CHAPTER 3

SEMANTIC WEB MINING

3.1 Introduction

The World Wide Web drastically changed the data representation and its accessing. The huge amount of data electronically accessed by the machines within no time. The computers were developed for numerical calculations but now a day they have been totally used for information processing, text processing, using data bases and playing games. There is a difference between the ways the information presented to user than actually it is present in the repositories.

The Internet contains more than 10 billion static pages of information to be used by more than 1000 million users spread over the world. It is difficult to access & maintain this enormous amount of data using natural languages. It is rather difficult to bridge the gap between the available information and the techniques used for accessing it.

The search engines used today are keyword based and have their own limitations for searching from the huge repositories. The problems faced by user are

- Along with required pages the irrelevant or merely relevant pages were retrieved in thousands,
- The search engines are keyword sensitive, if string of keywords is not found then it separates the words and presents the irrelevant matter,
- It is required to collect the information from several web pages if its spread over various documents.
- The output results for query are relevant pages or meaningful pages on the internet, and
The results web search is not accessible by other supporting software tools.

The web content is increasing at very faster rate and difficult for search engines to cope up with it despite new techniques of searching. The Semantic Web’s establishes machine understandable Web resources. Researchers in this area plan to accomplish this by creating ontology and logic mechanisms and replacing HTML with markup languages such as XML, RDF, OIL, and DAM [1].

The information retrieved term used for search engine is not suitable. It is required to interpret the queries or sentences and then search the required information from the web pool. The semantic web is an alternate approach to search web content which should be more readable and processable using intelligent techniques. The semantic web is the outcome of the existing web.

The semantic web is propagated by the World Wide Web consortium (w3c), an international standardization body for the web. The driving force of the semantic web initiative is Tim Berners-Lee, the very person who invented the WWW in the late 1980s. He expects from this initiative the realization of his original vision of the web, a vision of the web, a vision where the meaning of information played a far more important role than it does in today’s web [2].

The three broad categories of Web mining are:

**Web content mining** is the application of data mining techniques to content published on the Internet, usually as HTML (semi-structured), plaintext (unstructured), or XML (structured) documents.

**Web structure mining** operates on the Web’s hyperlink structure. This graph structure can provide information about a page’s ranking or authoritativeness and enhance search results through filtering.
Web usage mining analyzes results of user interactions with a Web server, including Web logs, click streams, and database transactions at a Web site or a group of related sites. Web usage mining introduces privacy concerns and is currently the topic of extensive debate [3].

3.2 The Semantic Web

“The first step is putting data on the web in a form that machines can’t naturally understand, or converting it to that form. This creates what is called semantic web- a web of data that can be processed directly or indirectly by machines”[4], Tim Berners- Lee, The inventor of the web

He visualizes the future of the web in two parts:

- Web a more collaborative and
- The web should be understandable, processable and easily accessible.

In the Semantic Web, adding semantics to a Web resource is accomplished through explicit annotation (based on an ontology). Humans cannot be expected to annotate Web resources; it is simply not scalable. Hence, we need to automate the annotation process through ontology learning, mapping, merging, and instance learning. Web content-mining techniques can accomplish [5].

With the propagation of internet, email, instant messaging, audio-video conferencing the explosion of information was out of control of searching technologies.

The pillars of semantic web are:

1. Knowledge Management
2. Business-to-customer e-commerce
3. Business-to-Business e-commerce
4. Semantic web agents
The semantic web aims to provide advanced knowledge management systems, which should provide knowledge acquiring, accessing and maintaining in the repositories [6].

3.2.1 Knowledge Management

Knowledge is the basis of realizing intelligent services. As Vannevar Bush pointed out, sharing and inheriting knowledge can be challenging. The Knowledge Grid, according to Fran Berman, will be a mechanism that can synthesize knowledge from data through mining and reference methods and enable search engines to make references, answer questions, and draw conclusions from masses of data [7]. The knowledge management is important for international organizations with geographically dispersed departments. The current knowledge management techniques lack in so many ways, they are:

i. Searching required information because of keyword base search engines,

ii. The current web agents are not that intelligent for extracting information,

iii. The web contents should be updated regularly, and

iv. The distributed databases lack in proper structuring of documents.

The advanced knowledge management systems are expected to provide much more than the current available techniques. The semantic web as a solution to above problems should provide:

i. Meaningful information by organizing knowledge using conceptual spaces according to its meaning,

ii. Intelligent tools for extracting information,

iii. Query based search instead of keyword based search,

iv. The user, the exact information required by answering the queries, and
v. Interoperability between the documents.

3.2.2 Business – to-Customer e-commerce

Business-to-customer (B2C) includes the online shopping, ordering products, e-billing and delivery status. The user can have several options of selection. On the basis of quality, price, products. But to visit all online shops and then taking a decision is time consuming for the user.

So, intelligent agents should visit the shops and should provide the range of products along with its types and prices for proper selection. The semantic web allows developing such intelligent agents that search the required product with all required parameters. The intelligent agent conducts the negotiations on behalf of customer.

3.2.3 Business –to-Business e-commerce

The business–to-business (B2B) is growing at very fast rate. The most of the online business uses the latest technologies. In B2B e-commerce internet plays the major role of communication. The new B2B portals charged the business world to great extent. The HTML is found be insufficient to elaborate the business needs, the XML comes with new standards to cope up with B2B e-commerce using vocabulary with its meaning. The intelligent agents carry out the business functions automatically in B2B e-commerce.

3.2.4 Semantic-Web-Agents

The semantic web agents designed for particular task performs the task as required by the user. Providing exact information from the pool and should be easily processable involving comparison, composition and mixing of web services. It is required to automate the semantic web technologies for such interoperabilities.
The web services and semantic services in the internet environment forces to form semantic web services. The web services fits into semantic web by using XML. The web service problems can be solved by using semantic web technologies.

### 3.3 Web Services

The XML platform is used to complete the web services which makes the data and processing more smarter than the usual one [8]. Annotation element a special ELEMENT of the XML SCHEMA language that is intended for commenting XML schemas and also for adding information for applications similar to processing instructions [9].

Web service (WS) technologies are gaining importance in the implementation of distributed systems, especially grids. One such example
is the Web Service Resource Framework (WSRF), which extends the current WS technologies by modeling the stateful services. Design and development of the service oriented distributed system is quite common and there are several emerging WS initiatives, which tries to automate the process of discovery, composition and invocation of services. The semantic web services are a typical example, showing the potential of how ontological modeling can improve the shortcomings of service oriented computing [10].

The web services using XML platform grow along with

i. Logical Assertions
ii. Classification
iii. Formal Class Models
iv. Rules
v. Trust

3.3.1 Logical Assertions

It is the smallest expression of useful information. It models the key parts of a sentence by connecting a subject to an object with a verb. The Resource description framework (RDF) captures these associations between subjects and objects. As Tim Berners –Lee states “It is the connections of letters to words, words to phrases and phrases to documents”. Thus the assertions are not free from commentary but instead adds logical statements to a resource or about a resource [11].

3.3.2 Classification

The things can be classified to establish groupings to make generalizations. The classification used to classify the resources on internet. It uses the taxonomy concepts and models for classification. The taxonomy is the classification of information entities in the form of a hierarchy, according to presumed relationship of the real-world entities that they
3.3.3 Formal Class Models

A formal representation of classes and relationship between classes to enable inference which is beyond the conventions used in OOP languages. The ontologies are used to represent such formal class hierarchies and relations between the classes. The ontology captures logical information in a manner that can allow inference.

E.g. if Raj is identified as leader, then we can infer that Raj is a person and he may lead an organization. To each of these basic concept the formalisms are added using properties and rules.

The Web Ontology Language (OWL) being developed by W3C will have UML presentation. The UML tools have wide availability of commercial and open source will support in creation of ontologies and significantly expand the number of potential ontologies [13].

3.3.4 Rules

The web can be transferred from a collection of documents into a knowledge base using XML, RDF and inference rules. An inference rule allows user to derive conclusions from a set of premises. A well known logic rule called “modus ponens” states that

If X is TRUE, then Y is TRUE.

