Future Generation Computer Systems 25 (2009) 687–694

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Ontology-based content organization and retrieval for SCORM-compliant teaching materials in data grids

Wen-Chung Shih[a], Chao-Tung Yang[b,*], Shian-Shyong Tseng[a,c]

[a] Department of Information Science and Applications, Asia University, Taichung, 41354, Taiwan, ROC
[b] High-Performance Computing Laboratory, Department of Computer Science, Tunghai University, Taichung, 40704, Taiwan, ROC
[c] Department of Computer Science, National Chiao Tung University, Hsinchu, 300 Taiwan, ROC

ARTICLE INFO

Article history:
Received 31 May 2007
Received in revised form
31 August 2007
Accepted 29 January 2009
Available online 11 February 2009

Keywords:
Data grid
e-Learning
SCORM
Globus toolkit
Information retrieval

ABSTRACT

With the rapid growth of e-Learning, a tremendous amount of learning content has been developed by numerous providers. Recently, the Sharable Content Object Reference Model (SCORM) has been widely accepted as a standard of e-Learning for users to share and reuse various teaching materials. Data grids, characterized by their goal to manage large-scale dataset, are promising platforms to support sharing of geographically dispersed learning content. However, current data grid standards have not provided complete solutions to content-based information retrieval. To increase the precision of content retrieval on data grids, our idea is to propose an ontology-based approach to organize and retrieve learning content in geographically dispersed repositories. We designed a layered architecture to enable learning content organization and retrieval on data grids, implemented in a metropolitan-scale grid environment. Experimental results show that the proposed approach can precisely retrieve SCORM-compliant learning content.

1. Introduction

With the flourishing development of information technology, e-Learning has grown as an important learning paradigm. To facilitate adaptive and individualized learning, teachers are encouraged to develop adaptive teaching materials for their courses and students. Recently, the Sharable Content Object Reference Model (SCORM) [1] has been widely accepted as a standard of e-Learning for users to share and reuse teaching materials. In addition, many learning content repositories are built by educational institutes and organizations. The issue of "content explosion" will soon be posed to us. Conventional e-learning systems are stand-alone, which manage only their own learning content. In this kind of system, teaching materials are usually stored in a database, named the Learning Object Repository (LOR). However, when several LORs are built in different sites on the Internet, there exists a need to share learning objects across the Internet.

During the past decade, Grid computing [2,3] has emerged to support resource-sharing and to overcome the limitations of computing power and storage capacity in conventional computing platforms. Particularly, data grids connect distributed resources for managing large-scale datasets [4–8]. Many data-intensive applications, such as high-energy physics, bioinformatics applications, etc., require data file management systems to manage replication, transfer and access of data. Therefore, grid computing technologies provide possibilities for supporting innovative applications, such as e-Learning. In fact, more and more effort has gone into the field of e-Learning grids, using grid technologies in the context of e-Learning. Among these, ELeGI (European Learning Grid Infrastructure, 2004–2008) is the most representative project [9].

When content is stored and shared on data grids, there exists a great demand to find desirable teaching materials from multiple repositories in data grids. The problem is similar to information retrieval, which has been widely investigated in the past. A straightforward approach is to apply existing information retrieval methods to learning content retrieval on data grids [10]. However, this approach does not consider the characteristics of learning content and data grids to improve the precision of information retrieval. On one hand, SCORM-compliant learning content is associated with three types of information: text, metadata and structural information, which is different from conventional documents and web pages. On the other hand, data grids are implemented in a layered manner, so applications have to consider collaborative aspects with underlying components to improve their performance. Data grids, characterized by the goal to efficiently manage large-scale datasets, are promising

[*] Corresponding author. Tel.: +886 4 23590415; fax: +886 4 23591567.
E-mail addresses: wjshih@asia.edu.tw (W.-C. Shih), ctyang@thu.edu.tw (C.-T. Yang), sstseng@cis.nctu.edu.tw (S.-S. Tseng).

0167-739X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2009.01.005

platforms for sharing geographically dispersed learning content. However, current data grid standards have not provided complete solutions to content-based information retrieval. The existing information retrieval methods are not designed for data grids, and are inefficient in the process of query dispatching and results collecting, which involves synchronization and communication overhead. Hence, there is an urgent need to design a tailor-made approach to content retrieval which is suitable for data grids.

We have proposed and implemented an ontology-based approach to organizing and retrieving learning content on Data Grids, aiming at fast and precise content retrieval. The main idea is to increase precision by an ontology-based semantic search, and to reduce search time by ontology-based indexing. This framework consists of two phases. In the index creation phase, a bottom-up method is designed to organize learning contents located in different repositories, according to a built ontology. An ontology-based global index is then created to facilitate a semantic search. Next, users' queries are interactively verified in the search phase, and the desired content is retrieved fast and precisely.

The contributions can be summarized as follows. First of all, an ontology-based approach is proposed to solve the learning content management problem on Data Grids, which is not completely investigated by other existing methods. Second, a layered architecture for learning content retrieval on a data grid is proposed. Finally, a prototype is implemented and evaluated on a metropolitan-scale grid environment. Also, experimental results reveal that this approach can improve the performance of content retrieval.

This ontology-based indexing is extended from the method in [11]. However, this paper is significantly different from that work. First, we propose a new architecture to enable data grids to precisely retrieve learning content. Second, the taxonomy-based index is extended to an ontology-based index. Next, the similarity functions of the two papers are different. Finally, the primary goal of this work is precision, while the previous work focused on search time.

