Content-based Information Retrieval Techniques Based on Grid Computing: A Review

Mohammed Bakri Bashir[1,2], Muhammad Shafie Abd Latiff[1], Aboamama Atahar Ahmed[1], Adil Yousif[1] and Manhal Elfadil Eltayeeb[1]

[1]Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia UTM, Johor, Malaysia,
[2]Department of Computer Science, Facutly of Science and Technology, Shendi University, Shendi, Sudan

Abstract

Distributed information retrieval methods are growing rapidly because of the rising need to access and search distributed digital documents. However, the content-based information retrieval (CBIR) is concentrated to extract and retrieve the information from massive digital libraries, which require a huge amount of computing and storage resources. The grid computing provides the reliable infrastructure for effective and efficient retrieval on these large collections. In order to build an effective and efficient CBIR technique, varieties of architectures were developed based on grid technologies. The goal of such architecture is to solve interoperability and heterogeneous resource issues, and increase the efficiency and effectiveness of information retrieval (IR) techniques by harnessing the grid computing capabilities. This paper reviews and analyzes latest research carried out in the domain of large-scale dataset IR based on a grid. The evaluation is based on scalability, response time, scope, data type, search technique, middleware, and query type. The contribution is to illustrate the features, capabilities, and shortages of current solutions that can guide the researchers in this evolving area.

Keywords

Content-based, E-learning, Grid computing, Image retrieval, Information retrieval, Medical image, Spatial data Video retrieval.

1. Introduction

Information retrieval (IR) [1] manipulate the description, organization, and storage of information and identify methods to access this information. The IR is considered as the heart of the search engines nowadays and led them to be popular and effective. However, the advent of the Web creates a new big area of applications for IR, which being responsible to utilize and efficiently access the information on the recent search engine systems. Moreover, the rapid growing of distributed information retrieval (DIR) [2] offers to the user an efficient method to search distributed document collections from different and independent sources. Consequently, the tools and methods of IR are improved and adapted to apply in the new distributed environment and create a framework to evaluate and test these tools and methods. Furthermore, the aim of content-based IR is to retrieve the relevant items from different heterogeneous collections. However, the retrieval process from very big collection consumes time and resources [3]. Hence, the best solution is to use parallelization to run an IR process and furthermore improve and speed up the response time. Moreover, different technologies apply to distributed retrieval tasks among several computers such as cluster computing, using Local Resource Management Systems (LRMSes) namely Condor [4]. Nevertheless, grid computing [5] provide a large number of storage computational resources federated among several organization centers. Furthermore, gird technology provides a method to connect a number of clusters regardless of LRMSes used, job execution system, and security mechanism to accomplish the distributed jobs.

In this work, we present a review of grid-based IR that to some extent varies from the traditional IR in both its design philosophy and implementation. Substantially, IR run on the grid infrastructure harnessing the computational resources and high bandwidth potential. Additionally, the IR system use grid-off-line-to extract, analysis, and index contents, while response for the user query performe online. The aim of this paper is to investigate the degree of incorporation of IR in grid computing. Moreover, we focus on numerous complementary aspects like scalability, retrieval technique, scope, response time, query type, data types, and middleware.

Scalability is associated with the capability of a system to execute well even the amount of the resources collaborated is high, or the size of the processing data is huge. Additionally, grid resources change frequently and unpredictable in terms of the availability and the attributes of grid resources [6]. Furthermore, the effectiveness is associated with the freshness of the indexed data that is extremely related to the harvesting efficiency, i.e., if the document is downloaded more regularly, which will increase the probability of the freshness of the document cached copies. In the retrieval process, the effectiveness refers to the precision and recall measures, which precision evaluates the accuracy while recall evaluates coverage of the results [7]. Moreover, the scope of the searching system affects the performance of the search result. There are two types of scope [8]: General-scope and special-scope. General-scope aims to provide the capability to search all documents on the Web. Google, Bing, and Ask are a few of the well-known of this category. Special-scope focuses on documents in specific domains; for example, documents in an organization or in a specific subject area. Furthermore, aspects related to response time for propagating query over the network, and the accumulated query processing times on each node are reviewed in this paper. Moreover, the query type that supported by searching techniques provide more flexibility for the user. Among the different types of the query are the keyword-based, Boolean and proximity, term-based, and image-based. Additionally, the data type of dataset affects the performance of the system. The text retrieval time is less than the time consumed to retrieve the image dataset. Finally, the grid community provides different middleware and tools to implement grid infrastructure. The suitable middleware provides several services and supports the content-based information retrieval (CBIR).

The remainder of the paper is organized as follows. First, Section 2 introduces some concepts related to IR used. The section talks about DIR, grid computing, and IR on the grid. This is followed by Section 3, a detailed classification discussion of the various approaches for CBIR based on grid computing. Section 4 analyzes and compares the different approaches as discussed in Section 3. Finally, conclusions and future work are drawn in Section 5.

2. Basic Concepts of Information Retrieval

2.1 Distributed Information Retrieval

The purpose of IR [1] is to return appropriate information related to the user query. The IR process starts when the user submits a query to the retrieval system. Then, the retrieval system searches the document index that contains part or all of the query text, calculates a score for all documents list before using their scores to rank the documents list. On the other hand, DIR is a system that retrieves information from different and distributed collections [9]. The goal of DIR is benefiting from the distributed locations of the collections and harnessing the computer networks to access these collections. However, DIR brought many challenges for IR to be addressed, such as the content description for each collection unsimilar, the collections to be searched must be deterministic, and the results to be merged and ranked is diverse. The process of the DIR and centralized IR are similar, except that a query is distributed among different location in DIR [9,10].