X is TRUE,

Therefore, Y is TRUE.

The semantic web can uses information in ontology with logic rules to infer new information [14].

3.3.5 Trust

The trust determination can be made better by adding semantics. Instead of making logical statements about resources, smart applications
will make inferences on statements that they can trust. Thus, verifying the source of statements in a key part of the semantic web.

The above discussed five things will move corporate intranets and the web into a semantically rich knowledge base where smart software agents and web services can process information and achieve complex tasks[15].

3.4 Semantic Web Technologies

The small domains and prototypes uses the semantic web in the internet. The organizations like W3C, academic institutions. Business oriented etc. have a wide support for semantic web and interoperability.

The support of XML-based technologies such as SOAP-based web services provides interoperability in the Internet. RDF provides a way to associate information and also it is the foundation of other ontology based languages of the semantic web.

The XML Topic Maps (XTM) provides mechanism to develop taxonomies of information to classify data. Web services provides mechanism for communication. Currently the semantic web technologies are growing at faster rate. The main focus of semantic web services is to provide interoperability.

The next big trend in web services will be semantic-enabled web services where we can use information from web services from different organizations to perform correlation, aggregation and orchestration. The companies are heavily investing in semantic web technologies. The documents have been managed using ontology-level power. e.g. The PDF files can be handled by other software’s even if the software doesn’t know what PDF document is and How to display it.

The Gartner’s report mentioned that “The semantic web ontologies will play a key role in 75 percent of application integration by 2005” [16].
The semantic web technologies can help in decision support, business development information sharing and automated administration. But the technological progress will lead to move advanced semantic web than today.

Some semantic web technologies that are necessary for achieving the functionalities are:

3.4.1 Explicit Metadata

Most of the web content is useful for user rather than programs. The HTML was used for writing web pages, but for user there is no problem in understanding the information but the machines may came across. The keyword-based searches will indentify the words by the intelligent agent might found some problems in distinguishing the data.

The semantic web approach for solving these problems by proposing to attack the problem from the web page side. If HTML is replaced by more sophisticated language then the web pages could carry their contents on the top and also produces the document for human readers[17]. Following example explains the metadata.

```xml
<University>
  <Nature of Work> Receiving Degree </Nature of Work>
  <Name of Section> Degree Section </Name of Section>
  <Faculty> Science </Faculty>
  <Clerk> Vivek </Clerk>
</University>
```

This information is easily processable by machines.

The metadata refers to such types of information like data about data. Metadata capture part of the meaning of data resulting in semantic in semantic web. The large organizations are interested in knowledge management and B2B e-commerce; will adopt XML and RDF along with semantic web W3C standards.
3.4.2 Web Ontologies

The ontology is a group of specific objects that can be grouped into abstract classes based on shared properties. It is an explicit and formal specification of a conceptual classes of a finite list of terms and relationships between the concepts of domain.

In a college the faculties, students, courses, other staff members, rules etc. are same important concepts. The relationships typically includes the hierarchies of classes. The members of subclasses are members of a class. Figure 3.2 is a hierarchy definition and they are defined as follows

![Hierarchical relationships](source)

**Figure 3.2: Hierarchical relationships**

(Source: compiled by researcher)

Web ontologies provides a shared understanding of a domain to overcome differences in terminology. Two applications may use the same term with different meaning. e.g:- In university, faculty refers to faculty of Science, Arts or Commerce where as in colleges or institutions it means teaching staff. Such problems can be solved by mapping particular
terminology to a shared ontology. And support semantic interoperability. The ontologies are useful in navigation of web sites and improving the accuracy of web searches. The search engines look for pages that refer to a particular concept in ontology instead of collecting all pages which contains keywords.

In Artificial Intelligence (AI) ontology languages have been used for developing and use. The most important ontology languages for the semantic web [18] are:

- XML provides a platform for structured documents, but imposes no semantic constraints on the meaning of documents,
- XML Schema is a language for restricting the structure of XML documents,
- RDF is a data model for objects (“resources”) and relations between them, with simple semantic for this data model in XML syntax,
- RDF Schema is a vocabulary description language for describing properties and classes of RDF resources, and
- OWL is a richer vocabulary description language for describing properties and classes, such as relations between classes, equality, properties, characteristics of properties and enumerated classes.

3.4.3 Logic

Logic simply means reasoning by which the principles of reasoning can learn. Using logic the knowledge can be expressed using formal languages. It provides formal semantics for well understanding, and automated reasoners can infer conclusions from the given knowledge. The implicit knowledge can be changed to explicit.
The following example shows the inference; suppose all professors are faculty members, all faculty members are staff members, and Vijay is a professor. In predicate logic the information is expressed as follows [19].

\[ \text{Prof}(X) \rightarrow \text{faculty}(Y) \]
\[ \text{Faculty}(X) \rightarrow \text{staff}(X) \]
\[ \text{Prof}(Vijay) \]

Then we can deduce the following

\[ \text{Faculty}(Vijay) \]
\[ \text{Staff}(Vijay) \]
\[ \text{Prof}(X) \rightarrow \text{staff}(X) \]

This example involves knowledge typically found in ontologies. Thus logic can be used to uncover the ontological knowledge that is implicitly given. It also helps to uncover unexpected relationships and inconsistencies.

The intelligent agents use logic for making decisions and taking actions. For example: A shopkeeper may decide to offer discount to his regular customers, then it can be fit into rule as

\[ \text{Regular customer}(Z) \rightarrow \text{discount (10\%)} \]

Where a shopkeeper have a database to determine a regular customer.

The advantage of using logic is that it can provide explanations for conclusions. The explanation are rather more important for the semantic web to increase the infer power of semantic web agents. The facts traces to web addresses, trusts should be verified (by agents), and rules are part of shared ontology.

Example:

\[ \text{purchase}(X, \text{Item}) \rightarrow \text{price (Item, price)} \land \text{delivered (Item,X)} \]
\[ \rightarrow \text{owes (X, price)} \]

The logic in connection with other data will be found more effective in semantic web. The XML is used for primary logic and rules but for more
sophistication RDF and ontological languages like OWL, DAML should be used [20].

3.4.4 Agents

The agents are software modules that work indigenously and self protecting. The agents were based on object oriented programming and component based software programming. A smart agent in the semantic web receives some tasks and preferences from the person, search information from web resources, communicate with other agents, compare information and produces results for the user.

Grid is a new paradigm of Internet computing to share distributed resources and collaborates among them. Several grid systems have been proposed in the last few years. The concepts of agent are coming from artificial intelligence, and the benefits have arisen as a result of the Web are grist to the mill of agent research and development. The distribution of information and associated technologies lend themselves almost ideally to use by, in and for grid agent systems [21].

![Diagram of Today's System](a)

![Diagram of Future System](b)

**Figure 3.3: Intelligent personal agent,**
*(source: borrowed from book [2])*
The above figure depicts the nature of working of today’s internet and the nature in near future using intelligent personal agents in the figure [22]. Semantic web agents will make use of all the technologies defined

- Metadata is used to identify and extract information from the web.
- Ontologies is used to assist in web searches, to interpret retrieved information and to communicate with other agents.
- Logic is used for processing retrieved information and for drawing conclusions.
- Trusts for determination of inferences on statements for trust and verifying the outputted data.

3.5 Semantic Web: Layered Approach

The semantic web develops in steps called layers one above the other. There by simplifying the way of development of the semantic web. The nature of semantic web should be such that the single user or concerning firm will add their contents, they can build their required tools. Automatically the semantic web goes on developing but it requires a long period to materialize the semantic web to its full extents.

A layered semantic web approach should be based on two principles [23].

**The downward compatibility**: Agents fully aware of a layer should be able to interpret and use information written at lower levels. For example: Agents aware of semantics of OWL can take full advantage of information written in RDF & RDF schema.

**Upward partial understanding**: Agents fully aware of a layer should take at least partial advantage of information at higher levels. For example: An agents aware of the RDF & RDF Schema semantics can interpret
knowledge written in OWL partly, by disregarding these elements that go beyond RDF & RDF Schema.