The rest of this paper is organized as follows. In Section 2, the background about e-learning and information retrieval is reviewed. Next, Section 3 formulates the model and the problem. Then, the proposed methodology is presented in Section 4. Experimental results are described and discussed in Section 5. Finally, concluding remarks are given.

2. Background review

This section reviews the related background. First, e-learning is introduced. Then, the concepts of information retrieval technologies are described.

2.1. e-Learning and SCORM [1]

With the development in communication and network technology in recent years, under the gradual improvement of network bandwidth and quality, the real-time transmission of high-quality video and audio becomes possible. Therefore, the transmission of multimedia and relative network application technologies have gradually been developed and become popular, such as the technology of Distance Education, Video Conference and Video on Demand.

The advantage of e-Learning is that it can overcome the obstacle of geographical location; making students on remote sites feel that they are like being in the environment of attending classes in a classroom. Moreover, it can save cost and time of the students for their commuting to and fro the classroom.

Although e-Learning has many advantages, the biggest obstacle is the investment on equipment. For instance, a teaching mode that intends to achieve a video/audio teaching effect needs to be established with a costly file storage system or server. To enterprizes or schools with an insufficient budget, it is almost impossible for them to establish the system.

Currently, SCORM is the most popular standard for learning contents, and it is proposed by the U.S. Department of Defense's Advanced Distributed Learning (ADL) organization in 1997. The SCORM specifications are a composite of several specifications developed by international standards organizations. In a nutshell, SCORM is a set of specifications for developing, packaging and delivering high-quality education and training materials whenever and wherever they are needed. In SCORM, the content packaging scheme is proposed to package the learning objects into standard teaching materials. The content packaging scheme defines a teaching materials package consisting of 4 parts, that is, (1) Metadata: describe the characteristic or attribute of this learning content, (2) Organizations: describe the structure of this teaching material, (3) Resources: denote the physical file linked by each learning object within the teaching material, and (4) (Sub) Manifest: describe this teaching material as consisting of itself and another teaching material.

2.2. Search engine technologies

Inverted file indexing has been widely used in information-retrieval [12–15]. An inverted file is used for indexing a document collection to speed up the search process. The structure of an inverted file consists of two components: the vocabulary and the posting list. The vocabulary is composed of all distinct terms in the document collection. For each term, a list of all the documents containing this term is stored. The set of all these lists is called the posting list. However, the structure of a document is not considered in this model.

Storage requirements of inverted indices [16] have been evaluated based on a B+-tree and posting list. Five strategies of the index term replication were discussed. This approach is extended to analyze the storage requirement of the proposed approach in this proposal. In [17], 11 different implementations of ranking-based text retrieval systems using inverted indices were presented, and their time complexities were also investigated.

The meta-search approach has been studied in the context of distributed information retrieval [10]. This approach consists of Query Distribution and Result Merging phases. Furthermore, the Document Retrieval problem is divided into two sub-problems: The Database Selection problem and the Document Selection problem. However, Distributed indexing is not suitable for common Grid architectures. Also, Distributed IR is less efficient in searching.

The use of ontology to overcome the limitations of keyword-based search has been put forward as one of the motivations of the Semantic Web since its emergence in the late 1990s. One way to view a semantic search engine is as a tool that gets formal ontology-based queries from a client, executes them against a knowledge base (KB), and returns tuples of ontology values that satisfy the query. In this view, the information retrieval (IR) problem is reduced to a data retrieval task. While this conception of semantic search already brings key advantages, our work aims at taking a step beyond.

A purely Boolean ontology-based retrieval model makes sense when the whole information corpus can be fully represented as an ontology-driven knowledge-base. But, there are well-known limits to the extent to which knowledge can be formalized this way. A Boolean search does not provide clear ranking criteria, without which the search system may become useless if the retrieval space is too big.

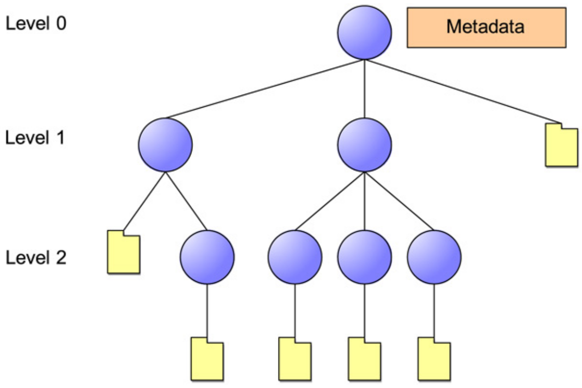


Fig. 1. The representation of an example of teaching material.

3. Problem formulation

Before the problem can be described, some concepts and models have to be introduced. First, a level-wise structural model of a SCORM-compliant teaching material is presented, which is intended to model the semantic concepts from general levels to specific levels. Then, definitions of a learning object repository, a query and a similarity function are described. Finally, the Grid Teaching Material Retrieval Problem is defined.

3.1. A level-wise model of teaching materials

Definition (*Feature Vectors (FV)*). We adopt the Vector Space Model [12] to represent a text document. A text document *td* is represented by a *Feature Vector*, which is a vector of keyword weights. For each keyword *k* appearing in a document *td*, td_k is a weight representing the importance of the keyword *k* in the document *td*. In this paper, this weight is computed by a scheme called TF-IDF (Term Frequency–Inverted Document Frequency) [14] as follows.

$$td_k = \frac{\mathrm{freq}_{k,td}}{\max\limits_y \mathrm{freq}_{y,td}} \cdot \log \frac{|D|}{n_k}. \tag{1}$$

In this formula, $\mathrm{freq}_{k,td}$ is the number of occurrences of k in td, $\max_y \mathrm{freq}_{y,td}$ is the most repeated instance in td, $|D|$ is the number of all documents, and n_k is the number of documents containing k.

