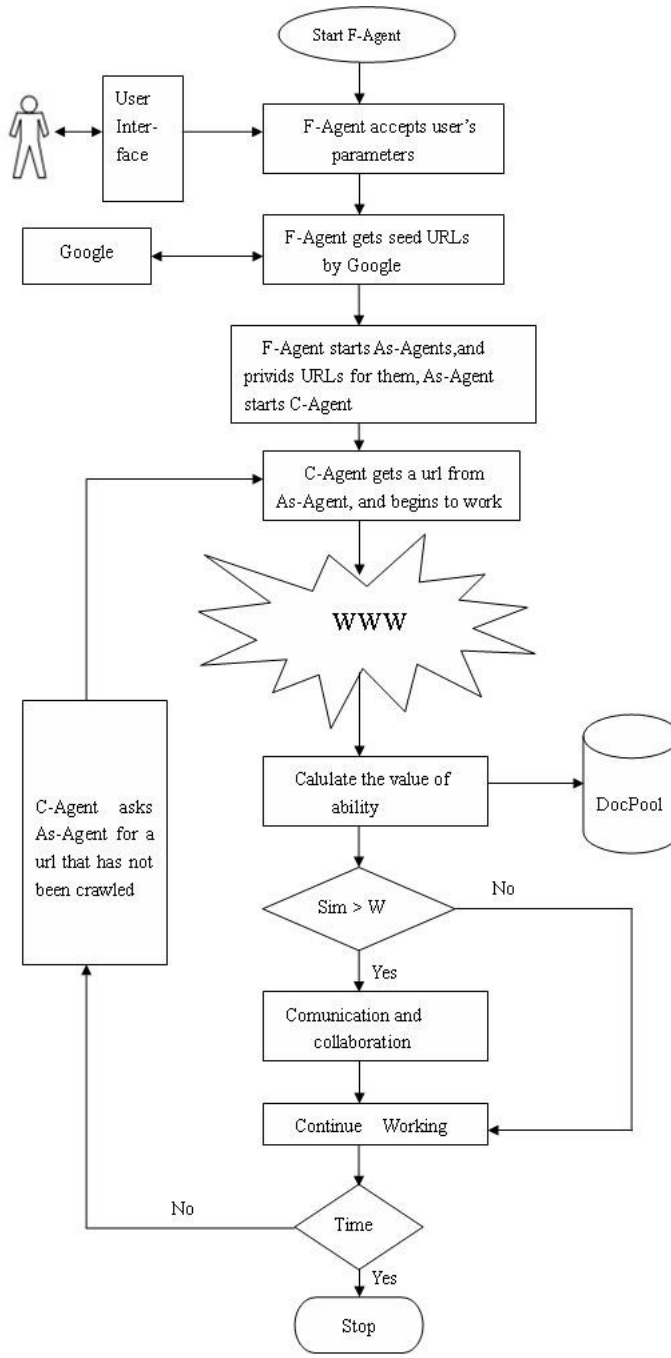


Figure 12. The process chart of our system

MAS. Figure 13 is the remote agent management GUI of JADE. We compared our proposed MAS crawler with other three kinds of focused crawlers on precision and crawling efficiency, these three kinds of focused crawlers respectively are: breadth-first crawler, best-first crawler, Wang [31] mentioned crawlers. All crawlers are implemented in Java, and ran in a computer with the 1M bandwidth and 1G memory. We chose "sports" as a topic, and obtained the initial URL seeds in our experiment by visiting the Google server. Figure 14 shows the search path tree of the proposed focused crawler based on MAS.