The following stack diagram shows the layered approach for the semantic web[24].

<table>
<thead>
<tr>
<th></th>
<th>Security + Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusts</td>
<td></td>
</tr>
<tr>
<td>Reasoning/ Proof</td>
<td>Inference Engine</td>
</tr>
<tr>
<td>Higher Semantics or Ontologies</td>
<td>OWL, DAML+OIL</td>
</tr>
<tr>
<td>Semantics</td>
<td>RDF/RDF Schema</td>
</tr>
<tr>
<td>Structure</td>
<td>XML Schema</td>
</tr>
<tr>
<td>Syntax : Data</td>
<td>XML</td>
</tr>
</tbody>
</table>

**Figure 3.4: Layered approach for the Semantic web**

(Source: Borrowed from book[4])

At the bottom XML language is used for writing structured web documents using user-defined vocabulary and for writing documents on web. The next higher layer uses the XML Schema for encoding data elements and description using types, elements, contents models, structures etc.

RDF is a basic data model used for writing simple statements of web resources. RDF does not depend on XML but it uses the syntax of XML language. The RDF Schema primarily used for modeling and organizing web objects into hierarchies. It uses classes, subclasses, properties, relationships, restrictions of domain etc. It is totally based on RDF & used for writing Ontologies. So as to represent the more complex relationships between the web objects more powerful and advanced ontology languages are required to expand RDF Schema.
The OWL builds some concepts and design of DAML+OIL. The OWL possesses classes, subclasses, properties, sub properties, restrictions etc. OWL uses the XML Schema class and data-type information.

The proof layer represents the proof in web languages and validation of data. The ontology languages capture logical information and allow inference.

The topmost is the trust layer also called “Web of Trusts” which automated proof and also providing security and identity features. It uses digital signatures and other kind of knowledge for security and identity purposes. The trust is a top-level and highly conceptual to achieve full potential in security operations and in the quality of Information supplied by web.

3.6 GRID COMPUTING

Grid computing uses geographically distributed computers collectively for high performance computing and resource sharing. The computers can be distributed throughout the globe irrespective of location. It also involves computer from various organizations and outside the organizations boundry [25]. The Internet is the backbone of networking which is much faster to communicate [26].

Data grids are grid-based infrastructures for working with data. A grid is a distributed collection of computing resources that appears as a single virtual computing infrastructure by sharing security services, including single log-on, an LDAP-based information infrastructure, and resource management services. A data grid adds data-specific services. These include grid FTP for moving data over a grid and data replication services, so that data might be efficiently cached over a grid [27].

The shared resources can also include the high performance and expensive laboratory equipments which are located at the place but shared
by the connecting computers in the network. Grid architecture can be visualized as a layered architecture. The topmost layer consists of the grid applications and the APIs from a user’s perspective. The middleware, includes the software and packages and used for grid implementation, for example Globus Toolkit, gLite. The third layer covers the resources available to the grid such as storage, processing capabilities and other application-specific hardware. Finally the fourth layer is the network, layer which deals with the network components like routers, switches, and the protocols used for communication between any two systems in the grid[28].

The Grid infrastructure supports the sharing and coordinated use for resources in dynamic global heterogeneous distributed environments. It involved resources that can manage computers, data, telecommunication, network facilities and software applications provided by different organizations. Namely, a Grid is a collection of distributed computing resources available over a local or wide area network that appears to an end user or application as one large virtual computing system [29].

The term virtual organization is used to describe the distributed team and the shared resources that are provided. The grid computing came into existence in mid 1990’s with high-speed networks and Internet which tends to generate numerous Grid computing projects across the world. The Grid computing uses the standard internet protocols and techniques. The Grids are set up locally, nationally and internationally.

### 3.6.1. Concept of Grid computing

Grid computing uses distributed interconnected computers and resources collectively to achieve higher performance computing and resource sharing. The high speed networks and Internet resource sharing. The high speed networks and internet allowed distributed computer systems to be readily interconnected. It in the most important technique in high
performance computing by providing the resource sharing among various disciplines.

The geographically distributed computers can be used collectively for problem solving due to high-speed Internets. The different organization can share the resources for problem solving and the Grid infrastructure can cross the boundaries of the organizations. These concepts of resource sharing have many benefits like:

- Due to limitations of computing resources some problems could not be solved, these problem can be solved to some extent by grid computing. For ex.- Finding solution for unknown diseases, searching solution for new drugs etc.
- People working in different areas of research now can work interdisciplinary to solve the problems by having expert of multiple disciplines.
- The laboratory equipment which is rarely available can be shared by the distributed computer in Grid Infrastructure.
- Collective data bases can be created to hold the huge amount of data.
- Computers can be more efficiently used in the Grid environment, and
- Re-implementation of business processes was found to be cost saver in Grid Technology.

3.6.2 Collaboration

The Grid computing provides the ability to conduct collaborative computing. Along with collaborative and resource sharing it also provides high performance computing. Collaboration is a core activity at the heart of large scale cooperative systems. However, with the advent of large scale and distributed computational resources, specialised collaborative support is
needed. The Open Collaborative Grid Service Architecture has raised as a framework built on top of the Open Grid Service Architecture (OGSA) providing a set of services that can be used by any collaborative application. The OCGSA framework extends the notion of Grid services into the collaborative domain, introducing the collaborative Grid Service concept[30].

*e-Science and hence the Grid is the infrastructure that enables collaborative science.* The Grid can provide the basic building blocks to support real-time distance interaction, which has been exploited in distance education. Particularly important is the infrastructure to support shared resources this includes many key services including security, scheduling and management, registration and search services and the message-based interfaces of Web services to allow powerful collaboration (sharing) mechanisms. All of the basic Grid services and infrastructure provide a critical venue for collaboration and will be highly important to the community [31]. Developing teams of geographically distributed people forms the Grid and concept of it became grown up in the high-speed Internet.

The vision of Grid computing makes high-performance computing as easy to access. The IBM Company took this vision in the early 2000’s by offering (Grid) on-demand computing by providing access to Grid computing resources whenever requested by the clients and paid for by the clients [32].

The utility computing term can also be used for using Grid resources in a similar way as utilities such as electrical gas and water are metered. The computing on demand led to “Cloud computing” in which companies open a “cloud” of services as a business model, a paid service. Cloud computing is a different concept than Grid computing.
The IT industry finds itself at yet another evolutionary cycle, with new technologies emerging with features and functions that were impossible only five years before. Such trends web services, open source, blade technology, commodity search, distributed computing, and the Grid offer substantial benefits. However, it is difficult to describe those benefits to our executives, who are inclined to say, “Just make it happen,” without an appreciation of cost, complication, or risk [33].

A well known example of Grid computing project that involved experimental equipment and geographically distributed Grid is at CERN (European Centre for Nuclear Research) Which straggles the Swiss/French border near Geneva Switzerland. The work there has centered around the search for new fundamental particles and understanding the universe for many years using state-of-the-art particle colliders. The data from their experimental large Hadron collider facility is sent to researchers around the world for analysis and collective research using Grid technology. This project involves the massive amount of data created at ultra high speed. (PBytes/sec) [34].

The important aspect of Grid computing is that it uses the same protocols & standards of internet, so that the software developed by various organizations can interoperate to develop and encourage the user community. The term e-science was coined by John Taylor, The director General of the United Kingdom’s office of science and technology, in 1999 to describe conducting scientific research using distributed networks and resources of a Grid computing infrastructure. Another more recent European term is e-Infrastructure, which refers to creating a Grid-like research infrastructure [35].

With the development of Grid computing tools such as Globus and UNICORE (UNiform Interface to COmputing RESources) a growing number of Grid projects began to develop computational applications.
These computational applications been categorized as:

**Computationally intensive**

It is the high performance computing category and for solving large problems traditionally. Some times a problem is required to be solved repeatedly using different parameters for obtaining solution.

**Data intensive** : The data intensive category includes the computational problems but more importance given to large amount of data to store and process. For example: The project of earth system Grid (ECG) for climate modeling and research. This provides the climate data among climate research community is an example of data intensive Grid application.

