LIBRARIES WEBZINE *High Energy Physics Libraries Webzine*

Issue 8 / October 2003

Developing a Grid-Based Search and Categorization Tool

Glenn Haya (*), Frank Scholze (*), Jens Vigen (*)

Abstract:

Grid technology has the potential to improve the accessibility of digital libraries. The participants in Project GRACE (Grid Search And Categorization Engine) are in the process of developing a search engine that will allow users to search through heterogeneous resources stored in geographically distributed digital collections. What differentiates this project from current search tools is that GRACE will be run on the European Data Grid, a large distributed network, and will not have a single centralized index as current web search engines do. In some cases, the distributed approach offers advantages over the centralized approach since it is more scalable, can be used on otherwise inaccessible material, and can provide advanced search options customized for each data source.

Introduction

Today's search engines are extremely centralized. In order to index a document it must be downloaded, processed and its index stored - all in one central location. However the centralized approach may not always be applicable. GRACE (Grid Retrieval And Categorization Engine) specifically addresses the situations where centralized indexes are unfeasible and proposes the development of a decentralized search and categorization engine built on top of grid technology.

The goal of the project is to build a tool capable of searching and automatically categorizing vast amounts of geographically distributed information, which in many cases could not be searched effectively, if at all, by a centralized search engine such as those available on the web. The types of documents that GRACE will process include unstructured text and heterogeneously structured text in the form of text files, web pages, and text stored in databases.

GRACE is a project in the Fifth Framework Program (FP5) of the Information Society Technologies (IST) initiative by the European Union. The partners in this project are Telecom Italia Lab (as project leader and manager), CERN (European Organization for Nuclear Research), Virtual Self, Sheffield Hallam University - School of Computing and Management Sciences, Stockholm University Library, and Stuttgart University Library. The project started in September 2002 and will end in February 2005.

1 What is Grid Technology

Grid technology has its roots in distributed computing. Distributed computing developed in an effort to generate processing power for meeting workload challenges. In order to boost processing power, institutions aggregated computing resources across locations or across the entire institution. The idea was to match the supply of processing cycles with the demand created by applications. This concept is now a ubiquitous solution practised by leading organizations around the world. It ensures continuous computing availability despite scheduled maintenance, power outages, and unexpected failures.

The same idea of sharing resources has paved the way for grid computing - a term coined in the mid-1990s. The grid approach uses untapped processing cycles from across geographical boundaries, much like distributed computing but with a far wider scale and scope. Grid computing, in effect, provides a global reach to distributed computing. It promises lower total computing costs along with on-demand, reliable, and inexpensive access to the vast, available computing resources that would otherwise go unused [1].

Grid technology can be divided into three groups. The first is data grids, which are grid environments designed to process huge data samples. An example of this is a particle physics experiment that requires the analysis of millions of particle collisions. The second group is the computing grid where the focus is on the execution of parallel algorithms rather than on the distributed processing of huge amounts of data. Finally, the concept of an application grid addresses the set of services available on sites distributed throughout the grid. The application grid is the primary focus of the GRACE project.

2 When is a Decentralized Search Engine Preferable to a Centralized Engine?

Centralized search engines such as those commonly used on the web can be effective. However, there are circumstances, such as those listed below, where a decentralized search engine such as GRACE is superior to the centralized alternative.

2.1 For material not accessible by a centralized search engine

Much has been written about the vast amount of material in the "invisible web", also called the "deep web", which cannot be searched by traditional search engines.

A decentralized search engine can potentially provide access to many of the sources that cannot be reached by a centralized search tool such as a typical web search engine. Below are three examples.

Problem	**How a decentralized search engine would help**
Dynamically generated text - e.g. from databases.	A decentralized search engine could search the database directly by interfacing with a local search engine.
Text to which the public is not meant to have access and so is not made available to centralized	Grid services provide a security system that allows authorized users to access material and

search engines on the web.	turns away unauthorized users. This enables the searching of sources that would not be made available for searching by a web search engine for security reasons.
Variations in links to documents: Centralized search engines index new material by crawling links. If a document is disconnected from all known pages or if it is linked in an unusual way such as with Javascript links, then it may never be picked up by the search engine's crawler.	Again, by interfacing directly with the local source, the decentralized search engine can cover this type of material which might never be picked up by web-based search tools with a centralized index.

