Research on the Database based on GRID Technologies

Hua-shuai Zhang[1],Xiao-wan Yang[2]

[1]Department of InformationEngineering, Environmental Management College of China, Qinhuangdao, 066000, China

[2]Department of Environmental Science, Environmental Management College of China, Qinhuangdao, 066000, China

Keywords: Database; GRID Services; Information Grid; Focus Spider

Abstract. Grid technology is widely used in recently. It has solved the real sharing of resources, making each node of unified command and using of resources. The information grid is on the basis of computational grid, using data mining. Information fusion, search engine technology and building are advantageous for the collection and sharing of grid resources. The goal is to create a build on OS and Web based on the new generation of Internet information platform. In this platform, the distribution of the information processing is the collaboration and intelligent. This paper illustrates the research on grid technology. This study selected tool is the theme crawler algorithm.

1 Introduction

Grid (Grid) technology is a hot research topic in recent years the rise of technology. It is the application requirements and the product of technology development drive. E-Mail as the main application of the first generation of Internet spread all over the world of the computer using TCP/P agreement together. The second generation of Internet using Web browsing, and the application of electronic commerce information service. It implements the global web of Unicom. The third generation of Internet will attempt to achieve comprehensive Unicom all resources on the Internet, including computing resources, storage resources, communication resources, software resources, information resources and knowledge resources, etc. This is a grid.

On a search engine is a Web application software system, it on the Web in a certain strategy to collect and find information, and organization in the information processing, Web information query service for the users, provides users with all the information on the Internet resources retrieval means, to the user with the most comprehensive the most extensive search results.

Topic crawler search technology is a kind of purposeful crawling algorithm, avoid blind search inefficiency, is currently widely used a crawling algorithm, it intelligently search theme resources, get rid of the dependence on experts, to improve the efficiency of the theme resources construction and quality.

Grid computing research under the pilot Ian Foster for grid computing is defined as: grid computing is coordination in dynamic, multi-institutional virtual organizations sharing of resources and the process to solve the problem.

Service Grid is an important product direction, it uses web services and Grid computing technology, follow the Open Grid service Infrastructure, the Open Grid service Infrastructure Grid service standard, enterprise integration, support service connection, management, integration, optimization and operation of the service Grid will become commercial Grid system is an important development direction of it in order to achieve more enterprises or departments wide-area distributed business application integration and synergy between provides on-demand service, system interoperability and can monitor the strong support of the respect.

This paper is divided into different sections. Section 2 surveys the acceptance of new technology—the theoretical framework. The focus spider model is illustrated in Section 3. Section 4 we discuss the focus spider design and implementation. Lastly, conclusions are described in Section5.

2 Theoretical Framework

In under the attention and efforts of many scholars, grid technology is becoming a hot spot research, at the same time it step by step to start the commercialization process. Clear in order to realize the goal of grid computing and design products in Oracle 10g. This marks the industry efforts to commercialize the grid technology reached a new height. Researchers have already no longer satisfied with just the technical level to discuss the grid. Researchers have begun to design all kinds of grid technology application solutions. The researchers have also gradually to the market at the same time.

2.1 Grid Architecture.

Grid architecture is a technology about how to build the grid technology. It includes the grid basic component and function of each part of the definition and description, the definition and description of each part function, the relationship between the parts and the integration method of the provisions of the grid, effective operating mechanism.

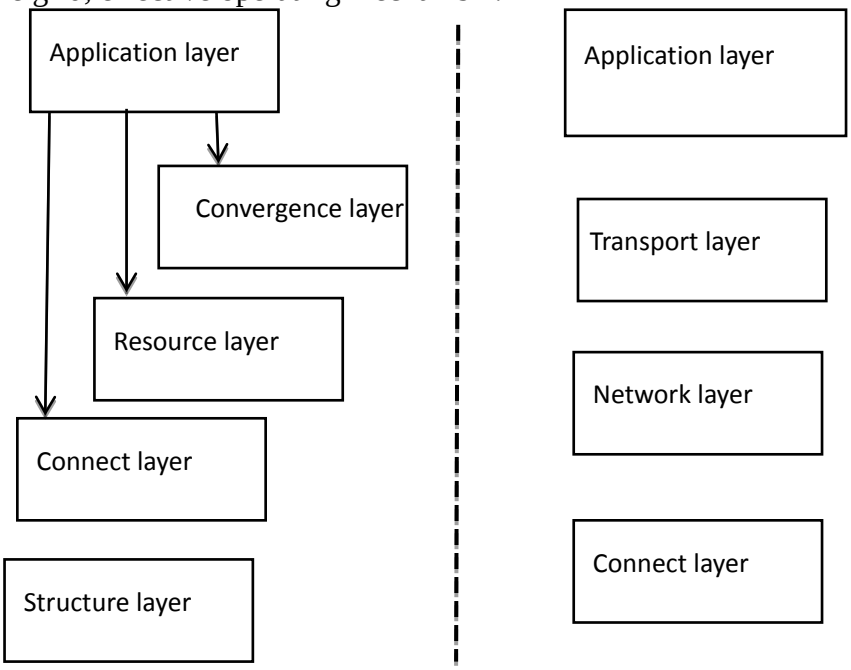

Fig.1 Contrast between Grid and Internet Architecture

The figure 1 describes the five layer sandglass model of Globus images, and put it with the Internet protocol model are compared.