Definition (*Teaching Materials (TM)*). A *Teaching Material* represents a SCORM-compliant teaching material, or referred to as a Content Package. It is a rooted tree. The leaf node represents a text document, which is characterized by a feature vector $= \langle w_1, w_2, \ldots, w_{|V|} \rangle$, where $|V|$ is the size of the vocabulary. In this paper, we assume that a controlled vocabulary, V, is used. Internal nodes represent structural information, or level-wise information, of this content. Each teaching material is associated with two types of attribute:

- Level i feature vector, L_i, $(0 < i < \text{Height})$ where L_i = Average of feature vectors of nodes at level i, and Height is the height of the rooted tree;
- Metadata, M_j, $(0 < j < num\_m)$ where M_j is the jth metadata of this teaching material, and num_m is the number of metadata of this content.

An example of a teaching material is shown in Fig. 1, where the tree has three levels.

In the SCORM standard, a learning object repository (LOR) is a database where the teaching materials are stored. We define the LOR as a set of teaching materials.

$$\mathrm{LOR} = \{TM_1^{;}TM_2^{;} \ldots, TM_{n\_TM}\} \tag{2}$$

where n_TM is the number of teaching materials in the LOR.

3.2. A composite similarity function

In order to decide the degree of relevance of two teaching materials, the similarity function has to be defined. It is difficult to choose a suitable similarity function for SCORM-compliant teaching materials, which are characterized by textual content, metadata and structural information. Here, a composite similarity measure for two teaching materials, a and b, is proposed:

$$\mathrm{Sim(a, b)} = \sum_{i=0}^{\mathrm{Height}} \alpha_i \times \mathrm{Sim}_i(\mathrm{a, b}) + \beta \times \mathrm{Sim_M(a, b)}. \tag{3}$$

where the sum of α_i, and β is equal to one.

The similarity function consists of two parts:

- Sim_i: level i similarity function, which is cosine function, $0 < i <$ Height; that is, the similarity between two teaching materials $TM_1 = \langle k_1, k_2, \ldots, k_{|V|} \rangle$ and $TM_2 = \langle p_1, p_2, \ldots, p_{|V|} \rangle$ is measured by the following formula:

$$\mathrm{Sim}_i = \frac{\sum_{i=1}^{|V|} k_i \times p_i}{\sqrt{\sum_{i=1}^{|V|} k_i^2} \times \sqrt{\sum_{i=1}^{|V|} p_i^2}}. \tag{4}$$

- Sim_M: Metadata similarity function, which is (# matched attributes)/ (# all attributes).

Definition (*Query*). A *Query Q* is a feature vector v_Q, where

$$v_Q = \langle q_1, q_2, \ldots, q_{|V|} \rangle. \tag{5}$$

The similarity between a query and a teaching material, which means the relevance of the teaching material to the query, is also determined by (4).

3.3. The grid teaching material retrieval problem

Based on the definitions above, the Grid Teaching Material Retrieval Problem (*GTMRP*) can be described as follows. Given a learning object repository *L*, a query *q* and a similarity function Sim, retrieve relevant teaching materials with respect to the query, ranking by sim. The objective is to improve precision of content retrieval.

This problem can be exhaustively solved by evaluating for each teaching material in the learning object repository. However, this approach will take a tremendous amount of time when the number of teaching materials in the repository is overwhelmingly large. Therefore, we propose a heuristic method, as described in the next section.

4. Methodology

In this section, this methodology is overviewed. First, the layered architecture is depicted. Then, the two phases, index creation and content search, are described.

4.1. Layered architecture

We propose a layered architecture for content retrieval on data grids, as shown in Fig. 2. This architecture is composed of four layers: Network Layer, Resource Layer, Middleware Layer and Application Layer. The content retrieval service is implemented in the Application Layer. The other underlying layers are in charge of dispatching an information retrieval job to the underlying data grids. These works include measuring system loading, disk-using status, analyzing learning content repositories, selecting the

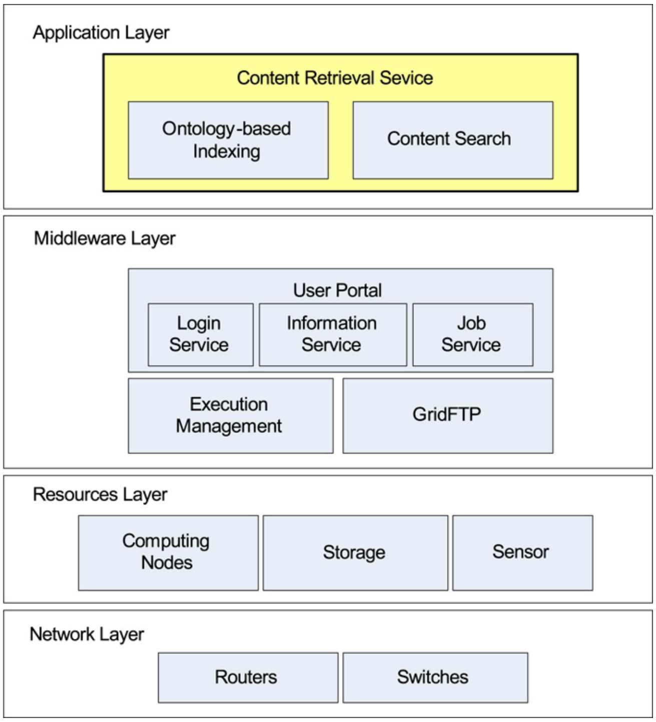


Fig. 2. The layered architecture of ontology-based content retrieval on data grids.

most adaptive grid systems for retrieving content and recording job information. The layered structure and modular design can ease the implementation process and increase the possibility of component reuse.

