Information Retrieval: Today and Tomorrow

Varsha Kumari
Dept of CSE,
NIT Delhi, India

Sandhya Pundhir
Dept of CSE,
GBPCE, Delhi, India

Supreet Kaur
Dept of CSE,
NSIT, Dwarka, India

ABSTRACT

World wide web (WWW), also known as web has become one of the most valuable resources of Information Retrieval (IR) and Knowledge Discovery. Due to the increasing amount of data available on the Internet, the effective information retrieval requires an efficient framework that can provide a communication between clients as well as for servers. Information Retrieval involves the searching and retrieval of knowledge from large database collections. In this paper, the analysis and comparison of the IR techniques is done. The analysis is based on the Information Retrieval in distributed environments, grid and the cloud environments. Different crawling techniques in these environments are also discussed in this paper.

Keywords

Information Retrieval, Distributed Grid, Cloud, Multi-Agents, Crawling

1. INTRODUCTION

Today wide adoption of Internet has become an integral part of human life in terms of communication, gathering information, conducting business etc. Also, the web has grown exponentially in size and contains a large amount of publicly accessible web document distributed all over the world on thousands of servers. As document collection grows larger, they become more expensive to manage. The different types of data have to be managed and organized properly so that they can be accessed efficiently. However, the retrieval techniques based on the Information Retrieval (IR) research have found their way into major information services and the World Wide Web (www).

The main goal of IR is to “finding relevant information or document that satisfies user’s needs”. An Information Retrieval process begins when a user enters a query into the system. Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user.

Providing timely access to the data collections both locally and globally is the key in making an IR system truly useful. A crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. Different crawling techniques are used for Information Retrieval. The concept of mobile agent in web crawler has increased the crawling speed and minimized the network overhead. Multi-agent based crawling technique is comparatively better than any other traditional crawling techniques [13][14][15].

Earlier the IR based on the distributed environment was mainly used. The application of parallelism and distributed computing can greatly enhance the ability to scale the IR algorithms.

In order to provide efficient and scalable IR services, different architectures for distributed and parallel Information Retrieval is discussed in this paper.

2. DISTRIBUTED COMPUTING & IR

A distributed computing is any computing that involves multiple computers remote from each other that each has a role in a computation problem or information processing. Information Retrieval using distributed computing is also distributed retrieval.

Distribution can be calculated as a special case of Multiple Instruction, Multiple Data (MIMD) parallel computing. Parallel computing is a form of computation in which many calculations are carried out simultaneously, operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently, in parallel.

Parallel IR algorithms can be approached from two different directions. One way is to develop new methods that directly use the parallel implementation. Another way is to adapt the existing well-studied IR algorithms to parallel processing. Parallel IR based on document partitioning fits well into distributed computing.

Distributed parallel technology and combination of IR has become a popular trend these days.

2.1 Architecture

Fig.1 shows the architecture of the IR in distributed environment. The query is entered by user/client through the distributed servers. From the collection of choices produced, user chooses the most appropriate data set for retrieval. The user’s query is then moved forward to the most appropriate server for the concurrent search. The results go through the process of consideration. The final search results are then returned to the user to complete a distributed search process [6].