2.2 Grid Computing

Grid computing is a technology that provides the infrastructure for applications handling huge volumes of data or need large computational resources [11]. The simulation and modeling of complex systems in a scientific research field are considered as an example of these applications [12]. Moreover, grid technologies offer tools and middleware to utilize the distribution and the diversity of resources to facilitate the interaction between the end user and the grid resources [13]. However, the requirement for systems that analyze and handle data stored in distributed and heterogeneous locations had motivated to develop the data grid. The current progress in the powerful middleware services development for grid computing provides a way to address the IR systems, whereas IR will provide the grid computing new methods for processing and accessing the information. Moreover, grid utilizes heterogeneous systems together into the mega-computer, and therefore, can dedicate to a task that requires big computational power. Furthermore, the grid virtualizes a different and distributed resources, thus the user will deal with the application in grid computing as a single and local computer with vast and powerful resources [14].

2.3 Information Retrieval on Grid Technologies

The Open Grid Forum (OGF) initiated an official document to create standards for IR based on grid computing. Furthermore, the Association for Computing Machinery (ACM) attempted to initialize standards for grid-based IR by adding to the agenda of the ACM's special interest group in IR [9]. Moreover, the IR can be deployed on the grid as a service or as a job. In one hand, the jobs are batched and it sends to grid computing to execute intensively using computational power [15]. On the other hand, the IR is implemented as a service when IR is considered as a part of the grid computing. These services are distributed among the nodes in grid computing, which provide retrieval functionality by preparing local indices in each node to replay for user queries. However, the IR model in general creates a search index, which contains all the information about the collections and the relation between the features and the documents. The process of retrieval is performed by searching these indices.

3. Grid-based Information Retrieval Techniques

IR is concerned with locating documents from single or several collections that are relevant to a user's needed information [1]. IR research deals with all aspects of different processes, including tools and methods for indexing, query processing, document representation languages and models, and crawling document collections (such as the Web), etc., [10]. However, the goal of multimedia IR systems is to aid the users to retrieve the information from multimedia databases taken into account-related features extracted during the preprocessing phase of multimedia databases. Furthermore, the quantity of items saved in the system makes the task complex because it is normally dealing with large volumes of data image or video databases [16]. However, grid computing helps the computer science researcher to use presented infrastructure to run the intensive computational jobs more fast such as extraction of image/video features from big multimedia databases, as illustrated in Figure 1. Furthermore, several techniques were proposed to provide flexible methods for IR as a vast distributed data by harnessing grid computing capabilities. The next sections will review and classify grid-enabled CBIR techniques.

3.1 Content-based Medical Image Retrieval

The current radiology departments usually produce a large number of images every day, which need infrastructure with capabilities for handling this huge data. Furthermore, physicians may require to treat and retrieve image files such as CT scans and X-ray images for analysis or to match up medical cases. As a result, the main objective of content-based image retrieval systems is to aid the physicians to diagnose cases by retrieving similar images/cases from different image collections. However, the grid technologies provide suitable environment to solve the problems of the image retrieval and visual feature extraction [17].

The GNU image finding tool (GIFT) [18,19] is a content-based image retrieval system based on grid computing. The main aim of the GIFT is to apply the grid technologies to harness the huge number of 6000 computers that exist in Geneva hospitals as infrastructure for research tasks. Moreover, the objective of this research is to identify potential of the grid technology and how the medical applications benefit from existing computing and storage resources. The GIFT system uses advance resource connector (ARC) [20], the grid middleware for federating computing resources offered by the KnowARC research project. Furthermore, the research tried to increase the performance speed of the system by assigning the computation intensive jobs such as visual feature extraction and indexing from image databases to grid computing. Furthermore, the harnessing grid will investigate further complex feature spaces and allow the image dataset size to grow to large scale. However, to test the middleware, a Linux operating system was installed on desktop windows machines as virtual machines using virtual machine ware (VMware), which will affect the speed of response time. Moreover, the number of computers available also depends on the idle time of the client computers. Furthermore, the research exploited a dataset provided by the ImageCLEF medical image retrieval task and almost 70 000 images in 2007 were stored in a server. However, the accessing of this server from desktop computer produces bottleneck problems and affect the response time and the scalability of the system.

