RETRIEVAL ON THE GRID. RESULTS FROM THE EUROPEAN PROJECT GRACE (GRID SEARCH AND CATEGORIZATION ENGINE) [1]

FRANK SCHOLZE, WERNER STEPHAN

ABSTRACT

GRACE (Grid Search and Categorization Engine http://www.grace-ist.org/) has been a project funded by the European Community under the Information Society Technology Programme 2002 Framework 5. It developed a distributed parallel search engine based on Grid Middleware. Search results from heterogeneous sources are categorized using key phrases extracted from the documents retrieved. Currently available Grid Middleware employed in the project (Large Hadron Collider Computer Group (LCG-2) und European DataGrid (EDG)) proved to be insufficient for interactive real time applications such as search engines. Therefore a combination of ontology based indexing (Grid based pre-retrieval indexing) was combined with categorization in order to provide quick response times. The paper gives a short introduction to the Grid technologies and the natural language processing tools used before it presents the results of the project.

INTRODUCTION

Internet computing and Grid-technologies promise to change the way we tackle complex problems. They will enable large-scale aggregation and sharing of computational resources, data and other resources across institutional and geographical boundaries.

Internet computing is just a special case of something much more powerful: the ability for communities to share resources as they tackle common goals. Business today is increasingly international and multidisciplinary. It is not unusual for corporations to span states, countries and continents. It is also not unusual for corporations to bundle together a variety of industries and to collect information and generate expertise in various areas of business, technology and science. E-mail and the World Wide Web provide basic mechanisms that allow such groups to work together. But what if they could link their data, computers and other resources into a single virtual office? So-called Grid-technologies seek to make this possible by providing the protocols, services and software development kits needed to enable flexible, controlled resource sharing on a large scale.

The World Wide Web has facilitated unprecedented ways of speedy global information sharing. The Grid-technology will build on this by allowing facilitating the global sharing of not just information but of tangible assets to be used at a distance. Very large databases - literarily terabytes and petabytes of information - that now are geographically confined will become Grid-sharable. This is why - in addition to the computational Grid technology recent efforts are directed into developing data-Grid infrastructures [2].

PROJECT GRACE

GRACE [3] – Grid Retrieval and Categorization Engine – is an information retrieval technology. Semantic Information Retrieval involves computationally intensive tasks that require extensive computational and storage resources. GRACE undertook the task of investigating whether the current Grid technologies can offer a practical solution to this demand for extensive computational and storage resources. GRACE is the first practical GRID application, designed as a comprehensive retrieval system, tailored specifically to the needs of researchers in any field. It is a unique solution which introduces the concept of knowledge domains, federated search and distributed processing to information retrieval. It collects, categorizes and presents information to the researcher through a simple and user friendly interface. The knowledge domains are defined within loadable ontologies in the form of domain specific Thesauri.

GRACE is based on the ability of distributed systems such as the Grid to offer controlled and authorized sharing of resources – computational, storage, human. This is precisely what the concept of knowledge discovery makes possible in the domain of the information retrieval (IR). It offers appealing and relevant content (documentation, knowledge basis, etc.), but also allows dynamic document publishing and storage. The solution combines a document management system found on many intranets with relevant content sources.

An important result has been gaining the experience of which parts of the process of the information retrieval workflow could benefit from being executed on the Grid. During the implementation it became evident that the grid middleware used as the platform provides a response time that is not suitable for the interaction with the end-user. Accordingly, the implementation took the direction of using the grid for computationally heavy pre-processing tasks.

Information retrieval on the whole involves many computationally intensive tasks that require extensive computational and storage resources. These computationally intensive tasks are typically performed as some kind of "text crunching" that may include various aspects of natural language processing, and are designed to transform the original document text into indices optimized for efficient querying. However, the "text crunching" performed by GRACE is even more complex and resource demanding. This is due to the advanced natural language processing functionalities offered by GRACE: categorization, named entity extraction, and concept indexing. Consequently, GRACE undertook the task of investigating whether the current Grid technologies can offer a practical solution to this demand for extensive computational and storage resources.