The framework of content retrieval service consists of two phases: index creation and content search, as shown in Fig. 3. Search Engine is a broker which processes the query by referring to the global index first and then returning the retrieved result to users. Learning object repository (LOR) is a database where local learning objects are stored. The Dewey Decimal Classification (DDC) system is adopted as the ontology [18].

The Search phase is carried out by the search engine component, which receives queries from users, processes the queries, and presents results to the users. After users specify the desired documents from the returned results, the search engine accesses the site where the documents are stored and retrieves these documents for the users. When a user submits a query containing keywords which do not belong to the vocabulary, the search engine will suggest other keywords in the vocabulary based on the ontology.

The purpose of this phase is to minimize the time of query processing and content transmission when retrieving SCORM-compliant documents in a grid. To speed-up the searching process, our idea is to use a centralized index, which is generated by reorganizing the existing documents based on a bottom-up approach, because this approach is suitable for the master-slave grid model and can effectively collect the information of existing documents from all sites in the grid. Furthermore, the indexing structure stores metadata and structural information, which increases the efficiency and precision of searching. To speed up the transmission process, the other idea is to present the ranked results with an estimated transmission time, which is derived from grid monitoring tools. In this way, a document which has high ranking score but has a long estimated transmission times (maybe due to a low-bandwidth link) can be avoided by users.

The process flow of the Search Engine is restated as follows.

1. The user submits a query to Search Engine.
2. The Search Engine refers to the global index.
3. The Search Engine gets the list of desired documents from the global index.
4. The Search Engine retrieves these documents from learning object repositories.
5. The Search Engine presents these documents to the user.

Once the related documents are retrieved, the search engine ranks them according to the similarity between the query and document, as shown in (4). In addition, we provide the other ranking option which is based on the estimated transmission time of these documents.

4.2. Index creation

O-Indices (Ontology-based Indices) are data structures that are designed to support SCORM documents management, which are enhanced from the taxonomic indices in [11]. The main idea is to reorganize SCORM documents according to their associated attributes (such as feature vectors, metadata, etc.), and to utilize the centralized indexing structure for fast search. The attributes and operations of O-Indices are described as follows.

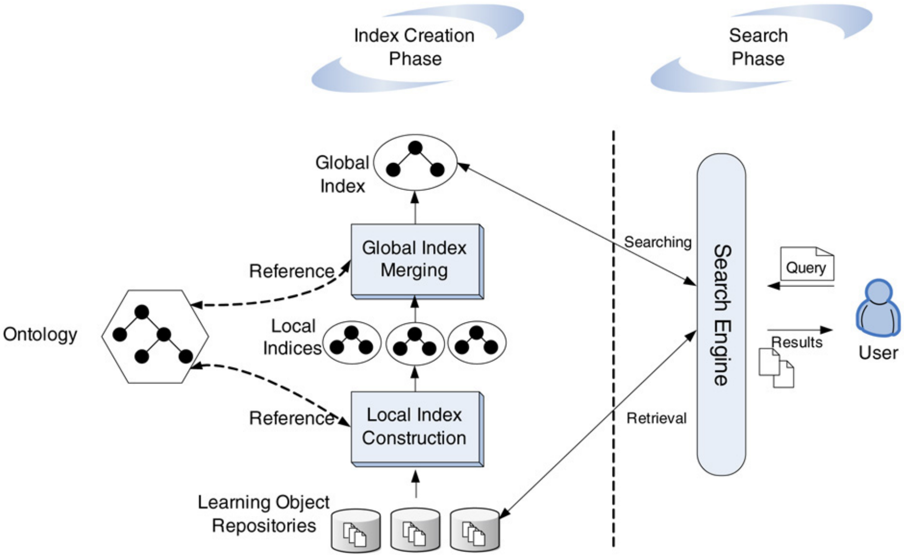


Fig. 3. The framework of ontology-based content retrieval service.

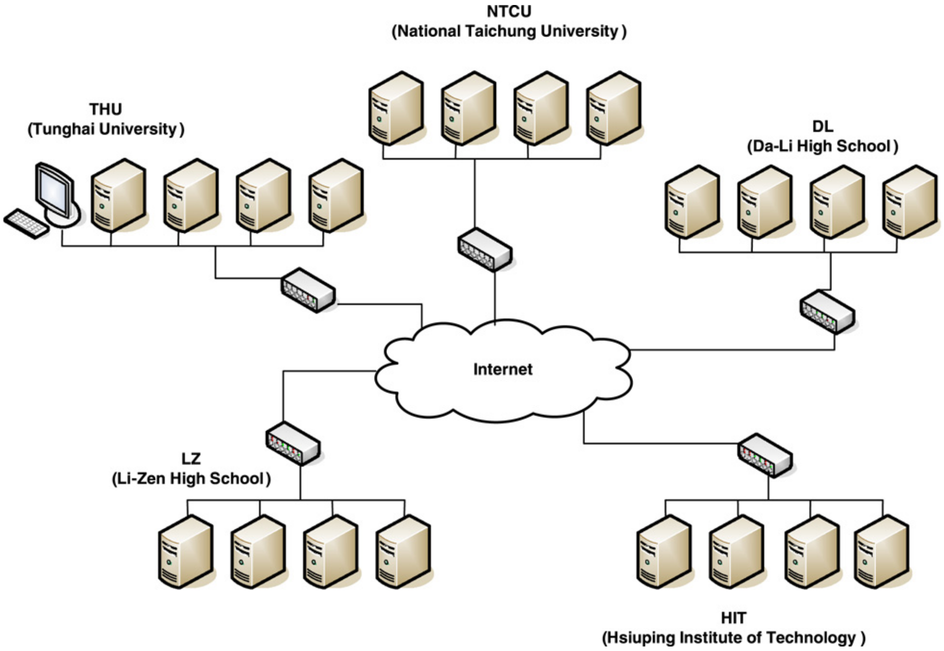


Fig. 4. The data grid test-bed.