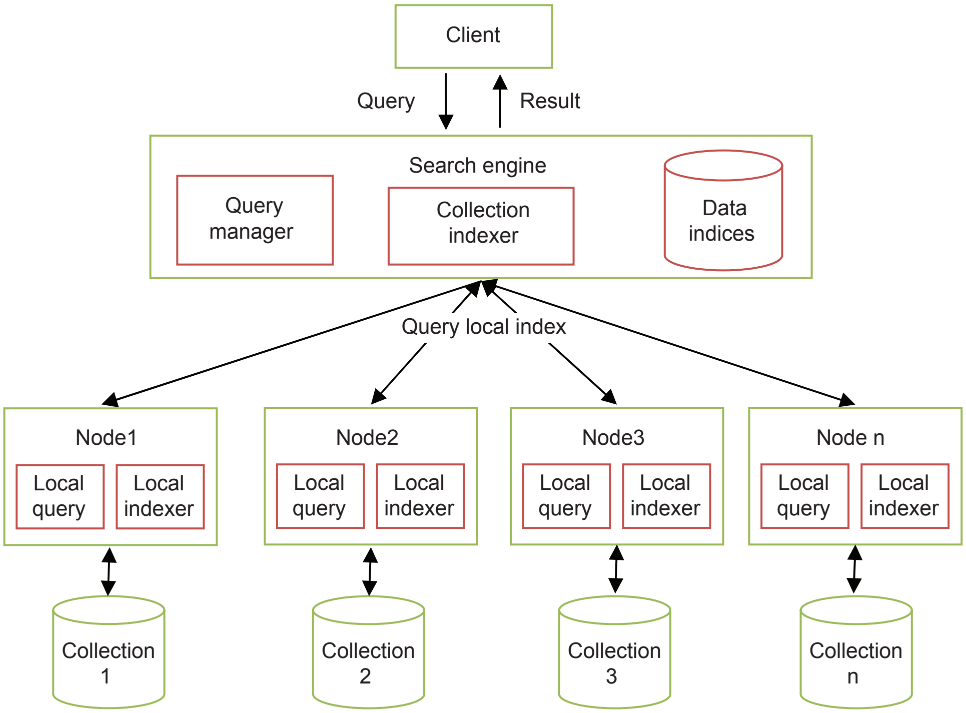

Figure 1: General architecture distributed information retrieval based on grid computing.

Yang *et al.* [21,22] proposed Picture Archiving and Communication System (PACS) [23] based on grid computing, and apply MIFAS (Medical Image File Accessing System) as a mechanism to retrieve and search medical image's databases from a co-allocation data grid. Furthermore, the research harnesses the MIFAS to reduce the transfer time of medical images, and use the co-allocation data grid infrastructure to implement the medical image technique. The experiment carried out on TIGER grid composes of ten clusters, and are distributed among seven institutes in Taiwan. Moreover, the experiments are conducted to test the performance of MIFAS by using the local grid node to simulate a WebPACS, and apply anticipative recursively-adjusting mechanism (ARAM) [24] to test and to compare it with the MIFAS. The system also implements the cyber agent service, tool used in grid to transfer data and manage the replica, that allow users to harness co-allocation data grid and to submit user queries and retrieve image results. The test was conducted to compare the performance of the retrieval process between cyber agent transformer and WebPACS by using small size of image data and the result shows that ARAM has better performance than WebPACS. The research focuses on exchange and retrieves medical images among several hospitals; unfortunately different institutions do not share their medical data with outside hospitals. Moreover, during transfer file's process, if any file change or update will cause incorrect results. Furthermore, the system uses replica mechanism to provide different data source and to increase the response time, but the growing of the replica system must be faced with increasing the storage sizes, and will affect the scalability of the system and managing of the data.

3.2 Content-based Image Retrieval

Content-based image retrieval techniques provide methods to choose data that look like a precise query over all information that exist in a database. Nevertheless, the growing cost of CBIR operations such as feature extraction and image processing affect the scalability of these techniques [25]. However, the large-scale distributed resources offered by grid computing provides best solutions to execute the CBIR tasks.

Town and Harrison [26] illustrated how a large distributed processing Grid (GridPP) is utilized to implement a variety of CBIR techniques (provided by Imense Ltd company) to a large number of images. The study performed high throughput with small overheads by federating the computation jobs among a very large number of grid resources. Moreover, the applying job management and submission framework such as Ganga [27] provides a way to deploy a large number of Imense's image analysis techniques to the grid computing. Furthermore, the Imense computational resources utilize GridPP to analyze the content and construct a searchable index from 25 million high-resolution images that are processed using Imens's computing infrastructure. The experiments were conducted using Condor processing pools [28], LCG [29], and gLite [30] middleware to compute the overhead caused by virtualization solutions. Nevertheless, the analysis of particle physics Grid architecture shows some problems related to the performance bottlenecks, which affect the job running time to go beyond the mean and prevent the system from being scalable. On the other side, the main performance criterion in content image analysis is an overall throughput performed by the technique. This criterion is measured by the number of images that could be processed in a specific time frame regardless of the time spent to handle any certain image. Moreover, the authentication mechanism in the grid proxy server causes network traffic problems that create some delay in response time.

Robles *et al.* [16] focused on measuring the ability to utilize grid computing resources to implement content-based image's retrieval systems by applying retrieval technique called wavelets [31]. The user starts the image search process by selecting an image as a search query and the system calculates its signature. This step is followed by comparing the created query signature with entire DB image's signature by applying metric based on the Euclidean distance, which produce a list P of the most similar images. The P list is then sorted and ranked by using the bubble sort algorithm with $O\,(NP\,log\,(P))$ order, whereas N is the number of images. In the last step, the system provides the most similar dataset images from the P list for the user. Moreover, the research conducts an experiment by using very heterogeneous nodes to measure the efficiency of the system and the overlap of execution time with communication overhead. The result of the experiment demonstrated that the amount of overlap was approximately constant, which proof the scalability of the system with regard to the size of the database. However, the node response time is less than the overhead values produced from grid infrastructure, which reduce the efficiency of the system.

