Cloud-Based Retrieval Information System Using Concept for Multi-Format Data

Ibrahim Alghamdi[1], Dhafer Ali F. Alshehri[1], Abdullah Alghamdi[2], Bedine Kerim[2], Rahmat Budiarto[2]

[1]Dept of Comp. Sc., College of Comp. Sc. & I.T., Albaha University, Baljurashi Campus, Albaha, Saudi Arabia
[2]SNoopRG, College of Comp. Sc. & I.T., Albaha University, Albaha, Saudi Arabia
tooman.aa@gmail.com, kau.0000@hotmail.com, agotmi@hotmail.com, bkerim@bu.edu.sa, rahmat@bu.edu.sa

ABSTRACT

The need of effective and efficient method to retrieving non-Web-enabled and Web-enabled information entities is essential, due to the fact of inaccuracy of the existing search engines that still use traditional term-based indexing for text documents and annotation text for images, audio and video files. Previous works showed that incorporating the knowledge in the form of concepts into an information retrieval system may increase the effectiveness of the retrieving method. Unfortunately, most of the works that implemented the concept-based information retrieval system still focused on one information format. This paper proposes a multi-format (text, image, video and, audio) concept-based information retrieval method for Cloud environment. The proposed method is implemented in a laboratory-scale heterogeneous cloud environment using Eucalyptus middleware. 755 multi-format information is experimented and the performance of the proposed method is measured.

Keywords: information retrieval, concept-based, multimedia cloud computing.

1. INTRODUCTION

Size and volume of multimedia content is growing exponentially. For example, more than 30 billion pieces of content such as web links, news stories, blog posts, notes, and photo albums are shared each month on Facebook. The nature of a document may be in the format of text (DOC, TXT, PDF, etc), image (JPG, PNG, GIF, etc), audio (WAV, MP3, etc), video (AVI, MOV, etc), and so forth. The functionality for retrieving documents relevant with information seeker's needs is provided by information retrieval system (IRS).

The need of effective and efficient method to retrieving non-Web-enabled and Web-enabled information entities is essential, due to the fact of inaccuracy of the existing search engines that still use traditional term-based indexing for text documents and annotation text for images, audio and video files. Previous works showed that incorporating the knowledge in the form of concepts into an information retrieval system may increase the effectiveness of the retrieving method. Unfortunately, most of the works that implemented the concept-based information retrieval system still focused on one information format. This paper proposes a multi-format (text, image, video and, audio) concept-based information retrieval method for Cloud environment. The proposed method is implemented in a laboratory-scale heterogeneous cloud environment using Eucalyptus middleware. 755 multi-format information is experimented and the performance of the proposed method is measured.

2. RELATED WORK

With the intention of retrieving documents of various formats effectively, several attempts have been proposed to index those diverse formats, such as [1-2] for text indexing, [3] for image indexing, [4] for audio indexing, and [5] for video indexing. Each format of document was indexed accordingly using content-based indexing. Whilst a document may exist in the form of multimedia (combination of different formats), the need to integrate the way the multimedia-document being indexed is imperative.

Many conventional IRSs use the bag-of-words (BOW) representation for queries

terms in indexed documents. This approach has several shortcomings, mainly the systems cannot find relevant documents that do not mention query terms explicitly, especially when users only enter very short queries, such as commonly found in web search. These may cause many irrelevant documents retrieved by the systems and make the search results not optimal.

The shortcomings can be improved by incorporating human knowledge in IRS. Instead of using keywords or terms only in the process of indexing, concept-based IR uses distinct basic units of meaning in indexing. By using the technique, IR systems can find many other related documents although they do not contain the query term [8].The retrieved documents are conceptually related even though they do not share the query terms [9].

Multi-format concept-based IRS capitalizes the documents that come from various sources with different formats but may share similar meaning and concepts. These meaning and concepts cannot be captured by conventional IRS. Multi-format concept-based IRS indexes those different formats of documents according to the nature of each document.

Multi-format concept-based IRS integrates the multi-format indexes of documents and concept of documents. Since the documents may come from heterogeneous systems that reside in different places and generate different formats, then the data cloud is needed.