Definition (*O-Index*). An *O-Index* is a rooted tree having the following properties:

Every node x has the following fields:

- *x.name*: Class name
- *x.id*: Class ID
- *x.num*: Number of documents
- *x.inv*: Pointer to the inverted index.

Let x be a node in an *O-index*. If y is a child of x, then there exists a relation IS-A between x and y. That is, y IS-A x, which implies that y is a specialization of x.

The operations supported by an O-index include Searching, Construction, Merging and Insertion, which are described as follows.

- Search (T_0, C)
  Given an O-Index T_0 and a class name C, this function returns a pointer x to a class node in T_0 such that $x.id = C$, or *NIL* if no such class belongs to T_0. A common operation performed on an O-index is searching for a query in a sub-tree. If the class of the sub-tree is not specified by users, the searching process will start from the root.
- Construct (LOR)
  Given a Learning Object Repository LOR, a Construct operation returns an O-index. The resulting O-index represents the content packages in the LOR.
- Merge (T_1, T_2)
  Given two O-indices T_1 and T_2, a Merge operation returns a new O-index which includes content packages of T_1 and T_2.
- Insert (T_0, CP)
  Given an O-index T_0 and a content package CP, an Insert operation inserts the content package into O-index T_0.

5. Experimental results

To verify the proposed approach, a prototype is implemented and is evaluated in a metropolitan-scale data grid test-bed. To begin with, our grid environment is illustrated. After that, an experiment with SCORM-compliant teaching materials is conducted, and the results are discussed.

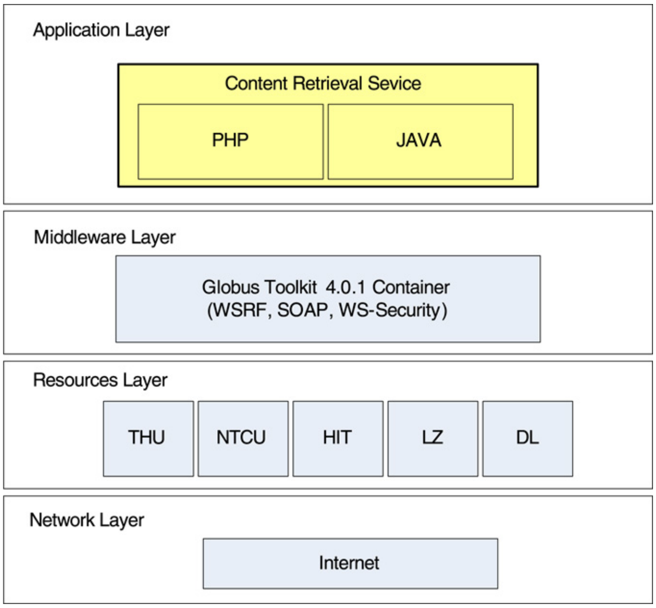


Fig. 5. Software architecture of the prototype system.

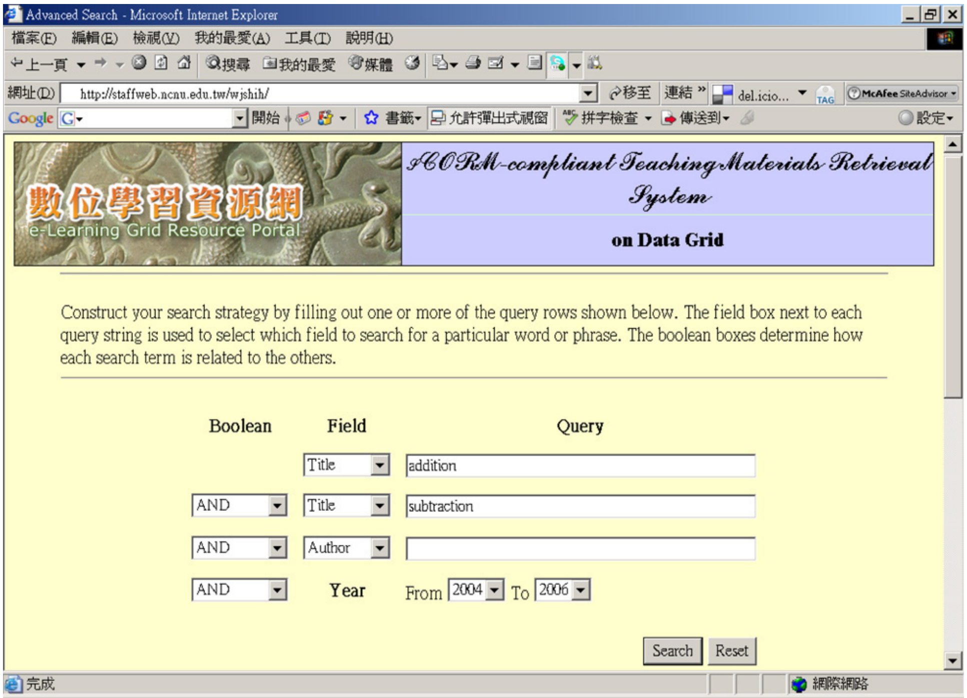
5.1. Experimental environments

We have built a data grid test-bed by the Globus Toolkit 4.0.1 [3, 19]. As shown in Fig. 4, this grid consists of five domains: Tunghai University (THU), National Taichung University (NTCU), Li-Zen High School (LZ), Hsiuping Institute of Technology (HIT) and Da-Li High School (DL). All these schools are in Taichung, Taiwan. The specifications of the nodes are listed in Table 1.