2.2 The Theme Crawler System Structure.

The crawler is an information collection system. Download it via a Web page and crawl along has the hyperlink to iterate through the Web page. It collect Web pages, it is usually used in the search engine, as page collection system. It is usually in the form of breadth-first traversal web, makes every effort to crawl in a limited amount of the cycle to collect as many web pages.

3 The Theme Crawler ZT spider Analysis Algorithm

The goal of the theme crawler algorithm is to make the crawler to crawl the page relevant to the subject matter in proportion as much as possible. The goal is to make the crawling to minimize the proportion of unrelated pages. The article designs the topic crawler to crawl in the process of learning the ZT Spiders crawler algorithm. These include hypertext classification algorithm and hyperlink evaluation algorithm. The two algorithms are used Web page relevance analysis. The latter link needs the feedback computing to be done.

3.1 The Structure of Theme Crawler Algorithm ZT Spider.

Crawl in order to achieve the target, general topic crawler is through such a process. The users provide the target topic information and sample pages for the system. The corresponding URL will

serve as a system of seeds. Topic crawler by learning target subject information can be more accurate expression of the retrieval requirements of users.

The whole system includes three components: the first one includes the web crawler and the analyzers. Web crawlers select the URL of the highest priority. And then it through the agreement to download the corresponding Web pages. At the same time, its extract information, to make its line. The second includes hypertext classifier. It is to classify the sample pages. It will be the results of the classification will be passed to the hyperlink evaluation. The third includes Hyperlinks evaluation apparatus. It is used to evaluate candidate URL in the queue.

3.2 The Analysis Algorithm.

Correlation calculation method uses the algorithm that is the cosine of the Angle algorithm. Formula 1 as the CI has crawled page and sample classification SV calculation formula of correlation, correlation of CI SIM (CI).

$$SIM(C_i) = \frac{\sum_{k=1}^{M} C_{ki} \times \varphi_k}{\sqrt{(\sum_{k=1}^{M} C_{ki})(\sum_{k=1}^{M} \varphi_k)}} \qquad (1)$$

The purpose of our improved model is to efficiently harness the use of information and adjust their model parameters. The main body of this action is adjacent nodes outside the area. The improved algorithm based on the above considerations. The final result is shown in equation 2.

$$S(Ci) = SIM(Ci) + \sum_{j=1, ci \in d_k}^{M} \delta(C_i) U(l_s) \Big/ M \qquad (2)$$

4 The Focus Spider Design and Implementation.

System USES tools are GT4.This tool is currently widely used, but also the grid standards implementation tool. It is the foundation of an open source grid platform. It is based on open structure, open service resources, and library. It supports grid and grid applications. At this level GT4 as grid service container, it builds the crawler algorithm implementation services, distributed crawler system architecture.

4.1 The Module Division.

In the overall structure, topic crawler algorithm ZT spiders test system uses hierarchy structure. Its function of each layer is different. It contains three main layers:

① topic crawler algorithm implementation layer (Grid spiders service)

② the GT4 layer (Globus Container)

③ the storage layer (Storage)

Topic crawler algorithm ZT Spider test system structure diagram as shown in figure 2:

4.2 The Improvement.

The crawl the web in the process of the crawler to URL is exploding. To avoid this bad situation, we introduce the distribute crawler system structure. In the GT4 grid service container, we create and deploy a WEB service. Create an endpoint using the URI service, and define the type of the port, the finally, we start the grid application. The key code about this is shown as follows:

```
Endpoint Reference Type endpoint=new Endpoint Reference Type();
Endpoint . Set Address (new Address (service URL));
Spider service Addressing Locator locator
=new Spider service Addressing Locator ();
%GLOBUS_LOCATION % bin /Globus —start—container
Determine the task list and initialize
        From E0 to E1 into a list RISK
            Sequential scheduling over the list
        For each task Ti belong to the LIST do
```

```
        Sct cost _Min (e Ej) = infinity
      Sct   temp_j = Mji, temp _k =0
  For each or do
        If cop (arj) >= dem()
       Then
          Calculate the MYT A(Ei ,ARj)
          Compute shortest route Pk by OPSE
          Calculate the Cot ()
          If      Pre   (Ei ,ARj) + MYT (Qk) < Things _ Min()
          Then
            Set     Things _ Min() =Pre A (Ti ,ARj) + MYT (Pk)
            Set     temp _ j = j , temp _ k =k
           Else
              Continue
           Endif
Else
      Neglect the cost
End if
End for
      If     MYT _ min (tj) not belong to the infinity
      Then
         Scheduling E to Ar temp _ j
          Setup the light path temp _ j
        Else exit
        End if
     End for
```

