Research on the Database Process Using GRID Technologies

Yang Yang

Assets Management Department, Shandong Women's University, Jinan, 250000, China

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

Abstract. Grid technology is widely used in recent years, it has solved the real sharing of resources, make each node of unified command and use of resources, the information grid is on the basis of computational grid, using data mining. Information fusion and search engine technology and building are advantageous for the collection and sharing of grid resources. The goal to create a build on OS and Web is based on the new generation of Internet information platform. In this platform, the distribution of the information processing is the collaboration and intelligent. Focus crawler sets according to the target subject. In the form of intelligent collection topics come from the Web page, and then adopt the method of machine learning. The information retrieval on the collected information intelligent need process and analyze. Finally, the effective retrieval way is to meet the demand of the user's information retrieval.

Introduction

Grid technology is a hot research topic in recent years the rise of technology. Grid as the third generation of the Internet has caused an unprecedented worldwide concern and attention. Grid is used to solve the real sharing of resources. It makes each node of a large number of idle computing resources. And storage resources are unified command and used to user. Grid is also trying to solve the problem of information island in all organic links on the Web server, no longer requires users to searching and sorting of useful information you need.

Grid is not to architecture, but an extension on the basis of the existing network. On a search engine is a Web application software system, it on the Web in a certain strategy to collect and find information, after the process and organization of information, to provide users with Web information services, provide users with the means to retrieve all information resources on the Internet, to the user with the most comprehensive the most extensive search results. Topic crawler search technology is a kind of purposeful crawling algorithm. To avoid blind search inefficiency, is currently widely used a crawling algorithm. It intelligently searches theme resources, gets rid of the dependence on experts, improves the theme resources construction.

The open grid service infrastructure is web-based service standards. It is a basic standard, is put forward by enterprise and the research community to work together, purpose is to realize the grid service. The open grid services architecture describes the concept of formal specification. The open grid service infrastructure includes how to manage tasks, assign tasks and how to describe the service provider and the specification of grid service. Web service is an important part of this specification, including the simple object access protocol (soap) and Web services description language.

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.

Theoretical Framework

Grid technology, after become a research hotspot in the attention of many enterprises and efforts, gradually started the commercialization process. Clear in order to realize the goal of grid computing and design products. This commodity is oracle109.Its emergence, marking the industry efforts to commercialize the grid technology reached a new height. Companies are no longer satisfied with just to discuss the grid technology, all kinds of grid technology application solutions have also been gradually to the market. Globally, IBM, Oracle, Microsoft are representative system provider. Sun, EMC, HP, Intel are on behalf of the architecture of the provider. Platform, Avaki is representative of the middleware and application providers. These are all for grid development of corresponding soft wares.

Grid Architecture.

Grid architecture is a technology about how to build the grid technology. Five layer sandglass structure of one of the most important thought is centered on agreement. It also emphasized the importance of service and API and SDK. It is similar to the traditional TCP/IP network protocol stack. The grid structure is divided into five interconnected and not equal level. In five levels, resource layer and the layer greatly extended the function of network application layer. This method is the encapsulation of traditional network application layer issues. In short, the function of the structure is becoming more powerful, and provide users with more transparent the use of the method.

Building block structure is actually a component structure. It is emphasized in the computational grid of different function modules are relatively independent, and connect with each other.

Compared with the above two kinds of system structure, the level of the architecture characteristics of concept space is not clear. It emphasizes the individual parts on the concept of correlation, agent technology and object-oriented technology is the main method to realize concepts associated.