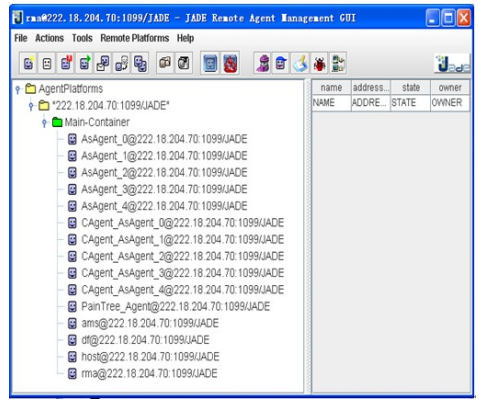


Figure 13. JADE remote agent management GUI

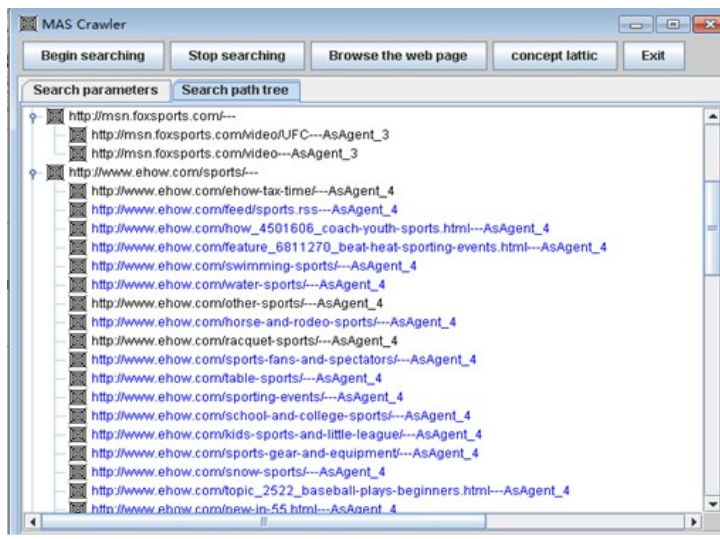


Figure 14. Search path tree of MAS crawler

6.1 Metrics for the Performance of a System

Mencze et al. [38] used the two evaluation indices of information retrieval to evaluate the performance of the focused crawler: precision and recall. However, crawling efficiency plays an important role in evaluating the performance of the focused crawler. Therefore, we use the precision (P) and efficiency (E) to measure the performance of focused crawler.

$$P = \frac{|R|}{|A|} \tag{6}$$

- $|R|$: The number of collection web pages which are on-topic.
- $|A|$: The number of collection web pages.

$$E = Num \tag{7}$$

- *Num*: The number of collection web pages after crawling on the web for some time T.

6.2 Thresholds

In order to make the focused crawlers achieve optimal performance, the ability threshold ω must be established to determine an agent whether it has ability to call for a proposal. Based on this, the agent contracts with other agents to enable collaboration. If the threshold W selected is too small, the number of contracts is bound to increase, it costs more time in building contracts and assigning tasks for the contractors, and may cause communication congestion. And with the increase of the number of contracts, part of contractors will give up their crawling path of high correlation topic to follow their manager's crawling path, so that the precision and the crawling efficiency will decrease. If the threshold W selected is too large, the number of contracts is bound to decrease. It makes the task of part of agents heavy, and unable to find contractors to work for them. In this case, our system cannot achieve a better collaboration between the multiple agents, and part of agents will crawl in a path of low topic correlation, so the precision will decrease. Therefore, choosing the appropriate threshold ω plays a vital role.

As can be seen from Table 1 and Figures 15 and 16, when the ability threshold is 0.25, the average precision is the highest, and when the ability threshold becomes larger or smaller, the precision will decrease. So in our experiments, we select 0.25 as the ability threshold.

6.3 Experimental Results

6.3.1 Crawlers Overlap

Different web pages could have the same hyperlinks, if a focused crawler cannot communicate with other focused crawlers, it will crawl the region that another crawler

Time(M)	0.05	0.1	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.5
5	0.74	0.93	0.80	0.88	0.92	0.83	0.81	0.82	0.94	0.91
10	0.90	0.89	0.88	0.90	0.94	0.92	0.89	0.92	0.91	0.92
15	0.91	0.91	0.89	0.93	0.94	0.92	0.90	0.94	0.94	0.94
20	0.92	0.92	0.91	0.93	0.95	0.93	0.92	0.94	0.94	0.94
25	0.93	0.93	0.92	0.94	0.95	0.94	0.93	0.94	0.94	0.93
30	0.94	0.94	0.93	0.94	0.96	0.94	0.93	0.93	0.94	0.93
35	0.94	0.94	0.93	0.94	0.96	0.94	0.93	0.93	0.93	0.93
40	0.95	0.94	0.93	0.95	0.96	0.94	0.94	0.93	0.93	0.93
45	0.95	0.94	0.93	0.95	0.96	0.94	0.94	0.94	0.93	0.93
50	0.95	0.95	0.93	0.95	0.96	0.95	0.94	0.94	0.93	0.94
55	0.95	0.95	0.93	0.95	0.96	0.95	0.94	0.94	0.93	0.93
60	0.95	0.94	0.93	0.95	0.96	0.95	0.94	0.94	0.93	0.93
Average P	0.92	0.93	0.91	0.94	0.95	0.93	0.92	0.93	0.93	0.93