2.2 When the scale is enormous

The decentralized search engine relies on processing power distributed over the network. As a result it is much more scalable than a purely centralized alternative since it is not limited to a single physical site with a limited number of servers.

Observers and researchers in the field of digital libraries have acknowledged that scalability is a challenge for digital libraries presently and in the future [2, 3, 4]. Researchers working with the Alexandria Digital Library Project identified the following four primary scalability issues related to digital libraries [4].

1. The amount of data itself is increasing in scale
2. The dimensionality of the data is increasing
3. Individual information sources contain very large collections
4. The size of individual information objects themselves is getting larger

These increases in scale create challenges that can be addressed by scalable computing and storage resources. This scalability of resources is one of the advantages gained through a grid-based search and retrieval system such as GRACE. Since both the computing and storage resources are distributed, the system can accommodate the increasing scale of information on all four of the axes described above. The same results would be difficult to achieve with a centralized search system.

2.3 When the frequency of updates is high

Centralized web search engines do not search the web in real time. Instead, they search through a cached full-text index which is often updated rather slowly. Google, for example, updates its index roughly once a month [5]. A decentralized search engine on the other hand searches directly using localized indexing where changes are reflected quickly.

2.4 When more advanced search features are desirable

A large centralized search engine can search through many different types of sources, however, it typically provides a generic full-text search of the documents indexed without offering advanced search features relevant to the individual source being searched. For example, the user of a web search engine can do a search which retrieves results from a university library, a government agency, and a local tire store. However, all the data will be searched in the same generic way. A more effective search is likely possible if one goes directly to the source, in this case a website. For example, to do a more focused search of the government material the user could go directly to the agency's website, where they might find a more advanced search engine which allows searching of the agency's collection of documents by title, author, year or document number.

A decentralized search engine such as GRACE can provide either the advanced search functionality of the local source or the large-scale free-text searching offered by a web search engine, depending on what is required. If the user is just interested in one category of data source such as libraries, then they can search that source using the appropriate advanced search features. If, on the other hand, they want to search many different types of sources with incompatible metadata or no metadata at all, then GRACE will allow them to do a full-text search which would be closer to that provided by the centralized search engine. By using the metadata and search capability inherent in the individual sources, GRACE will always provide the user with the most advanced search possibilities possible given the sources they choose to search.

One more exciting (from a librarian's perspective) possibility that GRACE will provide will be searching with thesauri or classification schemes that change depending on the sources selected. Once again, because the search engine will be distributed, it can provide the same functionality that you would expect when just searching the local data source including providing the users with a thesauri or classification scheme that is compatible with the data source. So if you were to search a database of papers from a source that focuses on astronomy for example, you would be able to start your search by selecting a term from a relevant tool such as the Physics and Astronomy Classification Scheme.

3 Automatic Categorization

The GRACE toolkit will also provide a categorization engine which will dynamically integrate and categorize results from the various data sources. The categorization engine will be based on Automatic Idiomatic Representation (AIR) technology which is designed by Virtual Self, one of the GRACE partners.