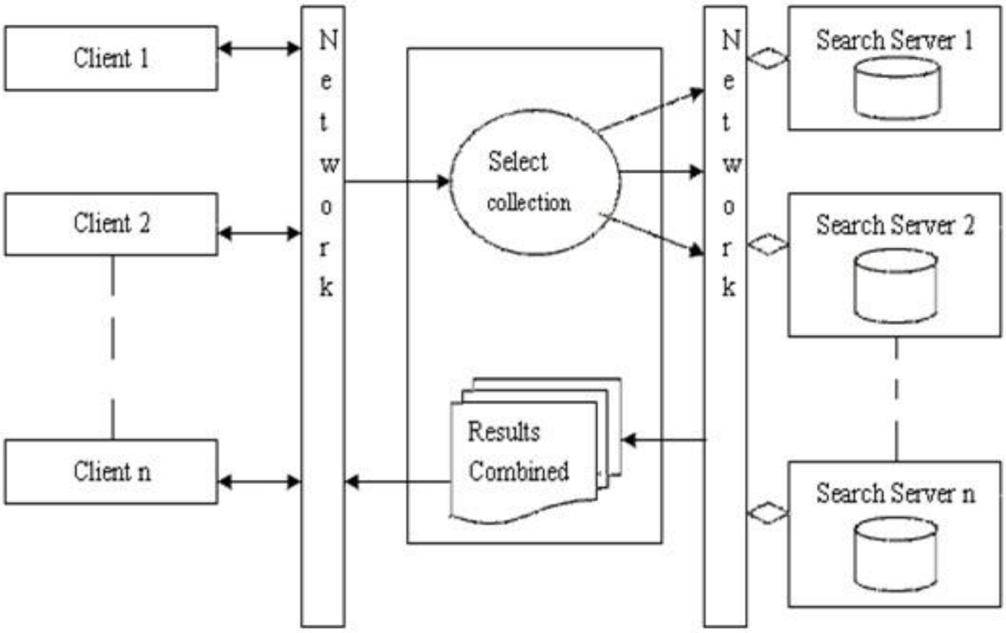

Fig 1: Architecture of Distributed Information Retrieval

2.2 Features

Distributed IR has the following advantages [1][6]:

- Can retrieve large amount of data/
- Improved retrieval speed/
- Machine failure does not affect the service/

Compared with the parallel search, the main features are [6]:

- Due to the heterogeneity of different systems, the realization of communication between processors is done through a shared memory.
- Data retrieval in each node has different capabilities, and some subset of data is selected to implement distributed search.
- Due to the expansion of web information resource and the need for efficient IR, the distributed parallel technology and combination of IR technology has become a future trend.

2.3 Challenges and Issues

Distributed systems are spread out over vast distances. Due to this, there are many issues and challenges surrounding such distributed systems.

- *Heterogeneity of data* - Describes a system consisting of multiple distinct components.
- *Transparency and Openness of system* – Property of each system to be pen for interaction with other systems.
- *Security* – Includes Confidentiality, Integration and Availability.
- *Scalability* – As the system, number of resources, or users increase the performance of the system is not lost and remains effective in accomplishing its goals.
- *Fault handling* – Need a way to detect failures, mask failures, recover from failures and build redundancy.
- *Concurrency* - Arise when several clients attempt to request a shared resource at the same time.
- *Transparency* – Must be able to offer Access transparency, Location transparency and Failure transparency to its users.

3. GRID COMPUTING AND IR

Grid is an infrastructure that enables the integrated collaborative use of high-end computers, networks, databases into a single mega computer that can achieve full connectivity and resource sharing available on the Internet.

Grid IR combines grid technology and IR, that makes full use of powerful computation and resource sharing, and it provides a new way of IR.

3.1 Architecture

Fig.2 depicts the architecture of Grid based IR. The retrieval processor receives the query entered by the user interface and the query is transferred to the search manager through the index manager. On the grid, the search manager extracts the metadata from the Internet resource through data mining techniques. Search manager gathers data and provides data to the index generator. Index generator build an index, that may be thought of as a machine-searchable representation of the data in the collection and then matches the data against a query and provides response set to the retrieval processor. Retrieval processor further presents the response set to the end users [2][4].