3. THE BAHA CLOUD-IRS

We propose the architecture for our multi-format concept-based information retrieval system as shown in Figure 1 shows the hierarchical view of the system implementation in cloud environments using the layered model as proposed in [14].The system comprises three components as follows:

a. Collection Manager (CM): this component is responsible for collecting documents and managing document collection intended to be indexed, searched, and retrieved by Indexer and Query Processor. The documents of all type such text, audio, image and video are stored in a repository along with their metadata. The metadata provide information about the physical and logical file names, physical and logical file locations, and other descriptive information such as authors, creation date, file type, etc. This component also provides user interface for administrator so that he/she is able to administer the document collections. When an administrator submits a new document or a document collection, CM should notify Indexer to begin the indexing process. In the case of documents retrieved from web, CM can be extended with web crawler, which is in charge of collecting web documents from many sites. CM can also be viewed as a virtual document collection to be indexed by Indexer.

b. Indexer (IX): This component is responsible to generate and maintain data structures called index that are used to provide searching capabilities. IX consumes the documents collected by Collection Manager for indexing processes. Our system design accommodates two kinds of indexing processes: term-based and concept-based indexing. Text documents can be indexed based on terms and concepts in order to facilitate searching by the term-based query with concept-based query expansion. Documents of other type (audio, image, and video) are only indexed using concepts-based indexing. These documents need to be annotated first before they can be indexed. The process Annotation should be done automatically based on training data set using algorithms such as proposed in [10-13]. Domain knowledge base is used to associate concepts to text. The concepts are then used as index.

c. Query Processor (QP): this component is responsible for managing query and search result sets. QP provides user interface for concept-based query. QP makes use of the same domain knowledge base with the one used in concept-based indexing process.

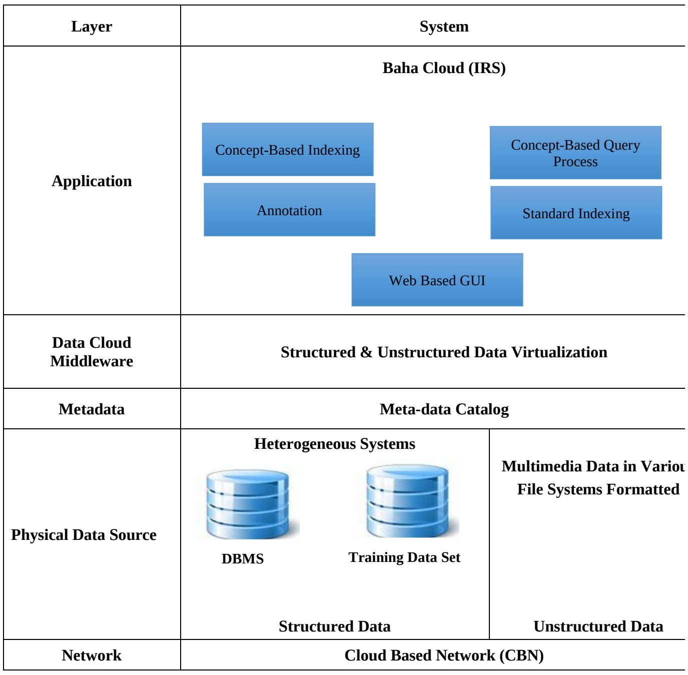

FIGURE 1. The proposed Baha Cloud-IRS architecture

The fundamental layer of the model is Data Layer. This layer represents all types of data: unstructured and structured data. The unstructured data means here are the collection of documents with all media types. These data are stored in heterogeneous file systems and storage system. They can be treated as a string of bits and usually managed by the underlying operating systems. The structured data include term-based index, concept-based index, domain knowledge base and training dataset.

These data have explicit or implicit schema that govern the structure of the data. Usually, data of this type are managed by special program called database management system (DBMS).

The second layer is Data Cloud Virtualization. This layer is built on the heterogeneity of file and storage systems used for storing the unstructured data and the heterogeneity of database management systems that manage the structured data. Based on the analysis of the file-oriented data grids, the sub layer Unstructured Data Virtualization provides the basic services [15]: Data Storage and Replication service, Composition and Relation service, Search service, and Metadata Storage service.

One object can be described in many metadata records. The metadata also records information associated with replication. These records can be utilized for the search service. Usually, database systems are used to store and manage the metadata. Furthermore, some database systems containing metadata can be integrated using structured data virtualization components of the data cloud middleware.

The sub layer Structured Data Virtualization provides the four basic services: data access, transformation, integration and delivery. This layer enables us to access structured data stored in distributed heterogeneous data resources such as relational and XML database, to transform data in one schema to another schema, to provide an integrated view of multiple databases, and to deliver data using many popular protocols, such as Web service, email, HTTP and FTP.