Table 1. Comparison of the precision under different ω and time

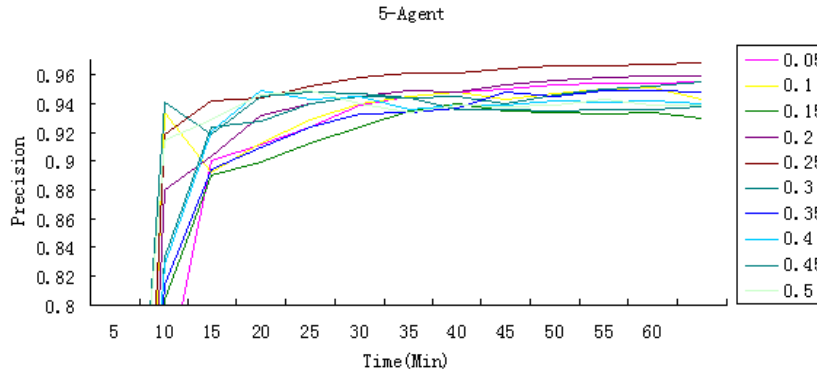


Figure 15. Comparison of the precision of 5 agents under different ω and time

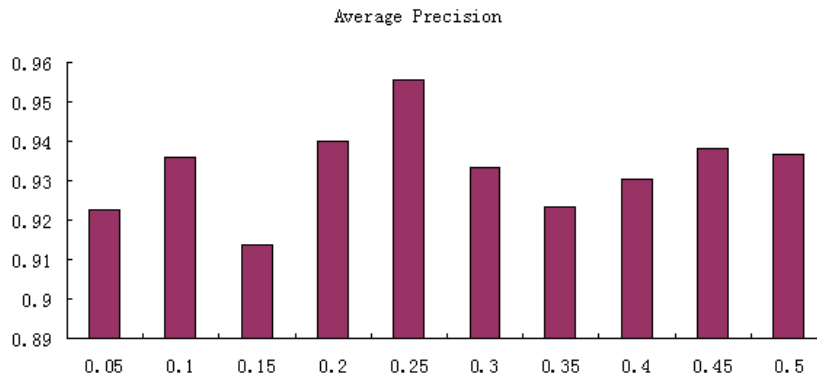


Figure 16. Comparison of the average precision of 5 agents under different ω

has crawled, so there exist overlaps between different focused crawlers. In this paper, all focused crawlers are organized by a new organizational structure, they can communicate and collaborate with each other to avoid the overlaps. In the new organizational structure of multi-agent, all focused crawlers are divided into three categories: F-Agent, As-Agent and C-Agent. The main role of F-Agent is to manage and guide the crawling of C-Agent, and F-Agent records all URLs that have been crawled to judge whether a new URL has been crawled or not. Before C-Agent is going to download a web page, it will ask As-Agent for a URL that has not been crawled, As-Agent will send the first URL from the crawling queue of C-Agent to F-Agent, and F-Agent will judge whether this URL has been crawled or not.

Table 2 demonstrates that our proposed MAS focused crawler can minimize the overlap between the activities of crawlers. With the increase of the number of focused crawlers, the overlap among the activities of crawlers is going to be larger. Our proposed MAS focused crawlers with communication and collaboration can greatly reduce the overlap (overlap rates are 0 under 5 agents and 10 agents), avoid different crawlers crawling the same web page, and improve the performance of our system.