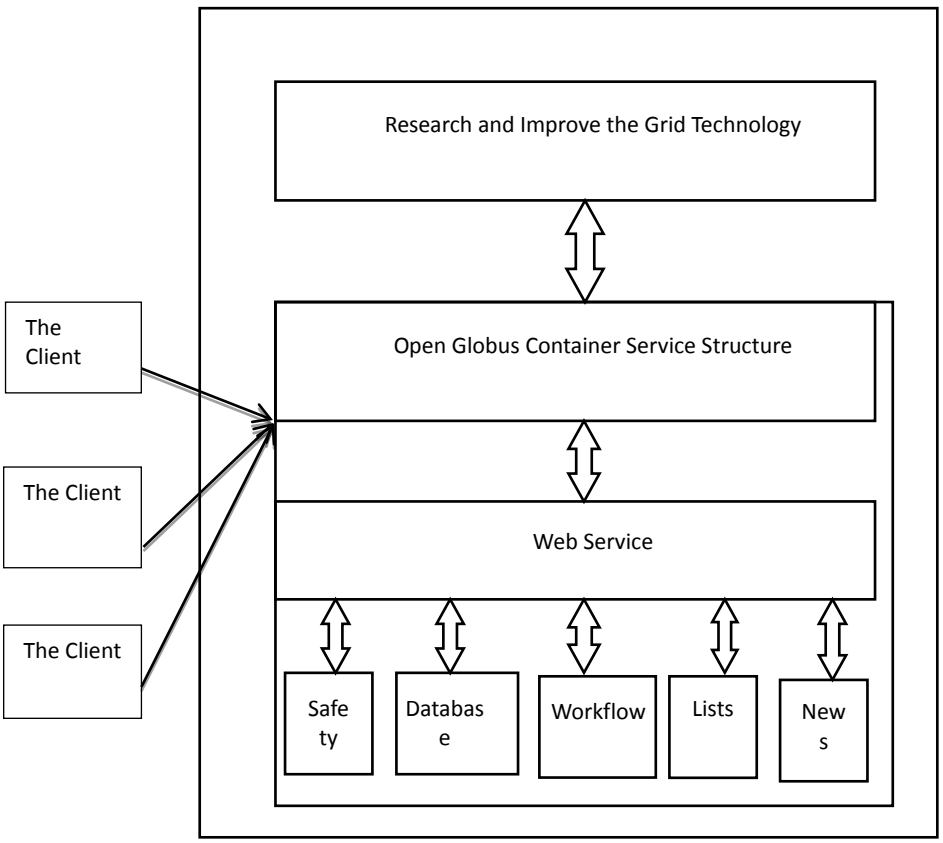

Fig.2 the Structure of the Algorithm

Then, we created and deployed a distributed grid services. This service is the method to use is the JAVA core services in the GT4.

5 Conclusions

This article mainly is divided into two parts of research work. The first part is the research topic crawler algorithm optimization. The second part is the study of information grid technology. The experimental work of the paper mainly includes two parts. The first part is to design and implement ZT Spider topic crawler algorithm. The second part is to build a distributed in the grid environment topic crawler system. Through distributed crawler system enables distributed in the grid nodes and super ZT spiders build communication mechanism. The purpose is to strengthen the effective information interaction, using the correlation analysis results with each other. Finally, we complete the Super ZT Spider depth feedback algorithm. The aim is to avoid a single Main Spider learning process into a local optimum.

References

[1] Hamilton BA. Understanding the benefits of the Grid – Grid implementation strategy. In: Hamilton BA, Miller J, Renz B, editors. United States: United States Department of Energy's National Energy Technology Laboratory; 2010.

[2] Moura PS, de Almeida AT. The role of demand-side management in the grid integration of wind power . Appl Energy 2010(87):2581–2588.

[3] Foster I，Kesselman C，Tueck S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations [J]. Super- computer Applications, 2001, 15 (3).

[4] R. J. Allan. A Globus Developers, Guide with Installation and Maintenance Hints, 2001.

[5] A Globus Toolkit Primer. Describing Globus Toolkit Version 4,2005.

[6] Yu C-Y, Muthén B. Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA; 2002.

[7] Henderson R, Divett MJ. Perceived usefulness, ease of use and electronic supermarket use. Int J Hum Comput Stud 2003;59:383–95.

[8] Holden RJ, Karsh B-T. The Technology Acceptance Model: Its past and its future in health care. J Biomed Inform 2010;43:159–72 .

[9] Broman Toft M, Schuitema G, Thøgersen J. The importance of framing for consumer acceptance of the Smart Grid: A comparative study of Denmark, Norway and Switzerland. Energ Res Soc Sci 2014;3:113–23.