At the application layer, data-intensive modules used in the system can utilize the two sub layers of Data Cloud Middleware in order to do some processes needed in information retrieval. The modules comprise term-based indexing and matching process, concept-based query expansion, concept-based indexing and matching process, and auto-annotation process for non-text documents.

Figure 2 shows an implementation of the structure that employs the two data clouds. We call the system prototype as BahaCloud-IRS system. In the design, documents of various formats are stored in four storage servers. One server functions as metadata catalog server which stores the metadata of all documents managed in the system. Each storage server also functions as indexer, which indexes each type of documents. Server A, B, C, and D focus on indexing text, image, video, and audio-typed documents respectively. Although all documents can be managed by different servers which may use various platforms, the difference is abstracted by unstructured data virtualization as such that one indexer server of a type of document can process a document with the same type stored in other server. For example, server A is responsible for indexing all text documents. It can do the indexing for all text documents, which are stored not only in its storage, but also in the storage media of other servers. It knows all text documents stored in other server using metadata information provided by the metadata catalog server. All index collection in every indexer servers are published as Data Resources, which can be accessed using web services. A server is dedicated to integrate all of the index collections.

This server has several Remote Data Resources that function as pointers to each Data Resource published by the indexer servers including the Data Resource of Metadata Catalog. The server also includes some activities that represent the operations regarding to the index collection and retrieval tasks, such as looking up query terms from the collection, merging the search result sets of every indexer servers, sending the merged search results to users, etc. The lab-scale cloud computing environment used in this work is illustrated in Figure 3

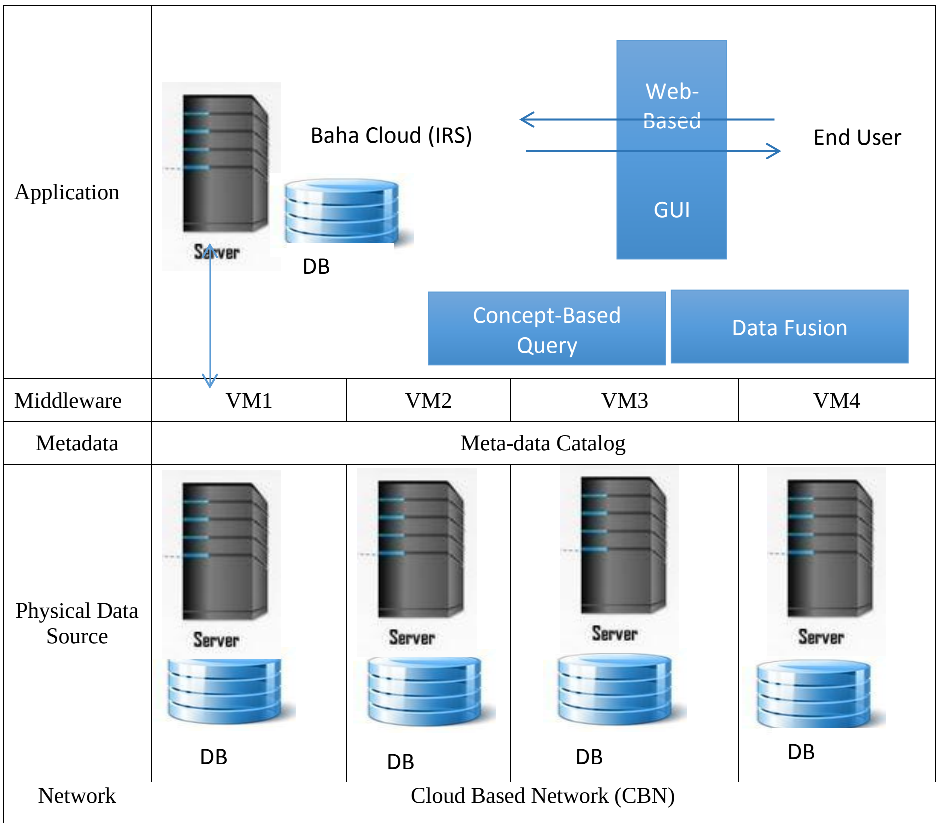

FIGURE 2. Hierarchical view of the structure of the proposed method BahaCloud-IRS

.

BahaCloud-IRS is developed using JDK 1.6 with Apache Axis 1.4. The PostgreSQL v8.3.9 as the backend database for ICAT Server and Apache Tomcat v5.5 as the Java Web container. Five computers with the minimum specifications: CORE i-7, RAM 4 GB, and HDD Storage 500 GB. Each computer used different operating system in order to simulate heterogeneous environments: Ubuntu 14.4, CentOS 2 and Windows 8.