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

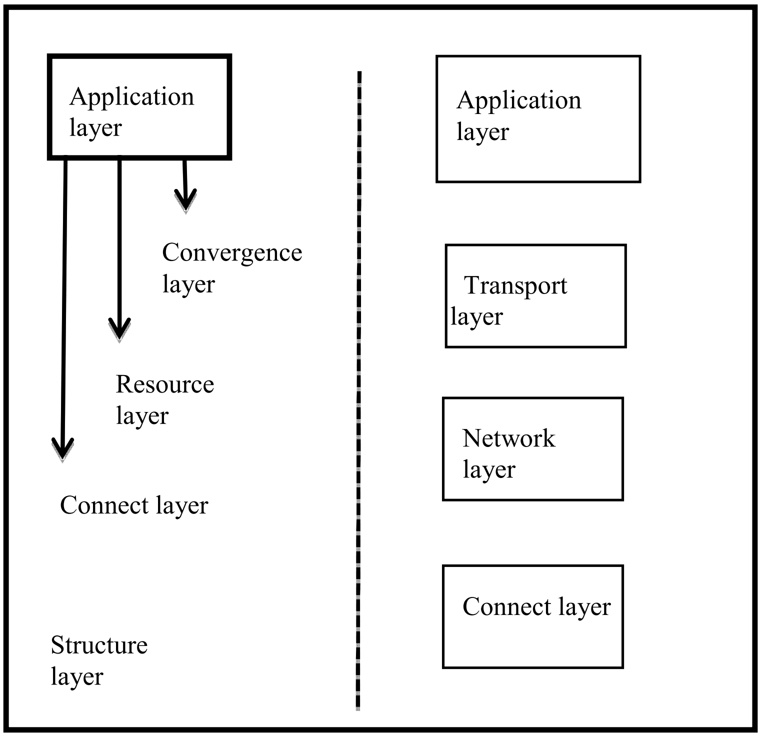

Fig 1 the contrast between the grid and the structure

Globus software as a grid, Globus Toolkit contains a set of implementation of the security, resource location, resource management, communication and other core service module, these modules in the form of building blocks to construct a system of grid computing.

Globus Toolkit of the relationship between each service component as shown in figure 2. In the five layer sandglass, each layer corresponding to different components, but the boundaries between layers are not obvious.