**Experimental collaborative projects** : In experimental collaborative projects the data is collected from experimental equipment. In this the computer shared by the virtual organizations formed for the project and the experimental data should be used for computations.

The collaborative Grid computing can also be classified into

- **Enterprise Grids** – Grids form within on organization form collaboration.
- **Partner Grid** – Grids form between collaborative organizations or institutions.

The enterprise Grids might cross the administrative domains of the departments and required to share their resources.

The key features of Grid computing are

- Shared multi-owner computing resources.
- Used Grid computing software Globus with security and cross- management mechanism in place.

The Grid computing software provides the tools for individuals and teams to use geographically distributed computers owned by other collectively.
3.6.3 Computational Grid applications

In the last decade there were large numbers of computational Grid projects working in science and engineering disciplines. These grid projects are in areas such as [36].

- Biomedical research,
- Industrial research,
- Engineering research (All branches),
- High energy physics,
- Chemistry,
- Medical, and
- Bioinformatics.

Large scale Grid computing project involving experimental equipment set up is the large Hadron collider experimental facility constructed at CERN for research into practical physics.

Another large scale Grid computing project involving experimental equipment is the NSF Network for Earth quake Engineering Simulation (NEES) for research into reducing the effects of earth-quakes. The Grid computing project focusing upon earth sciences in the Earth System Grid (ESG) for climate modeling and research.

A Grid computing project in the medical area is the UK e DiaMoND Grid computing project. The objective of this project was to build a national data base of mammographic images to aid Screening and diagnosis of breast cancer. It is categorized in both i.e. data intensive having data bases of images stored and transferred and experimental collaborative including experimental equipment for obtaining data and manipulating it.

The data being shared with other medical Grid computing projects involving human data, are legal and ethical considerations of maintaining confidentiality of personal records.
3.7 Data mining in Grid

Data mining relates to creation of formulations, analysis of data, implementation in the areas for processing specific data and the general patterns of data which provides the extraction of implicit and useful information from the repositories. The data mining uses the mathematical models, statistical methods, machine learning and representation for knowledge discovery.

The Data Mining Grid (DMG) is emerging area. DMG is the combination of the data mining and the grid computing, which can satisfy the needs of data mining and knowledge discovery for business in a grid computing environment. The Data Mining Grid is a system, which can solve the problem of business intelligence with scalable and reliable ways by sharing the solutions, algorithms, computing, data, and storage [37].

It is used for information retrieved and handling Grid systems. The data mining techniques are highly implied in the scientific research in various disciplines such as Biological, Mathematical, Biomedical, Chemical, physical, Social sciences, Governments, High-tech engineering and Process automation. Data mining also known as knowledge discovery in data bases.

A data mining Grid seeks a trade-off between data centralization and distributed processing of data so as to maximize effectiveness and efficiency of the entire process. A data mining Grid should provide a means to exploit available hardware resources in order to handle the data volumes and processing requirements of modern data mining applications. Further it should support data placement, scheduling end resource management [38].

In Grid data mining tasks and algorithms can be applied on distributed data. It should allow:

i. The distribution of compute-intensive data analysis among a large number of geographically scattered resources.
ii. The development of algorithm and new techniques such that the data would be processed where they are stored, thus avoiding transmission and data security issues.

iii. The investigation and potential solution of data mining problems beyond the scope of current techniques.

3.7.1 Challenges in Grid Computing

Grid computing allows the creation of comprehensive computing environments meeting the requirements of ground challenge applications. Sometimes this may be very complex. The complexity arises due to limitations of software and hardware resources, decentralized control, security and privacy mechanisms, local policies and usage pattern of the resources, etc. These complexities need to be addressed in order to fully exploit Grid’s features for large-scale data mining applications.

The monitoring and analyzing the information in a grid is a way of supporting the management. This also involves the system performance and operations such as throughput, network bandwidth and response time. The data mining technology is the solution to check this problem of complexity.

3.7.2 Data mining Grid and mining Grid data

The data mining and Grid technology when works hand in hand develops the concepts of data mining Grid and mining Grid data. Both the terms are different. The data mining Grid should be a grid that is specifically derived to provide the facility of data mining using data mining applications. Where as the mining Grid data provides the techniques to mined the data from the Grid resources for analysis and processing.

3.7.3 Data mining Grid

It is a grid which provides large scale data mining. It is the technology to perform data analysis within a particular domain. It is based on data to be mined, the algorithms for data mining, its methods and controls required for data mining. The data mining algorithm developed for
a particular data mining tasks are separate and have different software implementations. A data mining Grid is a system whose functions is to provide the sharing and use of data in order to improve existing data mining applications. Which facilitate the improved effectiveness and efficient solutions to data mining?

- A data mining Grid provides novel data mining applications than conventional one.
- It should facilitate the scalability by adding grid resources to accommodate user and applications without loss of performance.
- It should provide the execution of all kinds of data mining tasks there by increasing its scope of working.
- It should provide tools to define complex data mining processes.
- It should provide tools to monitor and steer data mining applications running on Grid.
- It should provide the extensibility maintenance and integration in the due course of time as per need.

The data mining Grid system should also required basic things like resource sharing and interoperation, its applications and resource management to match available Grid resources.

### 3.7.4 Mining Grid data

Grid provides the techniques to mined the data from the Grid resources and is a heterogeneous and decentralized environment and it deals with more complex problems. It is required to pool, analyze and interpret all the relevant information from the Grid. The data mining is found to be powerful tool for providing analysis and interpretation of large amount of complex data.
Mining Grid data is found to be a solution for improving the performance, operation and maintenance of Grid computing environment. The data from multiple sources can be gathered and analyzed using data mining techniques to learn new useful information about different Grid features. The mining Grid data is a supposed to be a methodology that could help to address the complex issues involved in running and maintaining large Grid computing environments.

Both these techniques work hand-in-hand and should benefited from each other. A data mining Grid can efficiently deploy large-scale data mining applications and mining Grid data implies techniques used to understand and solve the complexity of Grid computing.

### 3.8 Data mining techniques in grid

Grid computing basically designed to solved problems involving compute intensive applications. The ability of Grid to handle large amount of data which is in distributed manner as resources enlarged the horizon of the Grids. Being service-oriented, the data mining tasks and knowledge discovery processes provides services in Grid-based environment. The key technology for knowledge intensive and computing on demand perspective vision are SOA and web services, semantic web and ontologies, pervasive computing, p2p systems, grid computing data mining and knowledge discovery tools etc. So as to support the data-intensive applications the Grid environment must support smarter data management and data analysis tools and techniques using the resources, services and decentralized data access mechanisms.

The data mining techniques used for mining data in Grid along with data mining Grid and also for how the knowledge grid framework has been developed as a collection of grid services and how it can be used to develop distributed data analysis tools and knowledge discovery processes
exploiting the service oriented architecture model. The data mining techniques are discussed in terms of services for data mining, and data analysis in the knowledge grid are as follows [39].

**The data mining services are:**

i. Grid minor  
ii. ADaM services  
iii. GMS  
iv. FAEHIM  
v. Anteater service  
vi. DMGA  
vii. ESSE

### 3.8.1 Grid Minor

It is a framework developed to deal with all tasks of the knowledge discovery process on the Grid and integrate them to service-oriented Grid applications. It is based on cross industry standard process for data mining. It consists of two main components: technologies and tools. The grid minor provides a robust and reliable high performance data mining and OLAP environment.

The different services like data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation co-operate interactively in the grid minor architecture is the key productive e-science analytics.

The e-science analytics support to access very large data collections and very large scale computing resources. It includes scientific method of data pre-processing, integration, analysis, data mining and visualization associated with information extraction and knowledge discovery from scientific data sets.
Technologies and tools:

Grid minor includes services for sequential, parallel and distributed data, text mining, online analytical processing, data integrative, data quality monitoring and improvement based on data statistics and visualization of results. All these services supported by Grid minor effectively assists application developers to develop Grid enabled high performance analytic application. It also provides tools called work flow composition assistant based on semantic web technology to increase the productivity of analytic tasks.