Chatterjee *et al.* [32] proposed distributed multimedia data management architecture that is able to use the grid nodes to manipulate and save the multimedia data. Additionally, a K-NN-based algorithm was proposed for content-based similarity search and included in a distributed query management technique. Moreover, stochastic construct called markov model mediator is used to create a semantic relationship inside the query processing and index structure. Furthermore, to facilitate the implementation of the proposed framework in a grid computing, several grid services are introduced, such as load balancing and semantic relationship. Additionally, the query processing interface is implemented to receive the user query and distribute the query among the grid nodes. Moreover, the content-retrieval engine is implemented in each grid node to execute the user query algorithm and find a semantic relationship by using K-NN-based similarity search. However, the framework does not support auto failure detection as well as recovery of the multimedia data nodes.

3.3 Grid-enabled Search for Spatial and Geospatial Data

Spatial information consists of huge data, dynamic, and multi-dimensional produced by several and diverse organizations [33]. The heterogeneous attributes of spatial information such as data format, technology, and data type prevent the traditional search engine from searching these spatial data sources. The process of searching different spatial data sources is normally performed on desktop computers, which have small computing power compared with the size of the data sources. However, the grid infrastructure provides a large number of computing and storage resources, which offered new opportunities to be utilized by the scientific communities [34].

Zhang *et al*. [33] used the grid computing and grid services to develop and implement collaborative and rapid spatial search engine for distributed and large spatial data. Furthermore, the grid dispatch service is implemented using the ProActive grid tool, which was developed using Java libraries for distributed and parallel computing [35]. In addition, the proposed search engine uses multiple map servers to offer simultaneously services collaboration, because the single server fails to provide fast mapping services for massive data. Moreover, the search engine implements slice scheduling method to offer fast mapping services, which is dividing the map request depending on the map level applied to split the jobs that need long time to run small sub-jobs. Furthermore, the search engine decreases sub-jobs run time and raises the search response time by dedicating every sub-job to execute on one grid node in parallel mode. Additionally, the distributed map servers and the grid scheduler are structured as cluster of mapping services, which are parallelized among several servers based on grid to speed up map service. The search engine is developed to acquire spatial information distributed among 300 cities in China and hide the location and the source of the data from the users. However, the search engine applys master/slave architecture in which the master node is grid scheduler and responsible for job allocation mapping service, and collects the results from different grid nodes which represent the slave nodes.

Di *et al.* [36] and Chen *et al.* [37] proposed a research that combines open geospatial consortium (OGC) standard [38], web-based geospatial, and the Globus grid computing middleware [39]. The geospatial based on grid computing implemented in the research provides personalized, interoperable, on-demand data access and services for massive geospatial data repositories. Moreover, the geospatial grid can use the OGC Catalog Service for Web (CSW) as a geospatial data catalog service to provide data discovery, data search, and other catalog services. Therefore, the goal of developing Grid-enabled CSW (GWCS) service is to support geospatial modeling based on service and to offer a catalog service for the real geospatial data. Moreover, the CSW server relay on the XML means that the entire classes in catalog service is embedded in the requests and responses of the services that are XML classes. Additionally, the user starts geospatial data-accessing process by acquiring data catalog to locate the required data with exploit OGC CSW protocols. Furthermore, the CSW portal broadcast user requests GCSW catalogs on the grid, and then merges the results sent by different catalogs and return it to the user. Depending on the query results, the user creates a WCS data retrieval request to bring the data from grid nodes. However, the experiment conducts a test whether the way of execution and the size of request and response load affect the request-response time, by calculating and comparing the same OGC requests for both grid and web service request-response time. Consequently, the result of the comparison shows that the secure grid service takes long time more than web service because the grid service creates a performance overhead greater than the web service.

3.4 Content-based Video Retrieval

Content-based video retrieval (CBVR) techniques aid the users to search very huge video repositories that map to a specific topic, place, etc. and return the similar video sequence. However, the daily covering of media (mainly videos) increases the multimedia collection size to a very large collection. Furthermore, the main problems faced by the big multimedia collections, among the other problems, are storing and managing of video data. Consequently, the solution should consider requirements for a large storage space and federate management systems for this information that may belong to diverse organizations. Additionally, the preprocessing phase is required by multimedia data to perform a tagging and classification tasks request for particular hardware and software [40].

A grid and content-based video retrieval (GCViR) system proposed by Toharia *et al.* [40,41] is a system to utilize the grid infrastructure to manage and access video datasets by storing and retrieving the video contents. The system automate all processes such as the video uploading, segmentation, and storing. Moreover, the search start using image query, and the system retrieves the prior-stored sequences, which contain all the images based on user query. The search is conducted using video retrieval (VR) service, which receive the clip or image as input and return a group of most *N* shots similar to user input where a user can define before query the value of N. Furthermore, the system has many services to execute search queries such as feature extraction (FE), feature comparison (FC), and reliable file transfer (RFT). The function of FE service is to extract the features of the user query, then the RFT service sends and obtains shots concurrence from the user to grid nodes, and finally the FC service is compared with between the features of the shots stored in the grid resources and the input data, in order to obtain a similarity value that will be used to select the most similar shots. Furthermore, two experiments were conducted to test the system. The first one is video retrieve service VR to identify the number of grid nodes used to execute the user query that affects the response time of the system. However, the growth of the nodes is not considerable because the size of the database used in the experiment is increased depending on the number of participating nodes. The second experiment is conducted to examine an effect of the number of shots that exist in the system in response time, but the result is the same as the first experiment.