Number of C-Agents	5 Agents	10 Agents
Number of overlap (all) web pages without As-Agent	66 (724)	491 (1 464)
Overlap rate (%) without As-Agent	9.116	33.538
Number of overlap (all) web pages with As-Agent	0 (658)	0 (973)

Table 2. Crawlers overlap

6.3.2 Crawlers Evaluation

In this paper, we take two indices mentioned above into account, precision and crawling efficiency, and compare our proposed focused crawler (shorten for MyMethod) with other crawlers with breadth-first, best-first, XiangDan [28], WYY [31] strategies. We divide our experiment into two parts:

- 5 C-Agents, the seed URLs of these web crawlers are shown in Table 3;
- 10 C-Agents, the seed URLs of these web crawlers are shown in Table 4.

1	`http://en.wikipedia.org/wiki/Sport`
2	`http://sports.yahoo.com/`
3	`http://espn.go.com/`
4	`http://sports.com/`
5	`http://www.ehow.com/sports/`

Table 3. Seed URLs under 5 C-Agents

From the above Figures 17, 18, 19 and 20, we can see that the number of web pages returned by breadth-first crawler is the largest and precision is smallest under

1	http://en.wikipedia.org/wiki/Sport
2	http://sports.yahoo.com/
3	http://espn.go.com/
4	http://sports.com/
5	http://www.ehow.com/sports/
6	http://www.dmoz.org/Sports/
7	http://sportsillustrated.cnn.com/
8	http://news.bbc.co.uk/sport
9	http://www.olympic.org/sports
10	http://www.sportsdirect.com/

Table 4. Seed URLs under 10 C-Agents

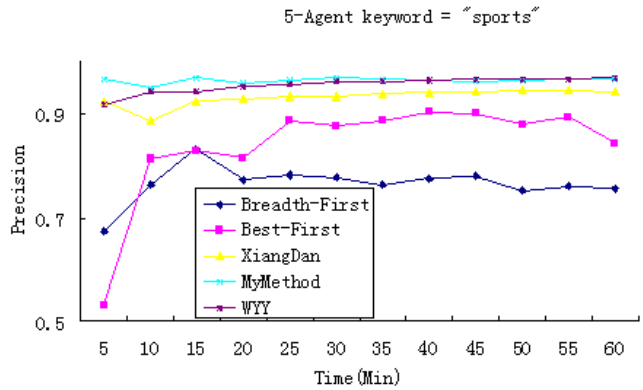


Figure 17. The precision comparison of the different crawled strategy under 5 agents

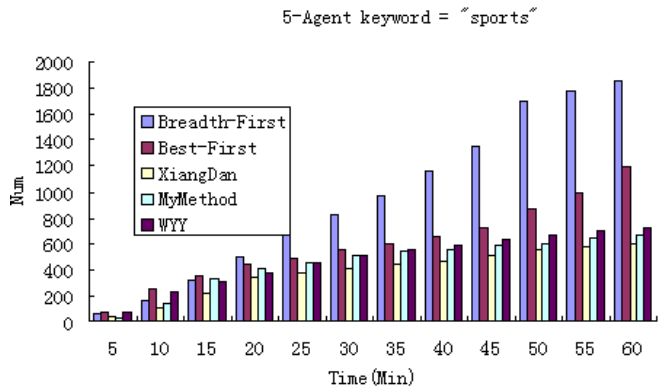


Figure 18. The crawling efficiency comparison of of the different crawled strategy under 5 agents

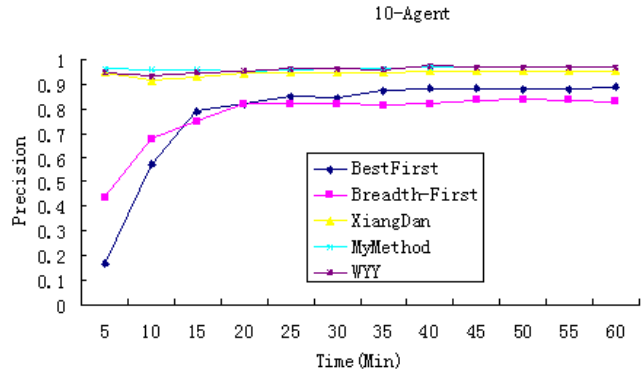


Figure 19. The precision comparison of the different crawled strategy under 10 agents