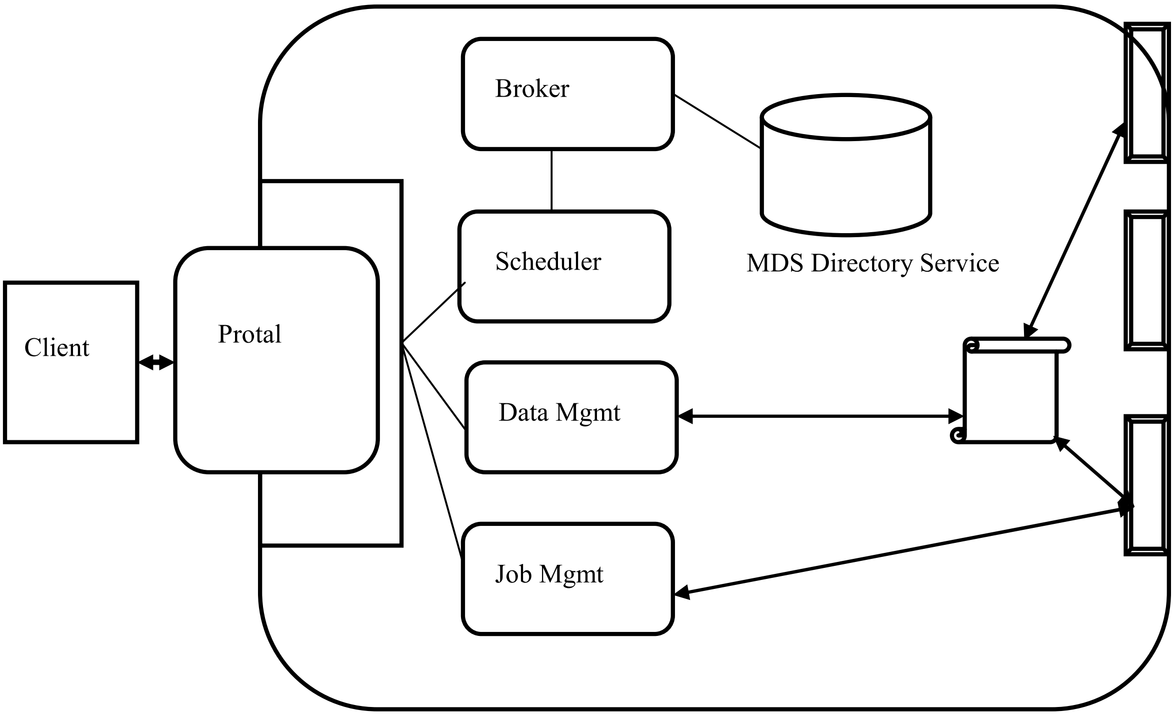

Fig.2 Grid Architecture in the Core Component

The System Structure In the Grid Technology.

All cores are distributed computing and grid resource management. A grid of a large number is heterogeneous resources, these resources are known and consistent way to interact and work. It is this interaction and interoperability between different resources provided by the components, led to the need of open standards, and promote the emergence of OGSA.

The open grid service infrastructure specification is a web-based service standard, working together by enterprises and research put forward the basic standard of grid service. The open grid service infrastructure specification is described the concept of OGSA formal specification. The open grid service infrastructure specification contains how to management tasks, assign tasks and how to describe the service provider and the specification of grid service. Web services, especially the simple object access protocol and Web services description language is an important part of this specification.

The Analysis Algorithm

The crawler is an information collection system, downloading it via a Web page, and crawling along the page links traversal Web. Collecting Web pages, it is usually used in the search engine, as page collection system. It’s usually in the form of breadth-first traversal web. It makes every effort to crawl in a limited amount of the cycle to collect as many web pages. Contrary to the crawler, topic crawler is selective when crawling, its aim is not to choose the page as much as possible, but it makes the page collection as much as possible topic, also is the biggest makes precision, precision of the calculation method is to look at the crawling to the percentage of page relevant to the subject matter of your web page.

The Structure of Algorithm.

Info-Spider is a topic crawler based on neural network. It by extracting the web hyperlink anchor text as the input of the neural network. It trained neural network to determine the selection strategy of hyperlinks. At the same time it has crawled web relevance as a further training feedback neural network. Mencze will be compared with Best-First algorithms, this method shows that Info-Spider has better performance.

Cora search engine, search engine is a computer science paper. It goes through topic crawler from computer science department of each web site to collect computer science papers. It is the theme of the crawler Q - learning to enhance learning algorithm. This algorithm through the mapping between text and Q value will return into classification problems. The experiment proved that by this method can crawl to a large number of computer science papers. The main advantage of this method is that it can in the short term of the expanding of hyperlinks reasonable compromise between returns and long-term returns. But it requires an off line training process, need artificial identification of a large number of training samples, the lack of online learning ability.

The Analysis Algorithm.

The crux of the problem is how to make a hypertext classifier and hyperlink evaluation can focus on learning from crawling to web pages, make the classification model and evaluation model of hyperlinks to update. Formula 1 as the SE has crawled page and sample classification CRcalculation formula of correlation, correlation of SE QWA (SE).

$$SIM(S_i) = \frac{\sum_{l=1}^{N} S_{li} \times \beta_k}{\sqrt{(\sum_{l=1}^{N} S_{li})(\sum_{l=1}^{N} \beta_k)}}$$

The Focus Spider Design and Implementation.

Super-ZT-Spider actually is also a topic crawler, its task is different from other theme crawler.It USES a certain interval Main Spider sent to crawl the result data, integrated with the method of the formula shown in calculation. It then feedback adjustment parameter, the result of the multiple crawler to crawl equalization processing, to achieve the purpose of deep feedback relevant information. After these it will recount the relevance of information is returned to each Main Spider, guiding the Main Spider crawling behind.