We have implemented a prototype of a content retrieval service for SCORM-compliant teaching materials. The software architecture is shown in Fig. 5. This prototype can access the data grid test-bed through a data grid portal. The user interface is shown in Fig. 6.

Table 1
Specifications of hardware resources on the test-bed.

Site	Host	CPU Type	Clock (Mhz)	RAM	NIC	Linux Kernel
THU	Delta1	Intel Pentium 4	3001	1 GB	1G	2.6.12
	Delta2	Intel Pentium 4	3001	1 GB	1G	2.6.12
	Delta3	Intel Pentium 4	3001	1 GB	1G	2.6.12
	Delta4	Intel Pentium 4	3001	1 GB	1G	2.6.12
LZ	lz01	Intel Celeron	898	256 MB	10/100	2.4.20
	lz02	Intel Celeron	898	256 MB	10/100	2.4.20
	lz03	Intel Celeron	898	384 MB	10/100	2.4.20
	lz04	Intel Celeron	898	256 MB	10/100	2.4.20
HIT	Gridhit0	Intel Pentium 4	2800	512 MB	10/100	2.6.12
	Gridhit1	Intel Pentium 4	2800	512 MB	10/100	2.6.12
	Gridhit2	Intel Pentium 4	2800	512 MB	10/100	2.6.12
	Gridhit3	Intel Pentium 4	2800	512 MB	10/100	2.6.12
NTCU	ntc01	AMD Athlon XP	1991	1 GB	1G	2.4.22
	ntc02	AMD Athlon XP	1991	1 GB	1G	2.4.22
	ntc03	AMD Athlon XP	1991	1 GB	1G	2.4.22
	ntc04	AMD Athlon XP	1991	1 GB	1G	2.4.22
DL	tc01	Intel Pentium 4	1800	128 MB	10/100	2.6.12
	tc02	Intel Pentium 4	1800	128 MB	10/100	2.6.12
	tc03	Intel Pentium 4	1800	128 MB	10/100	2.6.12
	tc04	Intel Pentium 4	1800	128 MB	10/100	2.6.12


Advanced Search - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D)
http://staffweb.ncnu.edu.tw/wjshih/
移至
連結
del.icio...
McAfee SiteAdvisor
Google
開始
書籤
允許彈出式視窗
拼字檢查
傳送到
設定
數位學習資源網
e-Learning Grid Resource Portal
完成
網際網路

SCORM-compliant Teaching Materials Retrieval System

on Data Grid

Construct your search strategy by filling out one or more of the query rows shown below. The field box next to each query string is used to select which field to search for a particular word or phrase. The boolean boxes determine how each search term is related to the others.

Boolean	Field	Query
	Title	addition
AND	Title	subtraction
AND	Author	
AND	Year	From 2004 To 2006

Fig. 6. User interface of the prototype for content retrieval.

5.2. Evaluation on precision and recall

In this experiment, we use two well-known metrics of information retrieval, precision and recall, to measure the performance of the proposed approach. We define precision and recall as follows.

$$Precision = R\_ret/Ret \tag{6}$$

$$Recall = R\_ret/R\_LOR \tag{7}$$

where

R_ret is the number of relevant documents in the retrieved documents;

Ret is the number of retrieved documents;

R_LOR is the number of all relevant documents in all repositories.

Two synthetic LORs are used in this experiment. The first LOR contains 1200,000 SCORM-compliant documents, which are converted from Web pages related to educational domains. After preprocessing, there remain 2570,623 distinct index terms. The size of this LOR is around 20 GB. The other LOR contains 2,400,000 SCORM-compliant documents, which are converted from technical papers related to computer science domains. After preprocessing, there remain 4730,384 distinct index terms. The size of this LOR is around 40 GB.

We compare the performance of a keyword-based search and the proposed method. As shown in Fig. 7, the proposed method attains better precision and recall.

We compare the performance of different similarity functions. As shown in Fig. 8, the composite similarity got the best performance. Also, the level-wise similarities have a larger impact

Table 2
Comparison of index creation time of different sites.

Sites	CPU clock (MHz)	RAM	No. of documents	Index creation time of LOR 1 (s)	Index creation time of LOR 2 (s)
LZ	898	256 MB	1200,000	205.1	421.8
DL	1800	128 MB	1200,000	239.7	483.9
HIT	2800	512 MB	2400,000	31.7	60.3
NTCU	1991	1 GB	2400,000	44.3	91.2
THU	3001	1 GB	2400,000	28.2	55.4

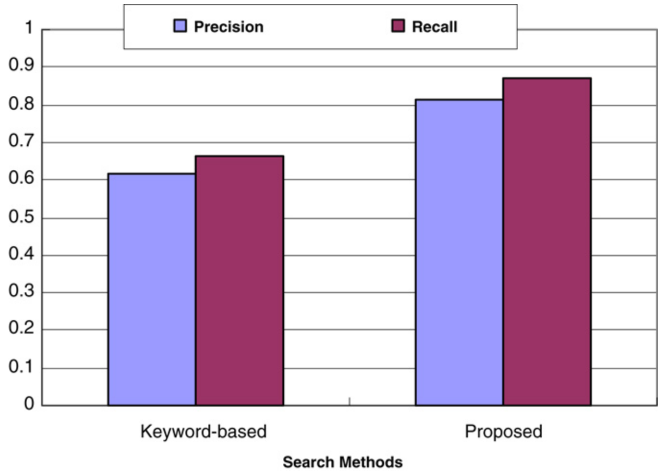


Fig. 7. Comparison with the keyword-based method.