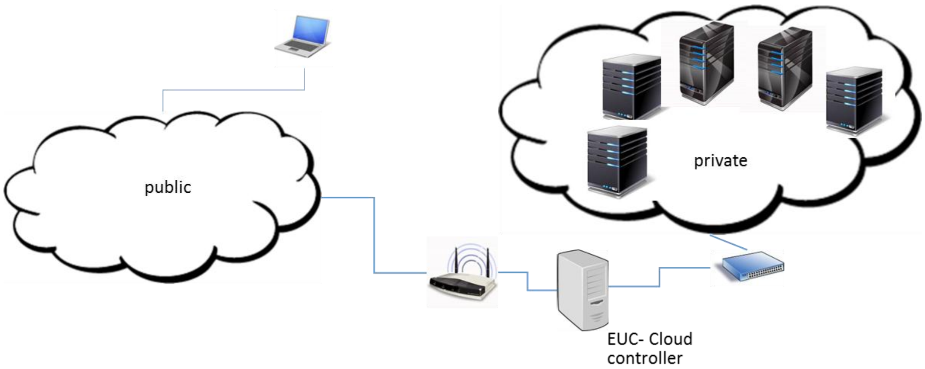

FIGURE 3. The Baha cloud environment

4. EXPEREMINTAL RESULTS AND ANALSYSIS

The dataset creation is limited into one topic, namely “Global Security Issues”. All of the following collections are gathered from several Web sites and filtered to be relevant to the selected topic. All text files were gathered using Google search engine with the keyword “global security issues”.

Image data set was collected using the similar method as text data set. We searches image files using Google, Yahoo, and Bing search engines with the keywords “global security issues” and downloaded 615 image files of various formats: JPG, PNG, and GIF. All image files were annotated manually using the textual annotation provided by the search engines. The annotations were usually limited in the number of words.

The data set used for audio retrieval were downloaded from some web sites. The audio files were recordings of lectures and radio broadcasts. Their contents are related to the determined topic “global security issues”. The descriptive metadata for each file was usually available although some of them were very short sentences. 85 audio files with two formats: MP3 and WAV are collected.

The sample data set for video were downloaded from YouTube. We choose some video of the same fields as the audio files in order to represent information entities with similar or related concepts. The available description of video will be used for indexing. We downloaded 55 video files along with the textual description of the video from YouTube using the search feature available in the Web site for the same keywords “global security issues” with the same FLV file type.

All files were managed by Coordinator server and distributed in some physical resources, namely repository1 to repository5.The system presents the retrieved documents of all types at once. At the first page, it selects one until three top documents for every type to be presented to users. Then, the user can freely view, play or download the documents from Data Cloud.

We measure the system performance based on the time needed for conducting certain operations. Figure 4 shows comparison of the time needed for uploading files via BahaCloud-IRS and directly to coordinator server. Uploading files via BahaCloud-IRS needs at least two steps to complete the process. The first is the process of posting the files from the user to the BahaCloud-IRS web application. The second is forwarding the bytes received from the user to the actual repository in data repository zone. It is shown that there is a significant delay caused by the additional layer (BahaCloud-IRS web application). Uploading large files via BahaCloud-IRS needs almost two-fold time compared with uploading directly to the coordinator server.