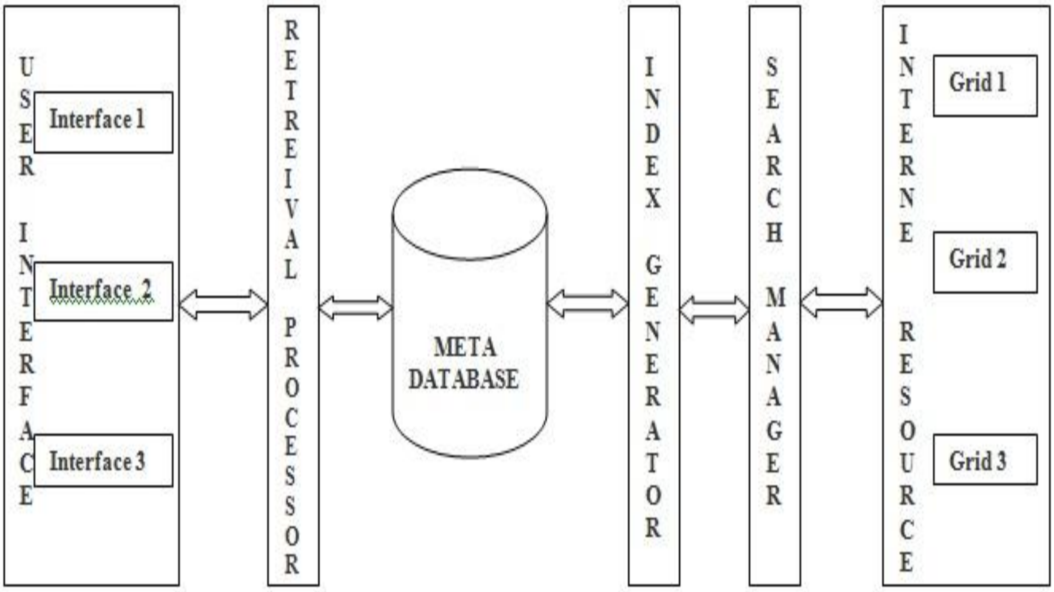

Fig 2: Architecture of Grid based Information Retrieval

3.2 Features

A grid retrieval system is an infrastructure that bonds and unifies globally remote and diverse resources in order to provide dynamic and computing support for a wide range of applications. Grid computing architecture has the following advantages [2][4]:

- Used to run an existing application on different machine.
- Massive parallel CPU capacity increases the computing power.
- Provide an environment for collaboration among wider audience.

The various features offered by the grids are [2][4][16]:

- *Large scale distributed supercomputing* – Able to deal with a number of resources and distributed computers to tackle problems that cannot be solved on a single computer.
- *Resource sharing and coordination* – Resources present on the grid belong to different organizations coordinates and it allows users to access various non-local resource.
- *Heterogeneity* – Grid contains a wide range of both software and hardware resources that can be data, files or any other software components.
- *Transparent Access* – Grid is visualized as a single virtual computer by the user.
- *Multiple administrations* – Grid is implemented with different administrative organization and their access policies.

3.3 Challenges and Issues

The challenges faced by the Grid architecture are [2][4]:

- Required to have fast interconnection between the systems and resources.
- Issue of scaling thousands of resources.
- Portability, interoperability and adaptability of resources on the network.
- Implement security with the environment.

Grid based information retrieval addresses the following issues [2][4]:

- Resource naming and representation by Universal Resource Identifier.
- Resource discovery and retrieval process concerned with very diverse grid are accessed through different service protocols.
- Definition and management of relationships should be proposed to describe grid-monitoring data.
- Grid should provide the users to navigate within the complex space of query result via a simple interface.

4. CLOUD COMPUTING AND IR

Cloud computing has evolved as a novel prototype used for extreme scalable, fault tolerant and compliant computing technique on enormous computers.

Cloud retrieval has become information retrieval based on the cloud computing services. Users having Internet terminal unit, enter the information to be retrieved. Cloud retrieval system then automatically retrieves the clouds layer quickly and finds the relevant information for the user.

Cloud computing architecture refers to the various components and sub components mainly required for cloud computing. The basic components required for cloud retrieval are [5][15]:

- Front end for various client types.
- Consist of Back end platform for storage.
- Homogeneous data.
- Network can be Internet, Intranet or an Inter cloud.
- Institutes or third party to share the services and resources.