Accordingly, GRACE combines three innovative technologies:
1. Federated search technology to access multiple, distributed content sources in parallel,
2. Natural language processing technology for highly advanced text indexing, and
3. Grid technology for execution of computationally intensive indexing tasks through flexible, just-in-time resource sharing.

The GRACE Toolkit prototype was integrated with the Gilda testbed [4] of the Large Hadron Collider (LHC) Computer Group (LCG) that maintains currently the largest data grid worldwide. LCG data grid uses a version of the European DataGrid (EDG) middleware. Gilda is available to other European projects through the courtesy of Istituto Nazionale di Fisica Nucleare (INFN - The Italian National Institute for Nuclear Physics) that is responsible for its maintenance.

The project duration was 30 months and it was led by Telecom Italia, one of the largest telecommunications operators in Europe.

GRID UNDERSTANDING

Grid is currently one of the most promising emerging technologies. It is designed to allow performing computationally intensive tasks with only limited resources by using additional computational resources that are temporarily made available on the Grid by members of a "virtual organization". In a "virtual organization" all members contribute their limited resources for the benefits of all partners. In this respect Grid is actually based on the timesharing of computational resources within a "virtual organization". This makes Grid in particular useful for collaboration between geographically wide-spread partners who can use all computational resources available on the Grid during their work hours while other partners don't work.

A special case of this scenario is the Data-GRID. Data-GRIDs are designed to share not only the computational resources, but also the storage space and the content. It is hereby assumed that the data stored at different places on the GRID, will be used by all partners in a "virtual organization".

Data-Grid-technologies address the problem of efficient storage of petabytes of data (typically generated by scientific experimental instrumentation) by distributing them across a Grid-network. Although due to their enormous size these data cannot be stored in any single central location, they are still required to be highly accessible. Data Grid-technologies attempt to coherently manage these petabytes of distributed data in order to enable their speedy and efficient manipulation.

INFORMATION RETRIEVAL

However promising the Grid technology seems to be, the question is whether it might be used for information retrieval applications. Information retrieval involves many computationally intensive tasks that require extensive resources. By information retrieval we understand retrieval of unstructured textual information, usually stored in various document formats. These computationally intensive tasks are typically performed as some kind of "text crunching" that may include various aspects of natural language processing, and are designed to transform the original document text into indices optimized for efficient querying. The resulting indices are also of significant size, and may be of the same order of magnitude as the original processed text, or even higher. This combination of the computationally intensive tasks with requirement for extensive storage space makes the Grid, and in particular the Data-Grid attractive.

REQUIREMENTS FROM GRID TO SUPPORT
AN INFORMATION RETRIEVAL APPLICATION

We will describe briefly some minimal Requirements which we believe must be satisfied, in order to ensure that Grid solutions indeed can improve information retrieval. The best approach to presenting these expectations is by analyzing a typical lifecycle of a Grid job.

Here are the major phases:
1. Grid job submission, certification and resources allocation
2. Upload of the computing application and the data
3. Queuing in the local queue

4. Initiation of the computing application
5. Processing
6. Monitoring
7. Download of the results

It is evident that phases 2, 3, and 4 could be omitted by running the processing application as a service on the targeted work node. Although running a service on the resources that belong to another member for the virtual organization seems to be primarily an administrative issue, it may also influence certification and the efficiency of resources allocation. The allocation of resources must guarantee that the required service is installed and indeed running on the targeted work node prior to launching the Grid job.

The certification on the Grid is already awkward, since it is repeatedly performed at every step of the Grid job submission. It would be a general expectation from the emerging Grid technologies to simplify the certification process significantly, i.e. to make it more similar to the Web Services certification. On the other hand, the members of the relevant virtual community will be forced to handle a stricter but simplified certification procedure. Another issue related to the phase two is the upload of the data that are to be processed. The design of Data Grids allows these data to be uploaded in advance, and than used by the computing applications as needed. However, this may be not always desirable for the information retrieval.