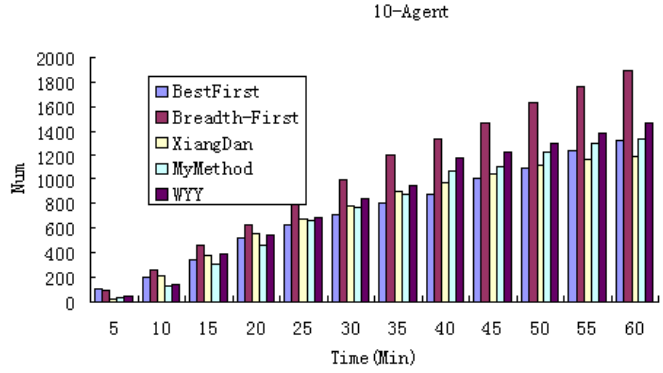


Figure 20. The crawling efficiency comparison of of the different crawled strategy under 10 agents

two cases (5-Agents and 10-Agents). This is due to the lack of judging and processing conditions. A lot of irrelevant pages are crawled and a lot of time and resources are spent in crawling web pages unrelated to the topic, so the crawling precision is not high. The number of retrieved pages and the precision by WYY, XiangDan and MyMethod focused crawlers with cooperation almost equal. Because the Agents in WYY and XiangDan focused crawler does not have a reasonable division of functions and competitions among agent crawlers, it results in an arduous task for the C-Agent and low efficiency. Thus, our multi-contract net protocol can deal with the issue of collaboration between agents and achieve better results. The crawling efficiency and precision can be improved by multiple crawlers' communication, collaboration and competition.

7 CONCLUSIONS AND FUTURE WORK

In this paper, a new organizational structure of multi-agent for focused crawlers is proposed, in which the agents were divided into three categories: F-Agent (Facilitator-Agent), As-Agent (Assistance-Agent) and C-Agent (Crawler-Agent). Based on this new organizational structure, an improved CNPIIS is used in MAS for collaboration and coordination between focused crawlers. At last, we compared its performance on the crawling precision and efficiency with other four kinds of focused crawlers. The experiment shows that our proposed MAS Crawler has the highest precision.

But there is still much to be desired, for the collaboration between agents has not reached the height of knowledge and reasoning, so the intelligence of agent is not high. Therefore, the next job in our focus is understanding in multi-agent, and thus to achieve a better collaboration between the agents.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 61271413 and 60872089), the Cultivating Foundation of Science and Technology Leaders of Sichuan Province, Chunhui Plan of Education Department of China.

REFERENCES

[1] Batsakis, S.—Petrakis, E. G. M.—Milios, E. E.: Improving the Performance of Focused Web Crawlers. Data and Knowledge Engineering, Vol. 68, 2009, No. 10, pp. 1001–1013.

[2] Diligenti, M.—Coetzee, F. M.—Lawrence, S.—Giles, C. L.—Gori, M.: Focused Crawling Using Context Graphs. 26th International Conference on Very Large Databases, Cairo, Egypt, 2000, pp. 527–534.

[3] Chakrabarti, S.—van den Berg, M.—Dom, B.: Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. Computer Networks, Vol. 31, 1999, No. 11–16, pp. 1623-1640.

[4] Yang, Y. K.—Du, Y. J.—Sun, J. Y.—Hai, Y. F.: A Topic-Specific Web Crawler with Concept Similarity Context Graph Based on FCA. Proceeding of 4th International Conference on Intelligent Computing, China, 2008, pp. 840–847.

[5] De Bra, P.—Houben, G.—Kornatzky, Y.—Post, R.: Information Retrieval in Distributed Hypertexts. Proceedings of the 4th RIAO Conference, New York, 1994, pp. 481–491.

[6] Hersovici, M.—Jacovi, M.—Maarek, Y.—Pelleg, D.—Shtalhaim, M.—Ur, S.: The Shark-Search Algorithm – An Application: Tailored Web Site Mapping. Proceedings of the Seventh International World Wide Web Conference. Brisbane, Australia, 1998, pp. 317–326.