4.1 Architecture

A cloud retrieval system consists of Cloud information layer, cloud retrieval cluster system, and User query box. The cloud retrieval cluster system consists of several layers – Cloud collecting layer, Cloud processing layer, Cloud index layer, Cloud query layer, Cloud interface layer, Data storage layer, cloud retrieval layer, cloud retrieval monitoring system, cloud management and Scheduling system. The architecture is shown in fig.3 [5]. All the layers work in collaboration with other layers.

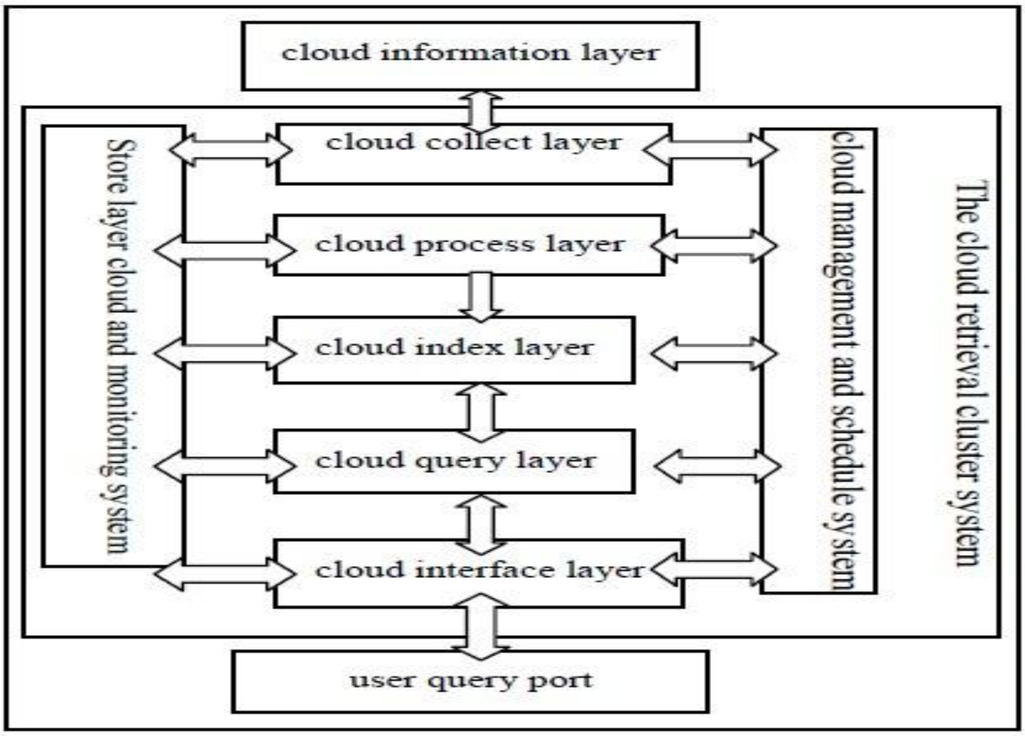

Fig 3: Architecture of Cloud based Information Retrieval

4.2 Features

A cloud computing architecture has various advantages [5][15]:

- Dynamically scalable
- Easier to manage.
- Faster application deployment.
- Has a data owner outsourcing data service where data is stored typically in keyword value form.

The cloud IR has following features:

- *On Demand Self-Service* – Provide needed service automatically without human interaction.
- *Broad Network Access* – Provide use of heterogeneous thin or thick client platforms.
- *Resource Pooling* – There is a sense of independence of provided resource to user.
- *Rapid Elasticity* – Capabilities can be provided and released automatically so appear to be unlimited to every user.
- *Measured Services* – Resource usage can be monitored.

4.3 Challenges and Issues

The challenges faced by the cloud architecture are [5][15]:

- Security and Privacy.
- Interoperability and Portability.
- Performance and Bandwidth Cost.
- Reliable system.

5. COMPARATIVE STUDY ON VARIOUS ENVIRONMENTS

Taking into consideration various performance measures and characteristics on distributed environment, a comparative study is done in Table1. The distributed environments discussed share a similar vision: reduced computing costs, increased flexibility and reliability by using third party operated hardware [1][6][8][9]. However, some features are different between various distributed, cloud and grid environment.