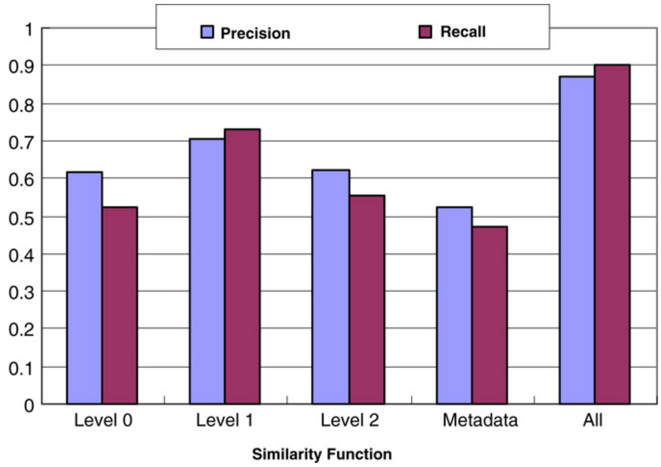


Fig. 8. Comparison of different similarity functions.

than metadata. Finally, the level_1 similarity has a larger impact than Level_0 and Level_2.

5.3. Overhead of local index creation

This experiment aims to address the cost of creating a local index. We store 1200,000 documents in LZ and DL, and 2400,000 documents in HIT, NTCU and THU, as shown in Table 2. At each site, the index creation program is executed on one local computing node. Fig. 9 shows the local index creation time for each site. We can see that the creation times in LZ and DL are significantly longer than those in the other three sites. In next section, we discuss this result.

5.4. Discussion

After the presentation of experiments, we discuss the implications behind the results. The purpose of the experimental design is to show that whether the proposed approach performs better than classical information retrieval methods. In addition, we want to know the cost of local index creation, which influences its feasibility in a real-world environment.

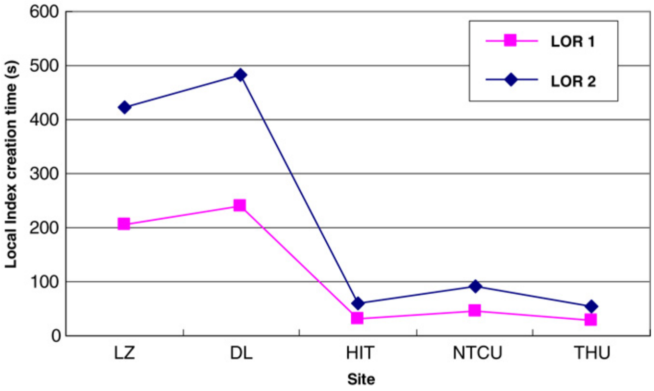


Fig. 9. local index creation time of different sites.

In Section 5.2, we see that the proposed approach is more precise than the keyword-based method. This mainly results from the proposed similarity function and ontology-based reorganization of the repository. Furthermore, we see that the level 1 similarity function attains better precision than other similarity functions. This suggests that the weighting of level 1 similarity can be increased in the proposed similarity function to get better precision.

In Section 5.3, we address the overhead brought by the proposed approach in index creation. From the experimental results, we found that the creation time is mainly decided by several factors. First, the creation time depends on the number of documents in the repository. Therefore, the creation time of LOR2 is longer than that of LOR1 for each site. Second, the creation time is also dependent on the CPU speed and main memory size. So, a site with a more powerful computer takes less time for index creation, such as HIT, NTCU and THU. Furthermore, when memory size is more than a certain level, the CPU speed becomes the dominant factor. For example, DL is slower than LZ because of the small memory size. However, when the available memory is more than 512 MB, HIT is faster than NTCU, even though the memory of the latter is greater than that of the former.

6. Conclusion

This paper describes a layered architecture to manage a large amount of learning content in data grid environments, aiming at fast and precise content retrieval. The main idea is to increase precision by an ontology-based semantic search, and to reduce search time by ontology-based indexing. The contribution of the proposed approach is two-fold. On one hand, the capability of resource sharing of data grids can be used to solve the learning content management problem on e-Learning grids. On the other hand, the information retrieval technology can enhance the resource discovery mechanism of data grids.

Content management on data grids includes many important and challenging issues. After the study of this work, future work will address replica management on grids to speed-up the content retrieval process. In addition, it is a promising way to use expert system technologies to facilitate the search process. Also, we plan to generalize this approach to include domains other than educational applications.

Acknowledgements

This research was partially supported by National Science Council of Republic of China under the number of NSC97-2511-S-468-003 and NSC96-2221-E-029-019-MY3.

References

[1] SCORM. Sharable Content Object Reference Model (SCORM). 2004 [cited 2006]; Available from: http://www.adlnet.org/.

[2] I. Foster, The grid: A new infrastructure for 21st century science, Physics Today 55 (2) (2002) 42–47.