3.5 E-learning Multimedia Retrieval

E-learning is a term used for computer-based guidance and teaching supplies, email, online meeting, discussion forums, and other related means [42]. The number of multimedia files involved in the e-learning is big that requires a big amount of storage capacity, high-speed network bandwidth for rapid access of e-learning resources, and efficient streaming-learning process to carry the learning materials to places away from a classroom. Among the previous requirements, grid computing deals with the first and second requirements [43], because grid computing can solve the conventional e-learning limitations like the availability and efficiency.

Taxonomic Indexing Trees (TI-trees) [44] is an indexing structure, designed to rearrange sharable content object reference model (SCROM) [45] documents depending on their related metadata, and to harness the centralized indexing structure of grid infrastructure. The TI-trees are constructed by a local TI-tree for every learning object repository and then merge these trees into a universal TI-true. The search phase utilizes the Grid Information Retrieval (GIR) algorithm by submitting a query to a grid portal from a user. After that, the TI-tree is searched for relevant teaching materials and required contents that are collected from the local locations and returned back to the user. After that, the results of the query are ranked depending on the similarities to the user's query. Furthermore, a centralized index that is created by reorganizing the current documents using bottom-up technique is utilized to accelerate the searching process. The main reason for using bottom-up techniques is appropriate for the master/slave grid model and can efficiently gather the information of document collections from all locations in the grid. Moreover, the reliability and efficiency of the search is growing, whereas the structural information and metadata of document collection are kept on the indexing structure. Additionally, the experiments were conducted to test the performance of GIR and conventional DIR by using the TIGER grid infrastructure. The experiments show that the query processing performance of the DIR is better than GIR. Nevertheless, the search process is slowing down because the query is distributed among different sites in different locations, and the results returned from these different locations are merged again.

The goal of the layered architecture proposed by Shih *et al.* [46] is to speed up the retrieval process by applying ontology technique to retrieve and organize the content from learning collections on Data Grids. Furthermore, the research used the ontology-based semantic search to raise an accuracy of the search, and applied ontology-based indexing to decrease the time of the search process. The index creation phase is to arrange learning contents kept in several repositories by designung a bottom-up technique, and based on a built ontology. A semantic search becomes easy by building global index based on ontology. On the other hand, the search phase is performed by using component of Search Engine, which receive the user's queries, process the queries, and return the results of the queries. When the users get the result about the documents location, the Search Engine retrieves the documents for users from the site that the documents are saved in. Furthermore, the Search Engine uses the ontology to propose the other keywords in vocabulary if the keywords in a query that is submitted by a user do not exist in the vocabulary. Additionally, the searching efficiency and precision is increased by storing the structural and metadata information in the indexing structure. The meta search approach that is studied in the context of DIR [47] has two phases; query distribution and result merging phase. Nevertheless, the search process is less efficient in DIR, while distributing indexing is not appropriate to apply on general grid architectures. The motivation of the semantic web is to utilize the ontology to solve the keyword-based search problems. If the entire information corpus can be completely represented as an ontology-driven knowledge, it is possible to use pure Boolean retrieval model based on ontology. However, the ranking criteria are not clear in Boolean search, which affect the system performance and become useless if the space of retrieval is very large. Moreover, the similarity function of the system that is used to measure the level of the relevant of the two teaching material must be defined.

The goal of the aided contents supporting service (ACSS) proposed by Wand *et al.* [48] is automated to augment learning contents in examination subsystem. The main task of the ACSS is to extract the exam item concepts beside query of the other search engine, such as Yahoo, to provide for the user further contents. The similarity module stores the results of the search as aided content and computes the similarity between aided contents and learning contents. Additionally, the ACSS use grid computing to run a search and similarity tasks as a job and apply first-come first-served algorithm to assign jobs to nodes. Furthermore, the word segmentation system provides term sets corresponding to every aided and learning contents. However, using one node as the word segmentation system produces a bottleneck problem and reduce the scalability of the system. Additionally, the experiment results show that the growing number of users will increase the waiting time beside and increase the makespan time for the search process.

4. Comparative Analysis

In this section, a comparative analysis of the reviewed grid-based searching techniques is elaborated.

Scalability: All those systems use grid computing for extracting features, indexing and searching collection's information, so as to speed-up query execution and to increase scalability of the techniques. However, some of these systems suffer from bottleneck problems, which affect the response time of the system and the scalability feature of the searching technique. As a result, the users spend more time in searching and harvesting the data, indexing the data, and the ratio of system failure will increase with the number of user queries.

Retrieval technique: Within this criterion, there are diverse ways to implement the search techniques. Furthermore, all the systems use IR tools and applications in a parallel way to search dataset based on the grid infrastructure. However, all these algorithms and techniques exist and are used in another field and are not designed and dedicated to be used in grid computing, but adapted to be suitable to work in grid computing. Moreover, the proposed solutions do not consider the grid infrastructure and attributes such as resource's heterogeneity and dynamicity, which affect the performance of the most systems.

Scope: The scope and the area covered and accessed by retrieval application is affected in performance and efficiency of the system. The scope of CBIR system is implemented using different methods. On one hand, the systems implemented in multiple geographic locations such as cities are affected with the network traffic problems. On the other hand, the systems implement in organization scale in small geographic area like one city obtain good performance and affect only the size of the datasets.

Response time: The good technique has fast response time, and the performance of the systems are measured based on their response time. However, most of the proposed researches do not support the real-time search engine, instead take a long time to replay the user queries. Moreover, the main factors that affects response time are number of node and dataset size. The increasing number of the nodes that participate in the search process will fast the response time, whereas the increase of dataset size will slow the response time.