Table 1. Various Environment and their Performance Measures

Performance Measures	**Distributed**	**Grid**	**Cloud**
Scalability	Scalable multiple processing nodes	Multiple working nodes scalability	Multiple working nodes, and hardware scalability.
Capacity	Huge heterogeneous data on the network	Large heterogeneous data	Large homogeneous data managed by data centers
Quality	Better than	Better than	High quality

	traditional IR	distributed IR	web services better than Grid IR

There are many important features that are common in various distributed environments. Various performance measures on different environments – distributed, grid and cloud environments are summarized in Table 2 [1][6][8][12].

The first performance measure *scalability* means the ability to do well with many machines and huge heterogeneous distribution of data. Second performance measure is the *capacity* that deals with the maximum data storage capacity. And the third one is *quality* describes quality of service provided by the particular environment.

Table 2 presents various features of distributed, Grids and Clouds, highlighting the similarities and differences among these paradigms.

The first feature, *Nature of environment* is dynamic in all these environments. The second feature is the *Resource sharing*, appropriate message passing mechanism is chosen for coordination among the nodes in a Distributed environment, Grids appear to be fairly sharing resources environment, whereas Clouds provide the resources and files that the Service Provider requires on demand.

Another feature is *communication technique*. Various distributed environment uses appropriate message passing mechanism. Mostly appropriate approach is used in simple distributed environment and high speed transfer mechanism among heterogeneous environments is used in grid but in cloud environment high speed message passing mechanism is done locally. In addition to these, communication overheads increase with the increase in data transfer among nodes on a huge heterogeneous data. Mutually, distributed and grids add the communication overheads attributable to heterogeneous data environment. But cloud environment contains less overheads because of homogeneous data retrieval processing locally [6][7][9][10][12].

Table 2. Various Environment and their Features

Features	**Distributed**	**Grid**	**Cloud**
Nature of environment	Dynamic	Dynamic	Dynamic
Resource sharing	Coordination and message passing mechanism	Remote data access approach	Coordinator make logical division, Assigned resource not shared
Communication techniques	Appropriate approach and message passing	High speed data transfer and message passing	High speed local area message passing
Communication overheads	Increases with volume of data	Complicated query request increases overheads	Less overhead because of homogeneous data retrieval
Security mechanism	Less secured	Security considered when the model is build.	Less secured because of third party involvement.
Efficiency	Better than parallel IR	Better than distributed system	Best comparatively

As far as the *security* is concerned, it has been a major issue of various distributed environments. Grids have not dealt with end user security but it is taken into account from the beginning when the grid is originally built; Whereas in clouds, each user is provided an access to it's individual local environment hence seems it is more secured, but because of the involvement of the third party cloud environment become less secured.

Cloud is facing a serious problem caused by lack of high level services; this may be a result of the low level of maturity associated to clouds environment. In contrast, Grids have a number of these high level services for instant data transfer and metadata search [8]. In terms of *efficiency*, IR in a distributed environment is better than a parallel IR.

Grid based information retrieval is comparatively better than simple distributed environment. Whereas, cloud based IR perfoms much more efficiently as per the user requirement than the grid environment as it is mainly concerned with the retrieval of data locally from a homogeneous database.

The essential objective for creating the Cloud is to provide a required set of capabilities to the user that is totally a design specific interface. In contrast, Grids are considered as a general purpose environment that provides a complete set of available system capabilities and interface for users.

The main feature of the cloud computing is the *homogeneity* within each data center in the infrastructure compared to grid computing. In case of any conflict among heterogeneous data center and/or different administration domains, it can become a serious issue for cloud interoperability. These environments support both traditional and multi- agent based crawling techniques [13][14][15][16].