Globus Toolkits used to own tools to service the binding and produce customer need the root class, then is to provide the implementation of the interface. Spiders are given Service Impl class, mainly realize the port type of interface, the implementation should use with spiders. In front of the WSDL file generated by the root class, defined as shown below. 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:

```
Public class Spider service impl implements spider port type
Public Create spider response create spider ( Create spider cs)
Throws java. Remote exception
Public process spider response process spider (process spider ps)
Throws java rmi remote exception
Public get spider response get spider
Endpoint . Set Address (new Address (service URL));
Value =’’org.globus.axis.providers.RPCProvisder’’/>
   </service>
Spider service Addressing Locator locator
=new Spider service Addressing Locator ();
```

When the program found the new link, the first to join them is waiting in the queue. It considering the program may need to crawl web pages is heavy, so the queue using the database. Program, a total of four queues, as shown in figure 4.3.

Waiting queue: URL in the queue is waiting by the program.

Processing queue: program starts when processing the URL, they in the queue.

Error queue: there was an error in the web page download, its URL will be added to the wrong queue, and the URL won't move into other queue.

Complete queue: download page is no error, will be completed U left to join the queue.

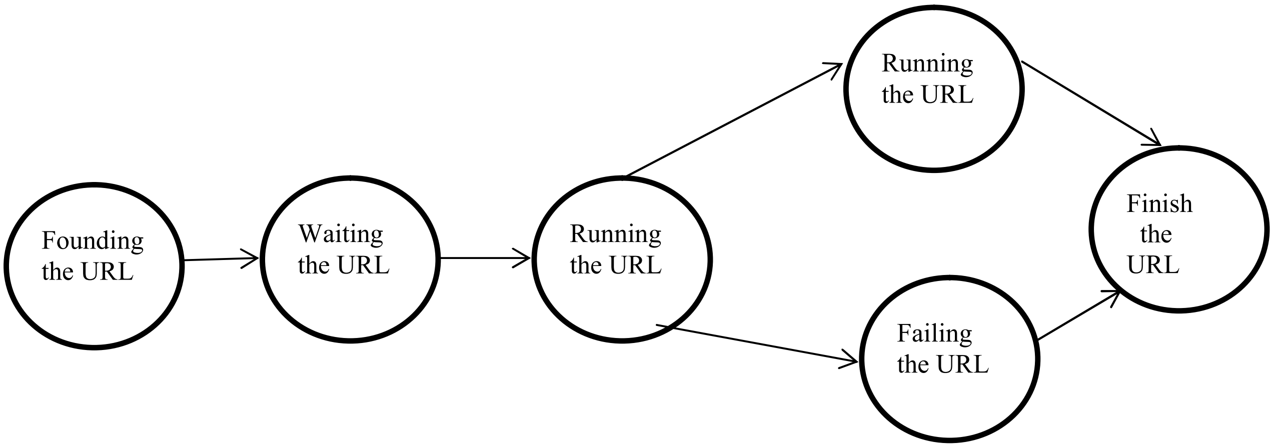

Fig.3 The Design of Queuein ZTSPider

Conclusions

ZT-Spider topic crawler algorithm mainly includes the web page analyzer and hyperlink correlation evaluation. Improve current topic crawler algorithm that enables the crawler to online learning, and through the web page of the parent and child nodes correlation study. It implements dynamic feedback of adjacent nodes correlation information, strengthen the classification of web page classification effect.

Distributed crawler system enables distributed in the grid node of each Main spider through and super ZT-Spider communication mechanism, strengthen the effective information interaction. To learn from each other correlation analysis results, and completed by super ZT-Spider depth feedback algorithm, 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] Bolderdijk JW, Steg L, Geller ES, Lehman PK, Postmes T. Comparing the effectiveness of monetary versus moral motives in environmental campaigning. Nat Clim Change 2013;3:413–6.

[7] Throne-Holst H, Strandbakken P, Stø E. Identification of households' barriers to energy saving solutions. Manage Environ Qual: An Int J 2008;19:55–66.

[8] 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.