In the context of information retrieval this data is typically the original text of an document. After the text crunching that is performed on the data, the original document text is of little importance, and does not need to be stored on the GRID any longer. This may suggest to prefer the upload of data as a part of the Grid job submission. For the scheduled Grid jobs the resulting time overhead during the Grid job submission is in any case of little importance. Nevertheless, it may be more efficient to separate completely these two tasks, and to pre-load the data into a temporary storage on the GRID. In this case the failing GRID jobs would not need to reload the data with every additional resubmission. Pre-loading of the data would also be more suitable if the processing application is run as a service that is than invoked without any additional uploads. It would, however, require an efficient system of data management that insures that the remaining data residues are eventually removed, and that the storage space can be reclaimed.

Although this issue seems to be more of concern to the efficient design of an application that runs on the GRID, it would be preferable if such storage management functionality would be offered directly by the Data GRID middleware, and handled seamlessly by merely replacing the original data with the processed data. If the processing application

is run as a service this would merely require invoking the desired service while pointing to the data stored on the GRID, which would be than automatically replaced by the processing results upon the successful completion of the GRID job. This scenario would completely eliminate the phase 7 of the GRID job submission.

The current GRID job failure rate does not seem to be of much concern for the GRID users. Nevertheless, it is of critical importance to timely detect such failures in order to handle them. For this, an interactive monitoring ability is required from the GRID. This would significantly improve the Phase 6 of the GRID job submission. Therefore, the greatest hopes from the emerging WSRF (Web Service Resource Framework) technology are that it will allow interactive messaging.

Besides monitoring, interactive messaging can improve significantly the allocation of resources by allowing their interactive inspection and selection. This could improve the phase 1 of the GRID job submission, and lower the overhead in time required. In future it is possible to envision even web services-resources orchestration that may be extremely interesting for the information retrieval. For example, various text crunching tasks could be exposed as services on various web services-resources, and different combinations of these tasks could be invoked in order to achieve results satisfying different needs.

The final requirement from the Data GRID (that is not related to the GRID job submission) is to simplify the system installation and configuration. Installation and configuration of current Data GRIDS seem to be extremely difficult, most probably due to the fact that they consist of different layers developed by various producers.

For example, a GRID testbed like GILDA developed and maintained by INFN uses Globus Toolkit 2 (GT2) as the basic middleware. On top of GT2 runs European Data Grid 2 (EDG2) that provides basic Data Grid functionalities, and on top of EDG2 there is still a Large Hadron Collider (LHC) Computing Group 2 (LCG2) installation that uses a specific configuration of EDG2 functionalities. It is obvious that this kind of configuration is extremely difficult to manage and maintain, and that a simpler middleware should be offered for the Data GRIDs.

KNOWLEDGE DOMAIN ON THE GRID

In order to access multiple, distributed content sources with the architecture first envisioned in the project it would be necessary to keep an index for each of them. It would be also necessary to perform some kind of post-retrieval processing in order to integrate content sources that are inaccessible except by querying them.

The concept of Knowledge Domain efficiently solves both of these problems. By eliminating the need for post-retrieval processing, and completely transferring the weight to pre-indexing of the information, Knowledge Domain efficiently harnesses the advantages of the Grid to the full extent. This is why ontologies are of critical importance for the Information Retrieval on the Grid, and must be, consequently, an integral part of the architectural design.

Nevertheless, ontologies are also used for querying. This approach employed in project GRACE is called concept based indexing which is similar to manual keyword annotation of documents but runs as an automatic process on paragraph level. Existing ontologies have to be used. Keeping them on the Grid necessitates the best possible response time in order to ensure that the retrieval of the search results matches the current Information Retrieval standards.