Resource selection algorithms help in selecting a small set of database that contains a lot of relevant documents. We have discussed the different distributed models. So to propose an efficient Information retrieval algorithms following things must be considered:-

1. To have less data transfer cost while query executing.

2. Have minimum query execution time.

3. Good query execution algorithm steps are also important.

4. To get the most relevant document.

As these databases on the internet are continuously increasing in real time so resource selection algorithm must be updated dynamically. There are many algorithms available for the dynamic update, but still much improvement is required. Evolutionary algorithms when used for local and then for global resource selection along with proper selection of the distributed environment may give good results.

6. CONCLUSION

Distributed computing architecture is the backbone of the most recent technologies – Grid computing and Cloud computing environment. However, clouds are considered as a new generation and user-friendly version of Grid computing.

The comparison of different environments shows that the cloud environment is the most appealing. It is dynamic, most scalable, provides high quality services and lower cost as compared to other environments.

As compared to other environments, Grid Environment is a general purpose, providing all features better than a distributed environment. Security is taken into account from the beginning when the grid is originally built.

Multi-agent based information retrieval is widely accepted and comparatively better as it overcome number of challenges existing in other crawling techniques as traditional and semantic crawler.

7. REFERENCES

[1] Eric Brown, Modern Information retrieval, chapter 10 Parallel and Distributed IR.

[2] Qing Chen,"Towards Web- based Information Retrieval in Grid Environment", IEEE 2010

[3] http://en.wikipedia.org/wiki/Information retrieval.

[4] Zhang Mei,"Information Retrieval Based on Grid",IEEE 2010.

[5] An Junxiu, " The Demonstration of Cloud Retrieval System Model", Journal of Software, Vol 6 No. 2 , February 2011, pages(249-256).

[6] Linping Shuang, Honjun Zhu,"Analysis of Distributed Information Retrieval", IEEE 2011(ICMT), pages(5297-5300)

[7] Parul Kalra Bhatia, Tanya Mathur, Tanaya Gupta ," Survey Paper on Information Retrieval Algorithms and Personalized Information Retrieval Concept", International Journal of Computer Applications, pages(0975-8887), Vol 66 No.6, March 2013

[8] Hosam AlHakami, Hamza Aldabbas, and Tariq Alwada'n,"Comparison between Cloud and Grid Computing: Review Paper", IJCCSA, August 2012

[9] L.M. Vaquero, L..R. Meino, J. Caceres, M Lindner,"A break in the Clouds: Towards a Cloud Definition", 2009, http://portal.acm.org/citation.cfm?id=1496091.1496100

[10] Members of EGEE-II, "An egee Comparative study: Grids and Clouds - evolution", Technical report, Enabling Grids.

[11] Dr. Sanskruti Patel, Dr. Priti Sajja, "A Grid- based Model for Integration of Geographically Distributed & Heterogeneous Educational Resources for Knowledge Extraction & Deleivery", IJIRSET, ISSN: 2319-8753, Vol No. 2, Issue 9, September 2013

[12] Jamie Callan,"Distributed Information Retrieval", Springer, The Information Retrieval Series Vol 7, pages 127-150, 2000 http://link.springer.com/chapter/10.1007%2F0-306-47019- 5_5#page-1

[13] Varsha Kumari, Preeti Rajput, Sandhya Pundhir, MQ Rafiq,"Web Crawler based on secure Mobile Agent", Research Journal of Computer Systems Engineering, Vol. 03,Issue 03, Page(419-423)

[14] Lou Junwei, Xiao, "Research on Information Retrieval System based on Semantic Web and Multi-agent",ICICCI, IEEE 2010

[15] Yu Mon Zaw, Nay Min Tun, "Web Services Based Infornation Retrieval Agent System for Cloud Computing", International Journal of Computer Applications Technology and Research, Vol 2, Issue- 1, Pages(67-71),201

[16] Miguel L. Bote- Lorenzo, Yannis A. Dimitriadis, and Eduardo Gomez- Sanchez,"Grid Characteristics and Uses: a Grid Definition" published in Springer http://link.springer.com/chapter/10.1007%2F978-3-540-24689-3_36#page-1