3.8.2 ADaM services

The Algorithm Development and Mining system (ADaM) was developed in early 1990’s with the goal of mining large scientific data sets for geophysical phenomena detection and feature extraction. It provides image processing and pattern recognition algorithms for science data sets. The ADaM system was developed at the university of Alabama in Huntsville to investigate new methods of processing large volumes of Earth observing system remote sensing data sets as a project for NASA. It was designed as a client/server architecture to support remote client applications communicating with the data mining server. The requirement for complete end-to-end mining capability and a specific mining algorithm working within the data analysis workflow, the ADaM system was developed as a toolkit to provide customized solutions and general applications and provide mining services.

Two data model were developed one for images and another for pattern data. The ADaM toolkit also includes utilities to translate back and forth between the two data models. The standard used and protocols referred maximizes the interoperability of ADaM services with other applications and tools. The scientists can mix and match various ADaM
services with other available services to create complex mining and analysis workflows to solve specific problems.

3.8.3 GMS-System

It is Grid Monitoring System (GMS). Which adopts a distributed data mining approach for the detection of misconfigured grid machines. The GMS non-intrusively collects data from sources available throughout the grid system. It converts raw data to semantically meaningful data and stores it on the machines from which it was obtained. A distributed outliers algorithm is employed to identify misconfigured machines. The algorithm itself is implemented as a recursive workflow of grid jobs and is especially suited to grid system in which the machines might be unavailable most of the time or often fail altogether.

Grid systems are difficult to manage because of there heterogeneous hardware and as they pool data from several administrative domains there is no single authority able to enforce maintenance standards, if required to update the software. There are two general approaches to automated system management they are white-box and black-box. The white box approach relies on knowledge of the system and its behavior. It interprets system events according to a set of rules. These rules specify exceptional behavior pattern and appropriate responses to them. It depends on domain experts for the definition of pattern is its drawback. So it is difficult to maintain the system when it develops and updates.

In black-box approach problems are detected and diagnosed with limited, if any, knowledge about the system are simple and economical to apply and maintain and the results are comparable to those of expert systems. The GMS is fully distributed. It processes and stores the data where they were created and the overhead it normally creates is negligible. The distributed algorithm only collects a portion of the data. Where as the
GMS does not process raw data but rather translates it first into semantically meaningful terms using an ontology.

3.8.4 FAEHIM Process

It is Federate Analysis Environment for Heterogeneous Intelligent Mining (FAEHIM) process. At the core of the process is the application of specific data mining methods for pattern discovery and extraction. This process is structured from interactive and iterative stages within a discovery pipeline and workflow.

The FAEHIM is a data mining tool kit that makes the use of web service composition with the widely deployed Triana workflow environment. The explosive growth in data and databases has generated an urgent need for new techniques and tools that can automatically transform the processed data into useful information and knowledge.

The FAEHIM enables composition of web services from a predefined tool box. Web services have been developed from the Java-based templates provided by the Weka data mining library of algorithms. Using Triana workflow system, data analysis can be performed on both local and remote data sets. A variety of additional services to facilitate the entire data mining process are supported for data translation, visualization and session management.

3.8.5 Anteater techniques

Data mining aims at extracting useful information from large volumes of data. Building scalable, extensible and easy-to-use data mining systems has proved to be a hard tasks.

The Anteater is a service oriented architecture for data mining, which relies on web services to achieve extensibility, simple abstractions for users and support computationally intensive processing on large amounts of data. Anteater relies on Anthill, a run-time system for irregular, data intensive, iterative distributed applications to achieve high performance. The
combination of a web services architecture and a parallel programming environment provides a rich environment for exploring different levels of distributed processing with good scalability. Anteater is operational and being used by the Brazilian government to analyse government expenditure, public health and public safety policies. The Anteater is very positive regarding the system usability, flexibility and impacts in view of users.

The Anteater architecture is based on the standard life cycle of knowledge discovery in database (KDD). It is designed as a set of distributed components that offer their services through well defined interfaces. The user interface is implemented over a web page and processed within the server. The Anteater maintains the information about the output format for the several algorithms and the input format for the several visualizations available in Anteater.

Anthill is the runtime support for system, a framework that was developed to support the efficient implementation of a significant class of applications on heterogeneous distributed environment. The anthill programming model used to implement the data mining algorithm in the context of Anteater. The Anteater is the free software available through the web site. The Anteater data mining service architecture which is implemented as a services-oriented architecture is easily extensible using web services as new resources may be added to it easily.  

3.8.6 DMGA architecture  

It is a generic brokering-based data mining Grid architecture for deploying data mining services in a Grid. The DMGA approach presents two different composition models.  

i. Horizontal composition, and  

ii. Vertical composition  

The Horizontal composition offers workflow capabilities and vertical composition for increasing the performance of inherently parallel data
mining services. The DGMA scheme is specially significant to those services accessing a large volume of data which can be distributed through diverse locations of Grid.

A Grid can be a suitable computing infrastructure in which data mining applications are applied and executed. So as to make the data mining successful the following things must be taken into account.

i. Some of data mining applications require access to large volumes of data,

ii. Data may reside in geographically distributed locations,

iii. Data may belong to several organizations, and

iv. Every organization may apply its security policies.

DMGA tries to address these issues by means of the definition of both a set of Grid services for data mining and usage patterns for their composition. The current version is based on Open Grid Services Architecture (OSGA). DMGA is a generic architecture, which defines services reflecting the main stages in a data mining process are pre-processing, data mining and post-processing. DGMA extends OSGA by the definition of new data mining services. These services use basic grid services of which data-related services are the most relevant for DGMA. The three stages which are designed to support data mining applications are the central element of the DGMA architecture and are linked to data mining techniques and algorithms.

The DGMA not only provides the new specialized data mining services, but also offers a framework in which composition of several Grid services helps to deploy an enhanced solution to a data mining problem there by creating workflows, which allows several services to be scheduled in a flexible and efficient manner. It is performed in two ways.

**Horizontal composition** It offers workflow capabilities by combining different functional services. This composition depends on the
decomposition of the overall data mining process into stages and is characterized by the DGMA.

**Vertical composition** Several outputs of the same replicated services are combined, such services have the same functionality but access a different data position. It is used for increasing the performance of inherently parallel data mining services. To maintain the global state of the resources needed by a grid service a centralized element called a broker is needed. The broker is vital for any grid infrastructure because its operation and performance determine to what extent the user requirements are met and how efficiently the underlying resources are utilized. A broker consists of three basic components.

a. **A resource discovery system**
   This system collects information about the Grid nodes. It must be aware of available services, such as software library or a statistical computation grid service, hardware details, CPU architecture, amount of memory and the presence of a specific device. Also current stage of nodes, disk space, CPU utilization, external bandwidth on free memory.

b. **A match making system**
   Based on the information provided by the resource discovery system, the match making system controls the selection of candidate nodes, matching all the clients' requirements.

c. **A decision making system**
   Following a pre-established logic or algorithm the system chooses the best nodes from the previously selected candidates.

**3.8.7 ESSE Search Engine**

It is Environmental Scenario Search Engine (ESSE) developed for Grid-based data mining. The increasing data volume in the grid needs more powerful data mining. The scientist needs an integrated and authoritative
representation of the natural environment in their analysis requires new approach to data mining, management and access.

The natural environment includes elements from multiple domains such as space, terrestrial weather, oceans and terrain. Systems such as Global Change Master Directory (GCMD) at NASA, Master Environmental Library (MEL) etc. provides ability to search metadata by keywords, the environmental data sets are distributed across the network but they are unable to search for specific pattern within the data themselves. The ESSE for data grids provides uniform access to heterogeneous distributed data sets for querying it. The ESSE data processor requires one or several time series as input and a set of instructions or program to process the data. The set of instruction can include arithmetic expressions, elementary functions, moving average, seasonal variations and time shifts. The result of ESSE is another time series that may contain fewer points than the original data. The ESSE applications from the environmental data sets are like

- Warning of the climate system is from observations of increase in global average air and ocean temperature, melting of snow and ice and rising global mean sea level.
- At continental, regional and ocean basin scales, numerous long-term changes in climate have been observed. This includes ocean salinity wind patterns, droughts, heat waves and the intensity of tropical cyclones like hurricanes and Typhoons.