In GRACE the High Energy Physics Keyword Index [5] was chosen as an example. It has some basic features like related terms, broader terms etc. which give HEP qualities of an ontology. HEP Index terms were used for batch querying Cern Document Server, Google Scholar, Scirus and ArXiv. Batch querying allows pre-processing of document text and is suitable for the Grid. Batch querying results in automatically tagged document collection relevant for high energy physics. Batch queries can be re-run on schedule to keep the concept based index up-to-date.
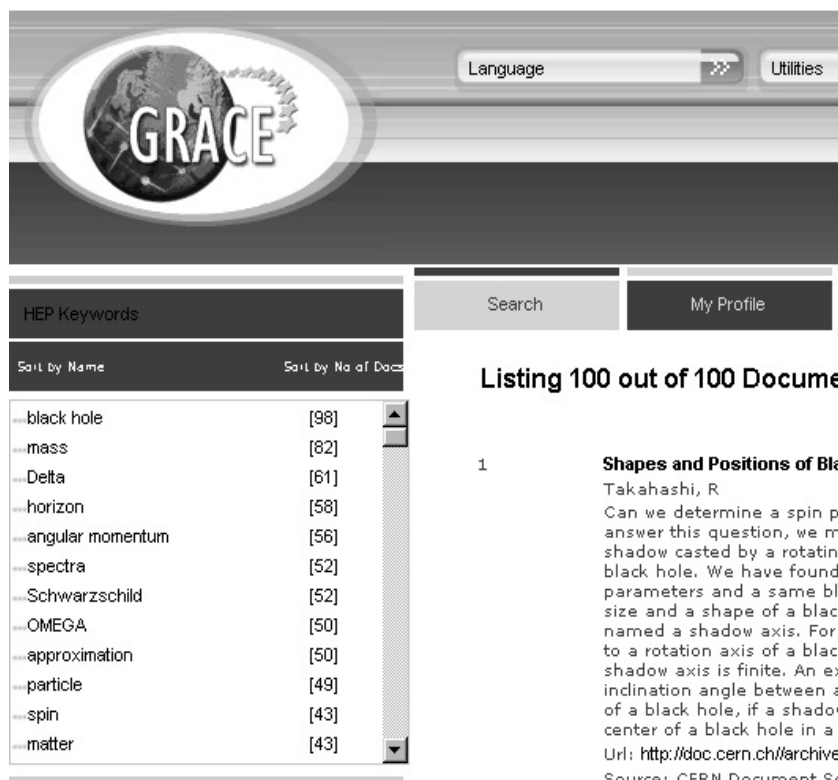

Figure 1: Results categorized with HEP terms

Queries posted by users are now run against the pre-indexed documents yielding immediate results categorized by using HEP terms (Figure 1). In the background grid jobs are now evoked which process the documents retrieved as outlined above. Thus additional key phrases are extracted and the documents are categorized additionally (Figure 2).

Figure 2: Results categorized with HEP terms (upper left) and key phrases extracted from the documents retrieved (lower left)

The most essential condition is to run the querying mechanism as a service on the Grid. This condition omits the downloading and the initiation of such mechanism – two operations that can add significant time overhead to the submission of a Grid job. The concept of the Knowledge Domain that implies strong consensus among the members of a Virtual Organization can efficiently solve the administrative aspects necessary for this condition.

The strong consensus among the members of a Virtual Organization could also allow simplification of the certification process, and thus further improve the procedure of the Grid job submission. How this simplification of certification could be accomplished is still an open research topic. It would be also required to research the scenario in which the certification mechanism could be used for billing according to the use the Information Retrieval system on the Grid.

The final obstacle to the rapid Grid job submission is the allocations of the Grid work node on which the querying Grid job would be launched. Fixing the number of Grid work nodes on which this particular service will run could efficiently solve this. Instead of submitting the querying Grid job to the Resource Broker module of the Data Grid, it would be submitted directly to the Grid work node on which the querying mechanism runs, and thus completely omit the resource allocation.