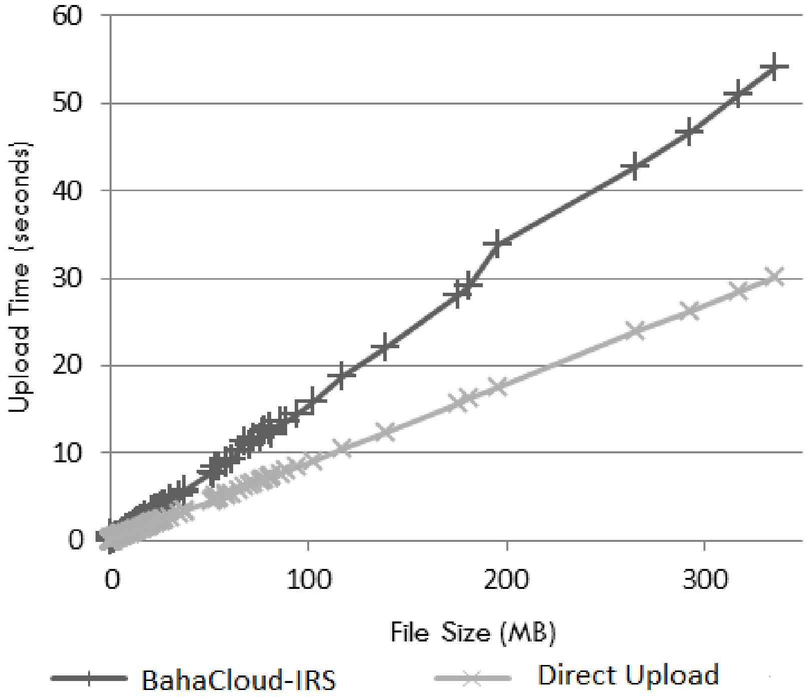

FIGURE 4. File upload time

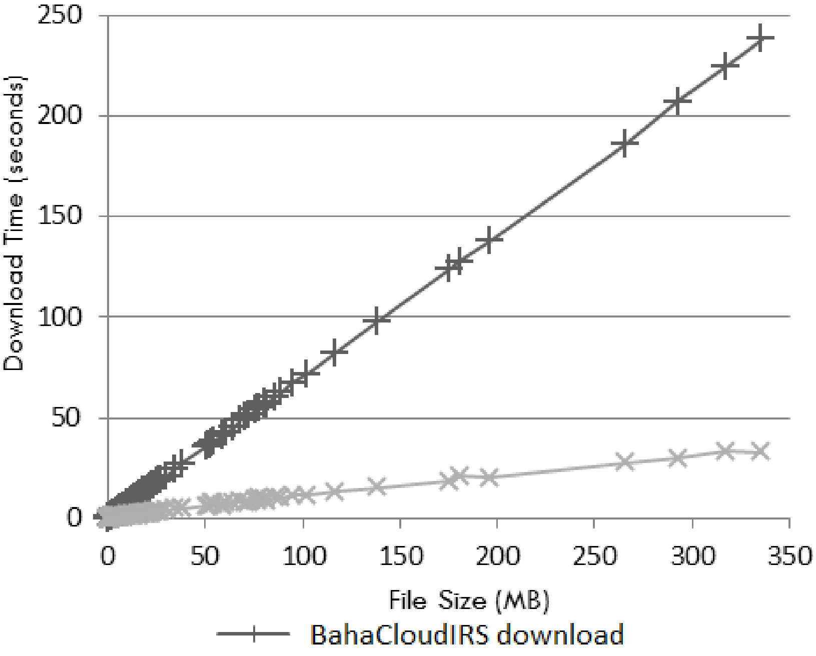

FIGURE 5. File download time

The next performance measurement is comparison of the time needed for downloading files via BahaCloud-IRS and directly from coordinator server as shown in Figure 5. In the first case, a user initiates to download a file via BahaCloud-IRS web application. The application then forwards the request to coordinator server. The server finds the location of requested file from coordinator Metadata Catalog server, retrieve the file from the physical resource and then send the file to BahaCloud-IRS server. The server forwards the file to the requested file.

In the second case, downloading files is conducted directly from coordinator server. Similar with uploading files via BahaCloud-IRS), it is shown that there is a significant delay caused by middleware layer (that is BahaCloud-IRS web application). Downloading large files needs almost five-fold time compared with downloading directly from the coordinator server.

As regards to query time measurement, a query “global security issues” is submitted into the system and measured the time needed to retrieve various numbers of documents starting from 1 until 300 documents. Figure 6 shows the measurement result of query time over various numbers of returned documents. Generally, the time needed for processing the query is proportional to the number of returned documents.