The combination of the ESSE weather reanalysis data sources with the ESSE fuzzy search engine and data processor became a powerful tool for quantitative formalization and rapid validation of such fuzzy statements.
3.9 Semantic Web Mining in Grid

Semantic Web was originally proposed by Tim Berners-Lee, James Hendler and Ora Lassila in 2001, with the objective of extend to the Web some of the intelligent behavior of humans [40]. Semantic Web is defined as an extension of Web technology in which the information has its own meaning. It makes possible for the web to understand and satisfy the requests of people and machines using web content. Besides, Semantic Web allows obtaining effective discovery, integration, automation and reuse of applications [41].

Semantic Web technologies are based on open standards. If a computer understands the semantics of a document, it does not only interpret the series of characters that make up that document but also understands the meaning of the information in the document. It is also defined as the extension of the today’s web where the information is given the better meaning and allowing users to work in hand-in-hand to use the resources of web. Also allowing the software agents and programs to interoperate for extracting the useful knowledge and solving the complex problems in the internet.

The need of metadata (which annotates and describes Web content) to allow machine-to-machine operation, it is required to have automated processing to give semantic meaning to each Web resource.

The aim of the Semantic Web is to provide machine processable metadata that describes the semantics of resources to facilitate the search, filter, condense, or negotiate knowledge for their human users. A core technology for making the Semantic Web happen, but also to leverage application areas like Knowledge Management and Ecommerce, is the field of Semantic annotation, which turns human-understandable content into a machine understandable form[42].
A layered model of the Semantic Web comprises of following

- A set of Web resources, with a unique, global identity, described by metadata in a common and shared formalism, and with rules for inferring new metadata and knowledge through ontologies.

- A set of basic services, such as reasoning and querying over metadata, and ontologies and semantic search engines. These services represent a great improvement over current Internet services, such as the Domain Name System (DNS) and keybased search.

- A set of high-level applications developed by using basic services. At this stage, major efforts are addressing the development of languages and technologies for the standard modeling and implementation of metadata and ontologies such as XML Schema and RDF Schema, DAML+OIL, and OWL.

- Ontology-building tools that let users define and build ontologies—for example, DUET (DAML UML enhanced tool), OilEd, Protégé and OntoEdit.

- Ontology-based annotation tools, for annotating Web resources according to an ontology—for example, UBOT (UML-based ontology toolset) DAML.

- Ontology-learning tools, for learning ontologies from natural language documents—for example, Corporum and Text-To-Onto.

- Ontology manipulation tools—for example, DAML and Jena APIs—that let users navigate and manipulate ontologies [43].

- Semantic Web technologies that integrate the Semantic Web are:
Uniform Resource Identifier (URI)

A Uniform Resource Identifier (URI) is a compact sequence of characters that identifies an abstract or physical resource. It is a global naming scheme. It is the way to identify any of those points of content, whether it is a page of text, a video or sound clip, a still or animated image, or a program. The most common form of URI is the Web page address, which is a particular form or subset of URI called a Uniform Resource Locator. It describes the mechanism used to access the resource.

Resource Description Framework (RDF)

It is a standard syntax for describing data. RDF is a specification that defines a model for representing the world, and syntax for serializing and exchanging that model [44].

RDF Schema

It is a data typing model for RDF which allows creating properties and classes. It could also be defined as a vocabulary for describing properties and classes of resources based on RDF and aggregating semantics descriptions [45].

Web Ontology Language (OWL)

It is designed to be used by applications that need to process the content of information rather than just presenting information to humans. It adds more vocabulary to describe properties and classes than RDF or RDF Schema [44,45].

OWL-based Web Service (OWL-S)

It supplies Web service providers with a set of markup language constructed to describe the properties and capabilities of their WS in a computer-interpretable form [46]. Semantic technologies represent meaning through the use of ontologies [47]. Ontologies, in terms of Semantic Web, are defined as meta-data schemas that provide a controlled vocabulary of
concepts, each of them with a complete and structured semantic [48,49]. They are descriptions of concepts for the purpose of allowing knowledge sharing and provide machine readable and understandable data [50]. OWL Web Ontology Language works with ontologies defining and instantiating them. OWL is also used to represent the meaning of terms in vocabularies and the relationship existing among these terms [51]. The use of Semantic Web in our applications has its benefits. The Semantic Web provides the ability to tag all content on the Web, describes each piece of information, and gives semantic meaning to each resource. Thus, search engines become more effective, and users can find the precise information they are looking for. Organizations providing services can tag services with meaning; using software agents based on the Web, these services can be found and used dynamically [52].

3.9.1 The Grid Services

Grid Services (GS) are software components which provide access to a set of Grid resources such as data sources, high performance equipment and computational resources [53]. GS are the base of the Open Grid Services Architecture (OGSA) [54], which aims to standardize all services on a Grid application. OGSA defines the GS concept based on technologies of Grid computing and WS. It also defines standard mechanisms for naming, creating and discovering GS [55,56].

Open Grid Services Infrastructure (OGSI) defines a set of conventions and specifications for the use of WSDL and XML Schema to enable GS [57]. GS solve the problems WS have for their use within Grid Computing (stateless and persistence) using a factory/instance approach to WS. It has a central service factory which maintains a set of service instances instead of having one big stateless service shared by all users. When a client needs to invoke an operation of the service, it would talk to the instance, not to the factory. In the same way, if the client wants to create
or destroy an instance, it will talk to the factory. Each client could have
access to more than one instance of the service, and one instance could be
shared by many clients. These instances are transients (the opposed to
persistent) because they have a limited lifetime which is not bound to the
lifetime of the GS container. It means that one client can create and destroy
instances of a GS instead of having one persistent service permanently
available. It avoids the risk that another client might destroy the operations
performed over the GS by the first client. [58].

**The main improvements of GS introduce are**

- **GS support portType extension**: It means we can define as a
  portType as an extension of a previously existing portType [59].
- **GS are stateful services**: It means they retain the state after each
  invocation and subsequent execution [60].
- **GS are transients**: It means they are not linked to the lifetime of the
  server, but they can be created and destroyed as required [61].
- **GS have Service Data**: Service Data allows to include a set of
  structured data to any service, which can be accessed through its
  interface. Service Data is an extension that allows including not
  only operations in the WSDL, but also attributes and any type of
  data (classes, arrays, etc.) [62].

### 3.9.2 The Semantic grid Services

As mentioned above, through WSRF, GS can be seen like WS. Moreover, we
have to remember that WS are described using WSDL definitions and advertised in
UDDI registries. The capability of a WS is defined through several aspects, such as
their automated discovery. The current discovery mechanism supported by
UDDI is not adequate for automated discovery. Its main limitations are the lack of
semantics in the
discovery process and that UDDI does not use information in the service descriptions during discovery; this reduces the effectiveness of UDDI [63].

Because GS are treated as WS through the WSRF standard, these limitations also affect them. The semantic discovery of Web/Grid services is achieved by adding semantics to WSDL and UDDI, and then we can find the required services using semantic matching algorithms.

In order to add semantics to WSDL of a service and find relevant operations, these operations must be mapped to concepts in appropriate OWL ontologies. The use of ontologies allow to represent service descriptions (WSDL) in a machine-interpretable form like OWL-S [64,65].

Additionally, WSDL contains message parts defined as input and output operation parameters, thus, can add ontologies and RDF concepts to annotate message parts in WSDL, and later can be perform more effective searches for Grid/Web Services and obtain better results. It can improve semantic service discovery and storing semantics in the existing structures of UDDI. UDDI contains the technical information that is necessary to invoke a service, information about companies that own the service, and business entities [66]. If added semantics to UDDI, it could also obtain good results in our search. The Semantic Web approach to GS, not only facilitates services discovery, but could also improves the selection, invocation, execution, composition and interoperation of Grid/Web Services as it does for services discovery[67] through the application of Semantic Web technologies.