[7] Najork, M.—Wiener, J. L.: Breadth-First Search Crawling Yields High-Quality Pages. Proceedings of the 10th International World Wide Web Conference. Hong Kong, 2001, pp. 114–118.

[8] Liu, D. Y.—Yang, K.—Chen, J. Z.: Agents: Present Status and Trends. Journal of Software, Vol. 11, 2000, No. 3, pp. 315–321.

[9] Zhang, H. X.—Lu, J.: SCTWC: An Online Semi-Supervised Clustering Approach to Topical Web Crawlers. Applied Soft Computing, Vol. 10, 2010, No. 2, pp. 490–495.

[10] Du, Y. J.—Hai, Y. F.: Semantic Ranking of Web Pages Based on Formal Concept Analysis. The Journal of Systems and Software, Vol. 86, 2013, No. 1, pp. 187–197.

[11] Park, S.—Sugumaran, V.: Designing Multi-Agent Systems: A Framework and Application. Expert Systems with Applications, Vol. 28, 2005, No. 2, pp. 259–271.

[12] Jiang, W. J.—Wang, P.: Research on a Grid Resource Allocation Algorithm Based on MAS Non-Cooperative Bidding Game. Journal of Computer Research and Development, Vol. 44, 2007, No. 1, pp. 29–36.

[13] Jiang, Y. C.—Jiang, J. C.: A Multi-Agent Coordination Model for the Variation of Underlying Network Topology. Expert Systems with Applications, Vol. 29, 2005, No. 2, pp. 372–382.

[14] Yun, H. S.—Li, Q. S.—Jiang, D.—Liu, H.—Mao, S. J.—Li, Y. P.: A Contract Net Protocol Based on Information Intermediary Service in Multi-Agent System. International Conference on Artificial Intelligence and Computational Intelligence, 2009, pp. 19–23.

[15] Gutiérrez, C.—García-Magariño, I.—Fuentes-Fernández, R.: Detection of Undesirable Communication Patterns in Multi-Agent Systems. Engineering Applications of Artificial Intelligence, Vol. 24, 2011, No. 1, pp. 103–116.

[16] Raza, M.—Hussain, F. K.—Hussain, O. K.—Chang, E.: Q-Contract Net: A Negotiation Protocol to Enable Quality-Based Negotiation in Digital Business Ecosystems. 2010 International Conference on Artificial Intelligence and Computational Intelligence, 2010, pp. 161–167.

[17] Chen, X. G.—Xong, H. G.: Further Extensions of FIPA Contract Net Protocol: Threshold Plus DoA. Proceedings Agents, Interactions, Mobility, and Systems (AIMS), 2004 ACM Symposium on Applied Computing (SAC'04), 2004, pp. 45–51.

[18] Yang, J.—Li, W. L.—Hong, C. Y.: Improvement to Contract Net Protocol Based on Threshold and Availability. Computer Integrated Manufacturing Systems, Vol. 15, 2009, pp. 5, pp. 1016–1022.

[19] Menczer, F.—Pant, G.—Srinivasan, P.: Topical Web Crawlers: Evaluating Adaptive Algorithms. ACM Transactions on Internet Technology (TOIT), Vol. 4, 2004, No. 4, pp. 378–419.

[20] Yang, S. Y.: OntoCrawler: A Focused Crawler with Ontology-Supported Website Models for Information Agents. Expert Systems with Applications, Vol. 37, 2010, No. 7, pp. 5381–5389.

[21] Luo, Y. W.—Wang, X. L.—Xu, Z. Q.: Design and Development of Financial Applications Using Ontology-Based Multi-Agent Systems. Computing and Informatics, Vol. 28, 2009, No. 5, pp. 635–654.

[22] Chau, M.—Zeng, D.—Chen, H. C.—Huang, M.—Hendriawan, D.: Design and Evaluation of a Multi-Agent Collaborative Web Mining System. Decision Support Systems, 2003, Vol. 35, pp. 167–183.