Query type: The majority of the systems use keyword-based query while the minorities of the system use different types of query such as image-based. Therefore, keyword-based type provides fast response time for the user queries and most of the systems use text to index and categorize the dataset even for the image and video datasets.

Data types: The searching systems search the information in a different format depending on the type of the dataset. Furthermore, the majority of the datasets are not database management systems, but it is a file (XML, HTML, image, video, etc.), meaning that the query processing will not be useful to search these files. Additionally, the searching technique is dedicated to certain dataset such as the GCViR system search for video documents, while some systems search for image documents. However, all of them use grid computing as a tool to facilitate the search and harvest the information across the federated locations.

Middleware: It is observed that the majority of the grid computing infrastructures use the de-facto middleware, Globus tool kit, because the Globus is implemented in many standards such as Open Grid Services Architecture (OGSA), Open Grid Services Infrastructure (OGSI), Web Services Resource Framework (WSRF), and Job Submission Description Language (JSDL) [49]. Furthermore, Globus provides most services and components required to implement application on grid or managing the infrastructure [49]. However, a few other systems use another middleware such as gLite, LCG, and ARC middleware's.

5. Conclusion and Future Work

In this paper, we reviewed and analyzed different grid-based, CBIR techniques for large dataset. First, brief description of DIR, grid computing, and IR on the grid are discussed. Then, we state in some details on how the existing CBIR systems can satisfy some very significant criteria such as scalability, retrieval technique, response time, data types, and deployed middleware. The paper presented classifies grid-based CBIR literature into medical image retrieval, image retrieval, VR, E-learning multimedia retrieval, and spatial image retrieval. Moreover, the paper explained every category in detail to provide insight into their distinctive methods of achieving grid-based CBIR.

The rapid growing size of digital collections produces several challenges in the field of IR such as collection's discovery, standardization of interfaces, collection's management, cost optimization, and privacy issues. Moreover, access to document collections needs efficient data management and searching techniques. To build an effective IR system, we need a huge amount of a storage resource. Furthermore, to index and extract document features require a very big computational resources to speed up the processes and to increase the performance of the system.

From the investigation and analysis carried in the paper, it can be concluded that existing CBIR approaches still suffer from major incompleteness; such as supporting more query types, delay in response time, and maintaining dynamic datasets. Consequently, systems that provide an easy way to access grid-based dataset should be further strong, efficient, and scalable. The reality shows most of the effort in the field of large-scale distributed dataset work with the parallel processing of small datasets at the same time as assurance data consistency has raised some concern as to whether grid-based artifacts are practical solutions for massive dataset CBIR. This is an area of further research. There are a number of grid-based systems that exist at this time, and IR requirements are to address those requirements to every type that differ as much as the types of grid systems themselves. This review paper will help in upcoming research by attempting to identify, investigate, and address these issues.

6. Acknowledgment

This work is supported by the International Doctoral scholarship (IDF) in Universiti Teknologi Malaysia. The authors thank the Research Management Centre (RMC) Universiti Teknologi Malaysia (UTM) for supporting this research.

References

1. C.D. Manning, P. Raghavan, and H. Schutze, "Introduction to information retrieval," Vol. 1, Cambridge: Cambridge University Press; 2008.
2. T.T. Avrahami, L. Yau, L. Si, and J. Callan, "The FedLemur project: Federated search in the real world," Journal of the American Society for Information Science and Technology, Vol. 57, pp. 347-58, 2006.
3. B. Mishra, and S. Dehuri, "Parallel computing environments: A review," IETE Technical Review, Vol. 28, pp. 240, 2011.
4. M.J. Litzkow, M. Livny, and M.W. Mutka, "Condor-a hunter of idle workstations," in 8th International Conference on Distributed Computing Systems, 1988., pp. 104-11, 1988.
5. I. Foster, C. Kesselman, J.M. Nick, and S. Tuecke, "The Physiology of the Grid," in Grid Computing, John Wiley & Sons, Ltd, pp. 217-49, 2003.
6. I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations," International Journal of High Performance Computing Applications, Vol. 15, pp. 200-22, 2001.
7. F. Can, I.S. Altingövde, and E. Demir, "Efficiency and effectiveness of query processing in cluster-based retrieval," Information Systems, Vol. 29, pp. 697-17, 2004.
8. W. Meng, C. Yu, and K.L. Liu, "Building efficient and effective metasearch engines," ACM Computing Surveys (CSUR), Vol. 34, pp. 48-89, 2002.
9. R. Baraglia, D. Laforenza, and F. Silvestri, "SIGIR workshop report: the SIGIR heterogeneous and distributed information retrieval workshop," SIGIR Forum, Vol. 39, pp. 19-24, 2005.
10. R.R. Pollán, and Á. Barreiro, "Enabling the Grid for Experiments in Distributed Information Retrieval," presented at the First EELA-2, Bogotá, Colombia, 2008.
11. K. Krauter, R. Buyya, and M. Maheswaran, "A taxonomy and survey of grid resource management systems for distributed computing," Software: Practice and Experience, Vol. 32, pp. 135-64, 2002.
12. W.E. Johnston, "Computational and data Grids in large-scale science and engineering," Future Generation Computer Systems, Vol. 18, pp. 1085-100, 2002.
13. T. Ma, S. Shi, H. Cao, W. Tian, and J. Wang, "Review on Grid Resource Discovery: Models and Strategies," IETE Technical Review, Vol. 29, pp. 213-22, 2012.
14. S. Bourbonnais, V. Gogate, L. Haas, R. Horman, S. Malaika, and I. Narang, *et al*., "Towards an information infrastructure for the grid," IBM systems journal, Vol. 43, pp. 665-88, 2004.
15. A. Chowdhury, and G. Pass, "Operational requirements for scalable search systems," in Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, 2003.
16. O. Robles, J. Bosque, L. Pastor, and Á. Rodríguez, "CBIR on Grids" in On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE. Vol. 4276, R. Meersman and Z. Tari, Eds., ed: Berlin/Heidelberg: Springer; pp. 1412-21, 2006.
17. H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," International Journal of Medical Informatics, Vol. 73, pp. 1-23, 2004.
18. X. Zhou, M.J. Pitkanen, A. Depeursinge, and H. Müller, "A Medical Image Retrieval Application Using Grid Technologies To Speed Up Feature Extraction in Medical Image Retrieval," Philippine Journal of Information Technology, Vol. 2, Issue. 1, pp. 3-9, 2009.
19. M. Niinimaki, X. Zhou, A. Depeursinge, A. Geissbuhler, and H. Muller, "Building a community grid for medical image analysis inside a hospital, a case study," in MICCAI-Grid, pp. 3-12, 2008.
20. M. Ellert, M. Grønager, A. Konstantinov, B. Kónya, J. Lindemann, and I. Livenson, *et al*., "Advanced Resource Connector middleware for lightweight computational Grids," Future Generation Computer Systems, Vol. 23, pp. 219-40, 2007.
21. Y. Chao-Tung, C. Chiu-Hsiung, Y. Ming-Feng, and C. Wen-Chung, "MIFAS: Medical Image File Accessing System in Co-allocation Data Grids," in Asia-Pacific Services Computing Conference, 2008. APSCC '08. IEEE, pp. 769-74, 2008.
22. C.T. Yang, C.H. Chen, and M.F. Yang, "Implementation of a medical image file accessing system in co-allocation data grids," Future Generation Computer Systems, Vol. 26, pp. 1127-40, 2010.
23. G.V. Koutelakis, and D.K. Lymperopoulos, "A Grid PACS Architecture: Providing Data-centric Applications through a Grid Infrastructure," in Engineering in Medicine and Biology Society, 2007. 29th Annual International Conference of the IEEE, pp. 6429-33, 2007.