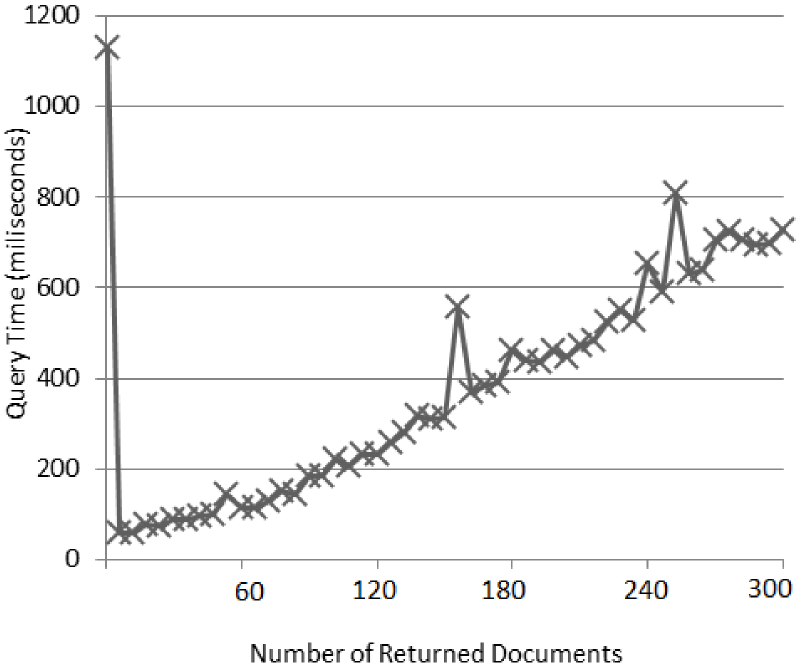

FIGURE 6. Query time

5. CONCLUSION

An architecture of multi-format concept-based information retrieval system for cloud environment is intruduced in this paper. The multi-format concept-based information retrieval system promises seamlessly unified document retrieval in the heterogeneous cloud environments.

As a future work, an automatic annotation system with well training capability to provide accuracy is considered. Another potential future work is a possibility to search documents of many media types using non-text query, such as usually found in Content-based Multimedia Information Retrieval (MIR). It enables the feature of multi-channel access, which can deliver the retrieval task provided by the proposed architecture to clients of many query and access types. The query is not only limited to the form of text in natural language, but also can be extended to another type of media. This feature needs indexing low-level features of multimedia and integrating with high level features (semantic) to solve the problem known in Content-Based MIR.

ACKNOWLEDGEMENTS

The authors thank to Albaha University for partially funding under research contract no. 95/1436.

REFERENCES

[1] V.N Anh and A. Moffat, Improved Word-Aligned Binary Compression for Text Indexing, IEEE Transactions on Knowledge and Data Engineering, 2006; 18(6): 857-861.

[2] K. Norvag and A.O. Nybo, DyST: Dynamic and Scalable Temporal Text Indexing, Proceedings of 30th International Symposium on Temporal Representation and Reasoning 2006, Budapest, 2006: 204-211.

[3] J. Huang, S.R. Kumar, W.J. Zhu, M. Mitra and R. Zabin, Image Indexing using Color Correlograms, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, 1997: 762-768.

[4] K. Thambiratnam and S. Sridhara, Rapid Yet Accurate Speech Indexing using Dynamic Match Lattice Spotting, IEEE Transactions on Audio, Speech, and Language Processing, 2007; 15(1): 346-357.

[5] J. Wei, S.M. Bhandarkar and K. Li, Semantics-based Video Indexing using a Stochastic Modeling Approach, Proceedings of International Conference on Image Processing 2007 (ICIP 2007), San Antonio, 2007: 313-316.

[6] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, 1st Ed. New York, USA: Cambridge University Press, 2008.

[7] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. New York: Addison-Wesley, 1989.

[8] Gabrilovich E and S. Markovitch, Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proceedings of 7th International Joint Conference on Artificial Intelligence, Hyderabad, India, 2007: 1601-1611.

[9] Z.A. Hasibuan, Multi-dimension Concept-based Information Retrieval System, Proceedings of ALL/ACH 2000 Conference, Glasgow, UK, 2000:

[10] P. Huang, J. Bu, C. Chen, K. Liu, and G. Qiu, Improve Web Image Retrieval by Refining Image Annotations, in Information Retrieval Technology. Berlin: Springer Berlin Heidelberg, 2008; 4993: 403-435.

[11] R. Shi, C.H. Lee, and T.S. Chua, Enhancing Image Annotation by Integrating Concept Ontology and Text-based Bayesian Learning Model, Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 2007: 341-344.

[12] G.J. Qi et al., Correlative Multi-Label Video Annotation, Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 2007: 17-26.

[13] Y. Song, W. Wang, and A. Zhang, Automatic Annotation and Retrieval of Images, World Wide Web, 2003; 6(2): 209-231.

[14] M. Kosiedowski, C. Mazurek, M. Stroiński, M. Werla and M. Wolski, Federating Digital Library Services for Advanced Applications in Science and Education, Computational Methods in Science and Technology, 2007; 13(2): 101-112.

[15] J. An, The Demonstration of Cloud Retrieval System Model, Journal of Software, 2011; 6(2): 249-256.