[23] Jennings, N.—Sycara, K.—Wooldridge, M.: A Roadmap of Agent Research and Development. Journal of Autonomous Agents and Multi-Agent Systems, Vol. 1998, No. 1, pp. 7–38.

[24] Mirchevska, V.—Lustrek, M.—Bezek, A.—Gams, M.: Discovering Strategic Behaviour of Multi-Agent Systems in Adversary Settings. Computing and Informatics, Vol. 33, 2014, No. 1, pp. 79–108.

[25] Liu, Z. Y.—Rong, L. Q.: Study on the Multi-Agent Model for PHM System. Management Science and Industrial Engineering (MSIE), 2011, pp. 1044–1048.

[26] Yu, Z. L.—Ji, H.: A Proxy Agent Cooperation Framework. 2010 International Conference on Computer Application and System Modeling, 2010, pp. 578–587.

[27] Hu, C. H.: Automated Negotiation Model of Supply Chain Management Based on Multi-Agent. Management Science and Industrial Engineering (MSIE), 2011, pp. 178–180.

[28] Xiang, D.—Du, Y. J.: Coordination and Communication among Topic Specific Search Agents. Proceedings of the Third International Conference on Natural Computation. Piscataway, IEEE Computer Society Publications, 2007, pp. 703–707.

[29] Du, Y. J.—Wang, Y. Y.—Chen, S. M.: The Understanding Between Two Agent Crawlers Based on Domain Ontology. Neural Network World, Vol. 12, 2012, No. 4, pp. 311–324.

[30] Chen, J. R.—Mathé, N.—Wolfe, S.: Collaborative Information Agents on the World Wide Web. Proceedings of the Third ACM Conference on Digital Libraries (DL '98), 1998, pp. 279–280.

[31] Wang, Y. Y.: Research on The Understanding and Cooperation of the Topic Crawlers Based on Multi-Agent System. Chengdu, Xihua University, 2010.

[32] Chung, C.: Topic-Oriented Collaborative Web Crawling. Master's thesis, University of Waterloo, Waterloo, Ontario, Canada, 2002.

[33] Chan, C. C. H.: Intelligent Spider for Information Retrieval to Support Mining-Based Price Prediction for Online Auctioning. Expert Systems with Applications, Vol. 34, 2008, No. 1, pp. 347–356.

[34] Jiang, L. H.—Zhang, H. B.: Multi-Agent Based Individual Web Spider System. World Automation Congress (WAC), 2010, pp. 177–181.

[35] Zhao, L. W.—Hou, Y. B.: Architecture of Multi-Agents System and Cooperation. Computer Engineering and Applications, Vol. 10, 2000, pp. 59–61.

[36] Raza, M.—Hussain, F. K.—Hussain, O. K.—Chang, E.: Q-Contract Net: A Negotiation Protocol to Enable Quality-Based Negotiation in Digital Business Ecosystems. International Conference on Artificial Intelligence and Computational Intelligence, 2010, pp. 161–167.

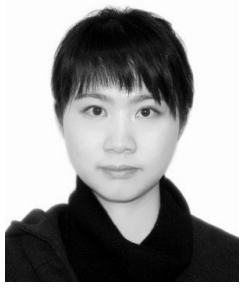
[37] `http://www.iro.umontreal.ca/galicia/`.

[38] Menczer, F.—Pant, G.—Srinivasan, P.—Ruiz, M. E.: Evaluating Topic-Driven Web Crawlers. Proceedings of the 24th Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval (SIGIR '01), 2001, pp. 241–249.

Yajun Du received his D.Sc. degree in traffic information engine and control from SWJTU (2005). Currently he is Professor at XHU (Xihua University) in computer science. He has published several papers and served in program committees of both Chinese and international conferences. His experience and research work focus on information retrieval, software engineering, search engine, web mining, and computer network.

Yong Xu received his M.Sc. degree in computer science at XHU (Xihua University) in 2012. His experience and research work focus on information systems, software engineering, knowledge graph, and web mining.



Min Wang is a lecturer of software engineering in School of Mathematics and Computer Science at XHU (Xihua University), Sichuan, China. Her main areas of research interest are artificial intelligence, information systems, software engineering, knowledge graph, and web mining.