[3] I. Foster, Kesselman C, Globus: A metacomputing infrastructure toolkit, International Journal of Supercomputer Applications and High Performance Computing 11 (2) (1997) 115–128.

[4] R.-S. Chang, J.-S. Chang, S.-Y. Lin, Job scheduling and data replication on data grids, in: Future Generation Computer Systems, Elsevier Science, 2007, pp. 846–860.

[5] R.-S. Chang, P.-H. Chen, Complete and fragmented replica selection and retrieval in data grids, in: Future Generation Computer Systems, Elsevier Science, 2007, pp. 536–546.

[6] H. Lamehamedi, B.K. Szymanski, Decentralized data management framework for data grids, Future Generation Computer Systems 23 (1) (2007) 109–115.

[7] X. Qin, Design and analysis of a load balancing strategy in data grids, Future Generation Computer Systems 23 (1) (2007) 132–137.

[8] M. Tang, et al., The impact of data replication on job scheduling performance in the data grid, in: Future Generation Computer Systems, Elsevier Science, 2006, pp. 254–268.

[9] M. Gaeta, P. Ritrovato, S. Salerno, ELeGI: The European Learning Grid Infrastructure, in: Proceedings of 3rd International LeGE-WG Workshop: GRID Infrastructure to Support Future Technology Enhanced Learning, 2003.

[10] C. Yu, et al., A methodology to retrieve text documents from multiple databases, IEEE Transactions on Knowledge and Data Engineering 14 (6) (2002) 1347–1361.

[11] W.-C. Shih, S.-S. Tseng, C.-T. Yang, Using taxonomic indexing trees to efficiently retrieve SCORM-compliant documents in e-learning grids, Journal of Educational Technology & Society 11 (2) (2008) 206–226.

[12] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, New York, 1999.

[13] A. Mittal, P.V. Krishnan, E. Altman, Content classification and context-based retrieval system for e-learning, Journal of Educational Technology & Society 9 (1) (2006) 349–358.

[14] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw & Hill, New York, 1983.

[15] I.H. Witten, A. Moffat, T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd ed., Morgan Kaufmann, San Francisco, California, 1999.

[16] Y.K. Lee, et al. Index structures for structured documents, in: The 1st ACM International Conference on Digital Libraries, 1996.

[17] B.B. Cambazoglu, C. Aykanat, Performance of query processing implementations in ranking-based text retrieval systems using inverted indices, Information Processing and Management 42 (2006) 875–898.

[18] Dewey, Dewey decimal classification, 2004 [cited 2006 Dec. 07]; Available from: http://www.oclc.org/dewey/about/default.htm.

[19] Globus, The globus project, 2004 [cited 2006]; Available from: http://www.globus.org/.

Wen-Chung Shih received the Ph.D. degree in Computer Science from National Chiao Tung University in 2008. Since 2004, he has worked as a librarian in National Chi Nan University library, Taiwan. In August 2008, he joined the faculty of the Department of Information Science and Applications at Asia University, where he is currently an assistant professor. His research interests include e-learning, ubiquitous learning, grid computing and expert systems.

Chao-Tung Yang is a professor of computer science at Tunghai University in Taiwan. He was born on November 9, 1968 in Ilan, Taiwan, R.O.C. and received a B.S. degree in computer science from Tunghai University, Taichung, Taiwan, in 1990, and the M.S. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 1992. He received the Ph.D. degree in computer science from National Chiao Tung University in July 1996. He won the 1996 Acer Dragon Award for an outstanding Ph.D. dissertation. He has worked as an associate researcher for ground operations in the ROCSAT Ground System Section (RGS) of the National Space Program Office (NSPO) in Hsinchu Science-based Industrial Park since 1996. In August 2001, he joined the faculty of the Department of Computer Science and Information Engineering at Tunghai University. He got the excellent research award by Tunghai University in 2007. In 2007 and 2008, he got the Golden Penguin Award by Industrial Development Bureau, Ministry of Economic Affairs, Taiwan. His researches have been sponsored by Taiwan agencies National Science Council (NSC), National Center for High Performance Computing (NCHC), and Ministry of Education. His present research interests are in grid and cluster computing, parallel and multi-core computing, and Web-based applications. He is both a member of the IEEE Computer Society and ACM.

Shian-Shyong Tseng received the Ph.D. degree in computer science from the National Chiao Tung University in 1984. Since August 1983, he has been on the faculty of the Department of Computer Science at National Chiao Tung University, and is currently a Professor there. From 1988 to 1991, he was the Director of the Computer Center at National Chiao Tung University. From 1991 to 1992 and 1996 to 1998, he acted as the Chairman of Department of Computer Science. From 1992 to 1996, he was the Director of the Computer Center at Ministry of Education and the Chairman of Taiwan Academic Network (TANet) management committee. From 1999 to 2003, he has participated in the National Telecommunication Project and acted as the Chairman of the Network Planning Committee, National Broadband Experimental Network (NBEN). From 2003 to 2006, he has acted as the principal investigator of the Taiwan SIP/ENUM trial project and the Chairman of the SIP/ENUM Forum Taiwan. In Dec. 1999, he founded Taiwan Network Information Center (TWNIC) and was the Chairman of the board of directors of TWNIC from 1999 to 2005. Since August 2005, he is the Dean of the College of Computer Science, Asia University. He is also the Director of the e-learning and application research center at National Chiao-Tung University. His current research interests include expert systems, data mining, computer-assisted learning, and Internet-based applications. He has published more than 100 journal papers.