An even better solution could be offered through the use of web services-resources interface. A user could simply invoke a web services-resource for querying, and the web services-resource would then seamlessly interact with the work node running the querying service.

CONCLUSION

The original view of GRACE was that it would be an interactive search and retrieval engine. During the course of the project much was learned by the consortium of the operating characteristics of the Grid as it is implemented in this instance. The view that the Grid will be the successor to the web as the internet infrastructure is a long way from realization and the limitations of the present iteration became apparent during the project.

Where GRACE originally envisaged a real-time interaction between the user and the data sources it became apparent that the grid response times did not allow for this. Grid is optimized for batch submission of multiple jobs and the GRACE architecture and operation were redesigned to cope with this characteristic. For this reason GRACE became a retrieval and categorization technology with pre-processing and indexing of queries. By this method of concept based indexing it became possible to give the end user a rapid response to the initial query, using pre-indexed documents, whilst the search and categorization was submitted and in process in the background.

This paper outlined a possible architecture of an Information Retrieval system based on the concept of Knowledge Domains that aims at efficiently overcoming the obstacles of present Grid implementations, and thus allowing the Information Retrieval to fully harness the power of Grid, without compromising its efficiency.

Frank Scholze, Werner Stephan

REMARKS

1   Main parts of this paper are based on the internal GRACE position paper - Information Retrieval on the Grid. The authors want to thank especially Nahum Korda from GL 2006, the author of the position paper. An extended version of the paper has been presented at the IATUL 2005 conference.

2   The most outstanding example is the European Data Grid project (EDG; http://web. datagrid.cnr.it) and its successor EGEE (Enabling Grids for E-science in Europe) headed by CERN and founded by the European Commission (http://public.eu-egee.org/).

3   http://www.grace-ist.org/

4   https://gilda.ct.infn.it/

5   http://www-library.desy.de/schlagw2.html

REFERENCES

Maurizio Cecchi, A Contribution on Knowledge Management on Grid. - http://www.grace-ist.org/docs/GGF-KM%20on%20Grid.pdf

Jawed Siddiqi, Requirements for Knowledge Discovery within the Grid Space - http://www.grace-ist.org/docs/GGF-knowledge%20discovery%20reqs.pdf

Nahum Korda, GRACE: Lessons Learned - In: GGF12 - The Twelfth Global Grid Forum September 20-23, 2004 Brussels, Belgium http://www.grace-ist.org/docs/GGF-Lessons%20learned.pdf

Frank Scholze, Glenn Haya, Jens Vigen, Petra Prazak, Project GRACE - A grid based search tool for the global digital library. - In: 7th International Conference on Electronic Theses and Dissertations, June 3-5, 2004,University of Kentucky, Lexington, KY USA http://www.uky.edu/ETD/ETD2004/scholze/etd2004.ppt

Glenn Haya, Frank Scholze, Jens Vigen, GRACE - Developing a Grid-Based Search and Categorization Tool. - In: HEP Libraries Webzine Issue 8 / October 2003 http://library.cern.ch/HEPLW/8/papers/1/

Jawed Siddiqi , Babak Akhgar and Mehrdad Naderi Towards a Grid Enabled Knowledge Management Services. - In: Proceedings of UK e-Science All Hands Meeting 2003, 2-4th September, Nottingham, UK http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/073.pdf

ADDRESS OF THE AUTHORS

Dipl.-Bibl. Frank Scholze M.A., Werner Stephan M.A.

Universitätsbibliothek Stuttgart

D-70174 Stuttgart, Holzgartenstraße 16

E-Mail: scholze@ub.uni-stuttgart.de, werner.stephan@ub.uni-stuttgart.de

Web: http://www.ub.uni-stuttgart.de/fachinfo/personen/scholze.phtml