24. C.T. Yang, M.F. Yang, and W.C. Chiang, "Enhancement of anticipative recursively adjusting mechanism for redundant parallel file transfer in data grids," Journal of Network and Computer Applications, Vol. 32, pp. 834-45, 2009.

25. A.J. Plaza, J. Plaza, and A. Paz, "Parallel heterogeneous CBIR system for efficient hyperspectral image retrieval using spectral mixture analysis," Concurr. Comput.: Pract. Exper., Vol. 22, pp. 1138-59, 2010.

26. C. Town, and K. Harrison, "Large-scale grid computing for content-based image retrieval," in Aslib Proceedings, pp. 438-46, 2010.

27. J.T. Mościcki, F. Brochu, J. Ebke, U. Egede, J. Elmsheuser, and K. Harrison, *et al.*, "Ganga: A tool for computational-task management and easy access to Grid resources," Computer Physics Communications, Vol. 180, pp. 2303-16, 2009.

28. J. Frey, T. Tannenbaum, M. Livny, I. Foster, and S. Tuecke, "Condor-G: A Computation Management Agent for Multi-Institutional Grids," Cluster Computing, Vol. 5, pp. 237-46, 2002.

29. M. Lamanna, "The LHC computing grid project at CERN," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Vol. 534, pp. 1-6, 2004.

30. E. Laure, F. Hemmer, F. Prelz, S. Beco, S. Fisher, and M. Livny, *et al.*, "Middleware for the next generation Grid infrastructure," Computing in High Energy Physics and Nuclear Physics (CHEP 2004), 2004.

31. O.D. Robles, A. Rodríguez, and M.L. Córdoba, "A study about multiresolution primitives for content-based image retrieval using wavelets," in IASTED International Conference On Visualization, Imaging, and Image Processing (VIIP 2001), Marbella, Spain, pp. 506-11, 2001.

32. K. Chatterjee, S.M. Sadjadi, and S.C. Chen, "A Distributed Multimedia Data Management over the Grid," in Multimedia Services in Intelligent Environments. Vol. 2, G. Tsihrintzis, et al., Eds., ed: Berlin/Heidelberg: Springer, pp. 27-48, 2010.

33. J. Zhang, C. Yang, Y. Zhu, and Y.C. Ren, "Research on massive and distributed spatial information collaborative search engine based on grid," in 8th International Conference on Geoinformatics, pp. 1-4, 2010.

34. G. Giuliani, N. Ray, and A. Lehmann, "Grid-enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities," Future Generation Computer Systems, Vol. 27, pp. 292-303, 2011.

35. L. Baduel, F. Baude, D. Caromel, A. Contes, F. Huet, M. Morel and *et al.*, "Programming, Composing, Deploying for the Grid" in Grid computing: Software environments and tools, J. C. Cunha and O. F. Rana, Eds., ed: London: Springer; pp. 205-29, 2006.

36. L. Di, A. Chen, W. Yang, Y. Liu, Y. Wei, P. Mehrotra, and *et al.* "The development of a geospatial data Grid by integrating OGC Web services with Globus-based Grid technology," Concurrency and Computation: Practice and Experience, Vol. 20, pp. 1617-35, 2008.