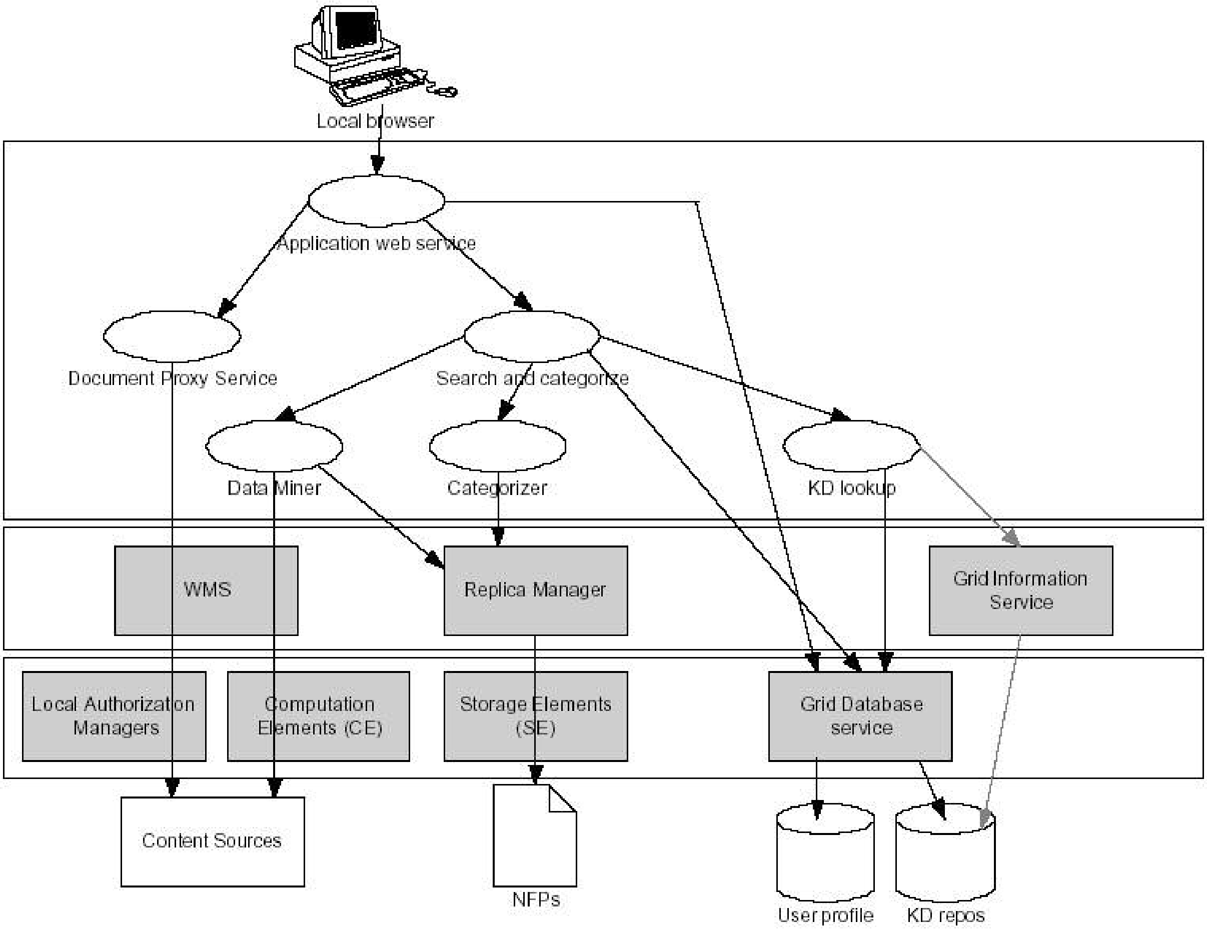
Results can be automatically categorized regardless of how they are formatted or whether they contain metadata. Moreover, the AIR technology is language independent, so with the aid of language lexicons it will work on sources in various languages. To begin with, GRACE will be capable of automatically identifying and then categorizing results in the following languages: English, German, Swedish and Italian. Additional languages may be added at a later stage.

The dynamics of how the categorization engine processes and automatically categorizes results is beyond the scope of this paper. Very briefly stated, it will use a combination of algorithms including a word-stemming algorithm enhanced by the use of language lexicons to automatically identify the language as well as identify stop words and language-specific exceptions that need to be taken into account in the algorithms.

4 GRACE Architecture

GRACE will be based on the European Data Grid (EDG) [6] which is in turn based on the Globus Toolkit (GT) [7]. Finally it aims for an Open Grid Services Architecture (OGSA) [8] compliant architecture. However, the implementation of the first version toolkit will not be able to use the full OGSA model, as there is not a good base for it either in the Globus toolkit (GT3 will be fully OGSA compliant, but is still unstable at the moment), or, of course, in EDG (based on GT2). GRACE will use existing grid services whenever possible (e.g. for configuration, security, managing storage) and will use an additional layer of services which will be built on top of these existing grid services.

The figure below is an overview of the services, both those inherent in the EDG and the ones built on top of it [9]. The two lower levels (grey boxes) correspond to underlying grid services and collective services, respectively (this displays a partial view for simplicity). Below are the 'fabric' elements of GRACE, which are the content sources, databases, and stored NFPs (which are actually data, not services). The top-most layer describes the GRACE proprietary services up to the application web service.



Developing a Grid-Based Search and Categorization Tool", High Energy Physics Libraries Webzine, issue 8, October 2003. URL: http://library.cern.ch/HEPLW/8/papers/1/