Consequently, it is observe that the use of Semantic Web and ontologies provide an appropriate mechanism to represent GS and to enhance their discovery, composition, selection, invocation and interoperation. The Semantic Grid seeks to incorporate the Semantic Web approach into the ongoing Grid: “As the Semantic Web is to the Web, so is the Semantic Grid to the Grid. The emerging Semantic Web Infrastructure
is an infrastructure for grid computing applications.” The Global Grid Forum supports the Semantic Grid through its Semantic Grid Research Group. The Semantic Grid’s research issues align with many aspects of the next-generation grid

- Full support of a grid’s three recognized layers: computation and data, information (where data produces information), and knowledge (where knowledge can be used to make decisions)
- Provision of seamless, pervasive, and secure resource use [68].

The Grid-Enabled Information Repositories generally refer to the storage of all information relevant to enterprise businesses and daily operations, including data and metadata. Moreover, the information produced by SGII (including management information, operational information, and metadata for data sources and services, etc.) is also stored into grid-enabled information repositories [69].

Support Service Space constructs common and basic services for SGII by providing Access Control Services and Semantic Enrichment Services. The Access Control Services are responsible for the guarantee of flexible information access control with the support of OGSA-DAI and Applications-DAI [70]. Based on OGSA, the Data Access and Integration (OGSA-DAI) infrastructure can model heterogeneous data sources as grid services and integrate them into SGII architecture. Analogy with OGSA-DAI, Applications-DAI (Data Access and Integration of application data) is responsible for wrapping data sources that stream data from applications as OGSA-compatible grid services and integrate them into SGII architecture. Furthermore, OGSA-DAI and Applications-DAI provide basic data access capability to SGII. Semantic Enrichment Services deal with relative topics of semantic enriching, which include Semantic Annotation, Ontology Creation, and Inter-ontology Mapping services. Semantic Annotation service is used to enrich data and service content based on pre-defined
ontologies by using automatic or semi-automatic semantic annotation tools. Ontology Creation service wraps a semi-automatic approach as OGSA-compatible grid service to create ontologies on the basis of a process of semantic discovering, explanation and user modeling. These ontologies constitute the conceptual backbone for information-level and semantic-level interoperability, which is the basis of semantic enrichment. To achieve pervasive interoperability among (extra-) enterprise information sources, SGII requires the support of multi-ontology. Accordingly, Inter-ontology Mapping service makes this to be reality by creating interconnection semantics based on semantic mapping between ontologies [71].

3.9.3 The Knowledge Grid Services

Next-generation grids must be able to produce, use, and deploy knowledge as a basic element of advanced applications. In this scenario, the KNOWLEDGE GRID(KG) system as a joint research project of ICAR-CNR (Istituto di Calcolo e Reti ad Alte Prestazioni-Consiglio Nazionale delle Ricerche) and the Universities of Calabria and Catanzaro, aiming at the development of an environment for geographically distributed highperformance knowledge discovery applications [72].

- The KNOWLEDGE GRID is a high-level system for providing grid-based knowledge discovery services.
- These services let professionals and scientists create and manage complex knowledge discovery applications composed as workflows that integrate data sets, mining tools, and computing and storage resources provided as distributed services on a grid. KNOWLEDGE GRID facilities let users compose, store, share, and execute these knowledge discovery workflows and publish them as new components and services on the Grid. One can use the KNOWLEDGE GRID to mine very large data sets available over grids, make scientific
discoveries, improve industrial processes and organization models, and uncover valuable business information [73].

The knowledge-building process in a distributed setting involves data and information collection, generation, and distribution followed by the collective interpretation of processed information into “knowledge.” Knowledge building depends not only on data analysis and information processing but also on interpretation of the produced models and knowledge filtering. Knowledge discovery includes mechanisms for evaluating the correctness, accuracy, and usefulness of processed data sets, developing a shared understanding of the information, and filtering knowledge to be kept in an accessible organizational memory. The KNOWLEDGE GRID provides a higher level of abstraction and a set of services based on the use of grid resources to support all phases of the knowledge discovery process. So, it lets end users concentrate on developing the knowledge discovery process without worrying about grid infrastructure details.

The KNOWLEDGE GRID architecture is composed of a set of services divided into two layers:

- The Core K-Grid layer that interfaces basic and generic grid middleware and services
- The High-level K-Grid layer that interfaces the user by offering a set of services for the design and execution of knowledge discovery applications.

The KNOWLEDGE GRID environment represents discovery processes as workflows that a user can compose using both concrete and abstract grid resources. Users define knowledge discovery workflows through a visual interface that shows resources (data, tools, and hosts) and offers mechanisms for integrating them in a workflow. The environment stores single resources and workflows using an XML-based notation that represents a workflow as a data flow graph of nodes, with each node
representing either a data mining task or data transfer service. The XML representation allows workflows for discovery processes to be easily validated, shared, and translated into executable scripts and stored for future executions. Implemented VEGA (visual environment for grid applications), a software prototype that implements the main components of the KNOWLEDGE GRID environment, comprising services and functionalities ranging from information and discovery services to visual design and execution facilities. We can simplify the design and execution of complex applications by exploiting the advantages that a grid environment offers in the development of distributed knowledge discovery applications. The application design facility lets users build typical knowledge-based grid applications in an easy, guided, and controlled way [74].

In developing the knowledge layer, the main issues are

- Synthesizing useful and usable knowledge from data
- Leveraging the Grid infrastructure to perform sophisticated data-intensive large scale computation.

To integrate knowledge discovery techniques in grid environments, we need an unambiguous representation of the knowledge base (through metadata and ontologies) that can translate moderately abstract domain-specific queries into computations and data analysis operations that can answer such queries by operating on the underlying systems [75].

Peer-to-peer model

Peer to peer is at once a set of protocols, a computing model, and a design philosophy for distributed, decentralized, and self-organizing systems. P2P is a set of methodologies and technologies that lets a group of computers collaborate in a network of equals (peers), without central coordination [76]. In spite of current practices and thoughts, the Grid and P2P models share several features and have more in common. Integrating the two computing models could bring many benefits in designing future
scalable grids. Grids used for complex applications include a large number of nodes, and decentralize their functionalities to avoid bottlenecks. P2P could help ensure grid scalability, designers could use P2P to implement nonhierarchical decentralized grid systems.

The models and challenges P2P systems face are not new: peers stay at the edge of a network in which everyone creates as well as consumes effectively, the Internet's original formulation [77]. P2P basic elements include.

- Action at the edge of the network (such as computing, resource sharing, and communication)
- Shared resources between peers (such as CPU idle cycles, disk space, computing power, network bandwidth, and content)
- Direct communication between peers, which takes place without great assumptions about the underlying network and protocols (such as DNS)

P2P’s main potential is its ability to exploit idle computing resources, facilitating information exchange (information discovery and content distribution). From the Grid point of view, P2P’s main interesting aspects are scalability, self-configuration, autonomic management, dynamic resource discovery, and fault tolerance. On the other hand, current P2P systems often lack the ability to deploy production-quality services, such as QoS negotiation, persistent and multipurpose service infrastructure, complex services (beyond simple file-sharing), robustness, performance, and security.