37. A. Chen, L. Di, Y. Bai, Y. Wei, and Y. Liu, "Grid computing enhances standards-compatible geospatial catalogue service," Computers and amp; Geosciences, Vol. 36, pp. 411-21, 2010.

38. Open Geospatial Consortium, "OGC reference model," in On-line Avaiable from: http://www.opengeospatial.org/standards, [Last accssed in 2013 Jan 22].

39. I. Foster, "The globus toolkit for grid computing," in First IEEE/ACM International Symposium on, Cluster Computing and the Grid, pp. 2, 2001.

40. P. Toharia, A. Sánchez, J. Bosque, and O. Robles, "Efficient Grid-Based Video Storage and Retrieval" On the Move to Meaningful Internet Systems: OTM 2008, R. Meersman, and Z. Tari, Editors., Vol. 5331, Berlin/Heidelberg: Springer; pp. 833-51, 2008.

41. P. Toharia, A. Sánchez, J.L. Bosque, and O.D. Robles, "GCViR: grid content-based video retrieval with work allocation brokering," Concurrency and Computation: Practice and Experience, Vol. 22, pp. 1450-75, 2010.

42. E.T. Welsh, C.R. Wanberg, K.G. Brown, and M.J. Simmering, "E-learning: emerging uses, empirical results and future directions," International Journal of Training and Development, Vol. 7, pp. 245-58, 2003.

43. M.B. Geetha Manjusha, S. Jaganathan, and A. Srinivasan, "gBeL: An Efficient Framework for e-Learning Using Grid Technology" Computer Networks and Information Technologies." In: V.V. Das, J. Stephen, and Y. Chaba, Editors., Vol. 142, Berlin/Heidelberg: Springer; pp. 63-68, 2011.

44. W.C. Shih, S.S. Tseng, and C.T. Yang, "Using taxonomic indexing trees to efficiently retrieve SCORM-compliant documents in e-learning grids," Educational Technology and Society, Vol. 11, pp. 206-26, 2008.

45. O. Bohl, J. Scheuhase, R. Sengler, and U. Winand, "The sharable content object reference model (SCORM)-a critical review," in Proceedings of International Conference on Computers in Education, 2002., Vol. 2, 2 pp. 950-51, 2002.

46. W.C. Shih, C.T. Yang, and S.S. Tseng, "Ontology-based content organization and retrieval for SCORM-compliant teaching materials in data grids," Future Generation Computer Systems, Vol. 25, pp. 687-694, 2009.

47. C. Yu, K.L. Liu, W. Meng, Z. Wu, and N. Rishe, "A methodology to retrieve text documents from multiple databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 14, pp. 1347-61, 2002.

48. C.M. Wang, M.C. Chiang, C.S. Yang, and C.W. Tsai, "Aided contents supporting service for e-learning systems," International Journal of Innovative Computing, Information and Control, Vol. 7, pp. 4005-26, 2011.

49. I. Foster, "Globus Toolkit Version 4: Software for Service-Oriented Systems," Journal of Computer Science and Technology, Springer Boston, Vol. 21, pp. 513-20, 2006.

AUTHORS

Mohammed Bakri Bashir received the B.Sc. from SUST University, Sudan, and M.Sc. from University of Gezira, Sudan. He is a lecturer at University of Shendi. Currently he is a PhD candidate at Faculty of Computing, Universiti Teknologi Malaysia, Malaysia. He is also member of PCRG research group. His research interests include computer network, distributed systems, grid computing, cloud computing, and distributed information retreival.

E-mail: mhmdbakri@gmail.com

Muhammad Shafie Abd Latiff received his Ph.D. from Bradford University, United Kingdom. His research interest is mainly in computer network focusing on grid computing and visualization technology. He is an Associate Professor and the member of Perversive Computing Research Group at the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia.

E-mail: shafie@utm.my

Aboamama Atahar Ahmed received Higher Diploma in computer programming and system analysis from Higher Institute of comprehensive profession in Libya1997 and M.S. degrees in Information and Multimedia Technology from Unitar University Malaysia 2004 and the PhD degree from Universiti Teknologi Malaysia 2010 and He is currently General Manager of Galactic Bridge Company and the head of the research Department. His research focuses on grid computing, distributed System, Scientific Visualization, robotic Development.

E-mail: aboamama@utm.my

Adil Yousif Received the B.Sc. and M.Sc. from University of Khartoum, Sudan. and his PhD degree from UTM University in Malaysia. He is Assistant Professor at Faculty of Computer Science and Information Technology, Kassala University. He is also a member of PCRG research group. His research interests include computer network, distributed systems, grid computing and optimization techniques.

E-mail: adiluofk@gmail.com

Manhal Elfadil Eltayeeb received his M.Sc. in Computer Science from University of Gezira, Sudan, in 2006, and his B.Sc. in Computer Science from University of Khartoum, Sudan, in 2000. He is a PhD candidate in Faculty of Computing, University of Technology Malaysia (UTM), Malaysia. He has worked as a lecturer in King Saud University (KSU), Saudi Arabia. His research interests include distributed algorithms and systems, applied parallel computing, high performance computing, cellular networking, and bioinformatics computing.

E-mail: manhalus@gmail.com

DOI: 10.4103/0256-4602.113511; Paper No. TR 469_12; Copyright © 2013 by the IETE