**Pervasive and ubiquitous computing**

Ubiquitous computing describes distributed computing services, such as personal devices, wearable computers, and sensors in the environment, and the software and hardware infrastructures needed to support applications on these computing devices. The terms “pervasive”
and “ubiquitous” are used interchangeably. As Mark Weiser described in his well-known article [78], ubiquitous computing means interconnected hardware and software so pervasive and so integrated in the environment that no one notices their presence. The characterization consider ubiquitous computing’s main dimensions [79] are:

- Mobility of users, devices (PDA, phones), and software (mobile agents)
- Degree to which devices are embedded into the environment

Mobile computing is about the ability to

**Knowledge discovery services**

These services extract knowledge from the data stored inside the grid knowledge base. It is able to use these services to build high-level knowledge discovery applications, such as the KNOWLEDGE GRID, and to enhance existing basic grid services. Two possible applications that require distributed data mining functionalities and accessing distributed partitions of a knowledge base are grid-based document management (that is, document classification and retrieval over the grid) and an enhanced version of the GridFTP protocol using data mining techniques to predict optimal transfer parameters.
Figure 3.5: Key technologies for grids
(Source borrowed from [43])

Figure 3.6: The next-generation grid’s main layers
(Source borrowed from [43])
3.9.4 Ontology-based Semantic Search

For this type of search, a user might query very detailed information about data mining resources annotated in DAML+OIL. The result set is very accurate because concepts from the underlying ontology clearly indicate the searched terms’ semantic content. Our ontology-based search engine supports several kinds of simple inference that can broaden queries, including equivalence, inversion, generalization, and specialization. For example if the query result set is empty, the user can at least find objects that partially satisfy the query: some classes can be replaced by their superclasses or subclasses. Moreover, in addition to finding out where and how to access the available data mining software, DAMON helps a user by searching all the available software that satisfies some user requirements, such as performing a given task (such as classification), implementing a given algorithm (such as CHAID, chi-squared automatic interaction detector), using a specific methodology (such as decision trees), and requiring a specific input format [80].

DAMON (data mining ontology) is an ontology that explicitely anages knowledge about data mining and related software tools. It offers users a reference for the different classes of data mining tasks, methodologies, and software components available to solve a given proble[81]. The choice of how to structure an ontology determines what a system can know and reason about. The ontology classifies data mining software that lets a user select the most appropriate software to solve a KDD (knowledge discovery in databases) problem. The ontology represents the data mining software’s features by classifying their main components and showing their relationships and constraints. The data mining knowledge base we use to support knowledge discovery programming has two conceptual layers:
at the top layer, DAMON gives general information about the data mining domain; specific information about installed software components, and

data sources are maintained where the resources reside.

From an architectural point of view, the ontology is a central resource, whereas specific metadata are distributed resources.

3.10 Interoperability: Weapon for Grid mining

In recent years, substantial effort has been made in researching and developing interoperable data management systems. As more and more special-purpose, non-interconnected software-intensive systems for application areas varying from embedded system software to logistic management systems are being used, the issue of interoperability is becoming increasingly important as a weapon for mining the Grid and also in software engineering research. Advances in computer communications technology, the recognition of common areas of functionality in related systems, and an increased awareness of enhanced information access can lead to improved capability and organizing data. The interoperable data management systems have major focus on data model mapping, communication protocols, and database integration, including schema integration, query processing, consistency management and security management [82].

The interoperability must also deal with real-time issues and justify the need for extending interoperability research into real-time. It is required to identify a number of research issues in integrating real-time data management systems. These issues do not exist in either traditional distributed database systems or multi databases; they must be addressed in the context of real-time system integration.
For interoperability between applications, the information processing must be supported by an architecture providing distribution technology and transparent access to information through heterogeneous networks and environments. The distribution of information has been done with a strong connection with current concepts and advances in the field of object oriented technology.

The advances in the Internet have emphasized the need for interoperability among different and heterogeneous systems, i.e. their ability to exchange data and services in a transparent way. The interoperability in a GIs context where systems are geographically distributed which needs to exchange spatial data. It is a unification of spatial data models and services in order to achieve interoperability among GISs. The interoperability at different levels of abstraction applied. At the highest level (the spatial data on the Internet) a set of common interfaces are adopted to represent both data and services and at the bottom level (database level), the concept of a specialized sewer (called OGIS/SQL server) is introduced, which is the entity that tries to obtain data integration among the data sets [83].

Grid interoperability involves not only the ability of systems to exchange information but also includes the capability for interaction and joint execution of tasks. Therefore the objective is to create such a system which provides system interconnectivity along with interoperability among systems. A prime difficulty in achieving interoperability among heterogeneous components of a system is that the component systems were developed independently, without any requirement for interoperation. Thus systems have different architectures, different hardware platforms, different operating systems, different host languages and different data models. Short of redeveloping a new system using the consolidated requirements from the various component systems and a common architecture, hardware platform,
operating system, host language, etc. (a cost prohibitive approach), a means must be devised to achieve the goal of component interoperability in the face of expected limited acquisition budgets. In order to overcome the limitations as above it is required to have new technologies and methods for achieving system interoperation and interoperability [84].

**Interoperable data exchange**

The great success of HTML language has led to development of so-called extensible Markup Language (XML). XML is a very flexible markup language for description of the content. The XML is so flexible, that it can suit applications of ubiquitous computing ranging from ecommerce, scientific applications, exchange of process data and visualization. The XML schemas in conjunction with SOAP (Simple Object Access Protocol) are superb tools for interoperable data exchange. The XML schemas can be used to create schema for description of exchanged data, while the SOAP technology enables to perform Remote Procedure Calls (RPC) using the HTTP protocol. And every field bus, in fact, is just a tool for exchange of data using explicit or implicit Remote Procedure Calls. However XML itself does not guarantee interoperability.

Without a common template it is possible to describe a single entity in many different and incompatible ways. For this reason there exist XML templates (XSL) that define method for encoding the data into a XML document. However there is missing a widely accepted universal template for description of process data. There have been developed many different templates designed to fit some special needs, however there is no process automation template that would cover all physical quantities, objects and data structures needed in process automation. For conversion among data represented using different XSL templates, there had been developed XSLT (extensible Stylesheet Language Transformation) technology. Using XSLT it is possible to convert data among competing XSL standards.
Similarly, networks will utilize network data systems, high-level programming languages or graphics support systems, and database technologies to create, provision, and administer new and enhanced services. The data systems will be distributed among work groups and network elements through techniques of interoperability and Distributed Relational Databases (DRDBs). Such capabilities are important to telephone companies worldwide that want to remain competitive. Responding quickly to changing market demands with fast service enabling and dynamic provisioning will help to ensure their survivability. Interoperability helps to ensure the survivability [85].

The interoperability among databases is of increasing interest IABGA88, Bati86, BSTSi, BrSi88, BST89, CRE87, Demi89, ElHe88, Hei189, Lars89, LiAbSG, LiAb87, RuE189, Rusi88, Shet89, Kim89, SiMa89, Wang90, WeKe89, Wied87, YuSc89. Understanding the real semantics of data is essential in schema integration in fully integrated heterogeneous database systems as well as in the interoperability of database systems. It has been pointed out that success in schema integration depends on understanding the semantics of the schema components and the ability to capture and reason with the semantics [86].

References


[21] Lu Quan, Chen Jing, “A Grid Agent Model for Information Retrieval based on Topic Map and Knowledge Elements Mining”


[26] Ian Foster Argonne, Carl Kesselman, “The Grid: Blueprint for a New Computing Infrastructure”.

[27] ROBERT GROSSMAN AND MARCO MAZZUCO, “DATASPACE: A DATAWEB FOR THE EXPLORATORY ANALYSIS AND MINING OF DATA” 1521-9615/02 © 2002 IEEE.


[29] Shu-Tzu Tsai Chao-Tung Yang, “Decision Tree Construction for Data Mining on Grid Computing”.


[31] FRAN BERMAN, GEOFFREY FOX, TONY HEY, “Grid Computing Making the Global Infrastructure a Reality”.


79

[34] Barry Wilkinson,“Grid Computing, Techniques and Applications”.


[39] Werner Dubitzky University of Ulster, UK ,"Data Mining Techniques in Grid Computing Environments”.


[83] Aphrodite Tsalgatidou, Eleni Koutrouli “E-SERVICES INTEROPERABILITY ANALYSIS AND ROADMAP ACTIONS” Department of Informatics & Telecommunications, National & Kapodistrian University of Athens, Greece.
[86] Aphrodite Tsalgatidou, Eleni Koutrouli “E-SERVICES INTEROPERABILITY ANALYSIS AND ROADMAP ACTIONS” Department of Informatics & Telecommunications, National & Kapodistrian University of Athens, Greece.