An MPEG-7 Compatible Video Retrieval System with Support for Semantic Queries

Quan Zheng
Department of Automation
University of Science and Technology of China
Hefei，China
qzheng@ustc.edu.cn

Zhiwei Zhou
Department of Automation
University of Science and Technology of China
Hefei，China
jordan@mail.ustc.edu.cn

***Abstract*—In this paper, we design and implement a distributed video retrieval system with support for semantic queries named XUNET. It firstly uses a self-made video semantic processing tool, which is based on a video analysis tool and added with semantic graph annotation and natural language processing functions, to parse videos and generate corresponding MPEG-7 description files. Subsequently, it establishes distributed index of the MPEG-7 files and distributed storage of video files separately. The system provides numerous web query interfaces, including keywords semantic expansion query, semantic graph query and natural language query, when clients' query intentions have been submitted, specific interested videos or segments will be quickly exhibited through web browsing or VOD as query results. Because of adopting a distributed architecture, XUNET is highly scalable and supports mass video information index and retrieval.**

Keywords-MPEG-7; Distributed System; Semantic; Lucene

I. INTRODUCTION

With the development of computation, digital device, multimedia, database, web technology and so on, large amounts of video information are produced quickly. Therefore, faced with large amounts of video information, how to classify, organize and index in order to achieve fast retrieval of video information, it is an urgent problem. At present, traditional video retrieval technologies could be divided into two ways: text-based queries and content-based queries[1]. Text-based video retrieval technology is widely used on Internet applications. Large amounts of video on the video sharing sites are generally embedded into web pages which also includes descriptive information about the video, such information allows general-purpose search engine to get video information with the help of web crawlers. However, video retrieval based on meta-information relies on manual editing of text annotation. How to fully understand the video content information is complex, and search results have certain limitations. Content and image-based video retrieval technology mainly uses computer graphics, vision and other areas, through video content analysis and processing, the video will be divided into two levels of the scene and the lens fragments, After extracting color, texture , shape, motion and other low-level features in video segments, video retrieval could be finally achieved by matching characteristics. Representative systems in this area are: Fischlar system of Dublin City University [2] and Multimedia Search and Retrieval System of IBM[3][4]. Problems of such technologies are too complex when we compute the intensive index, slow retrieval speed and low retrieval accuracy.

At present, new research field of video retrieval is based on semantic information, It focuses on relations between words and understanding of the whole sentence: How to make computers automatically extract video semantic information with high accuracy, establish mass video semantic index, and achieve the objective of semantic video retrieval processing, these problems are the hot topics of the current video retrieval technologies. There are lots of works to be done to focus on the automatic mapping from low-level features of videos to semantic description of videos, which attempts to solve the semantic gap problem, the representative samples in this area include: Digital Library Project joint funded by US.NSF,ASF and NASA, CueVideo Project of IBM[5], Informedia Digital Video Library of Carnegie - Mellon University[6], MediaMill of Intelligent Systems Lab of University of Amsterdam, The Netherlands[7] and so on, Nowadays, such technologies of computer automatically understanding, extracting video objects and their relations, and setting up video index are still immature. Video semantic object recognition accuracy rate is only 60%, far less than the degree of commercial applications.

Currently a developing direction of video retrieval technology is to combine low-level video content features with advanced features to establish a joint index, in order to provide users with new search tool, a representative work includes: BilVideo-7.

Presently without a breakthrough in video machine understanding, the most practical video indexing in commerce is still based on manual annotation, and setting up the video index using text indexing techniques of artificial mark on the video. Through this approach, there are many research institutions that provide automatic segmentation and labeling of video editing tools, such as Pepsky Video Splitter V5.2 and Semantic Video Annotation Suite Which developed by JOANNEUM RESEARCH Company of Austria.

For the standard video description of low-level features and advanced features, MPEG-7 is formally called "Multimedia Content Description Interface", it is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) in 1998,

The National High Technology Research and Development Program ("863" Program) of China（2008AA01Z147); Natural Science Research Project in Anhui Province's Universities（The research on the key technologies of building video semantic index based on the analysis of annotating）

978-1-61284-459-6/11/$26.00 ©2011 IEEE

MPEG-7 standard (Multimedia Content Description Interface) can use the structured text to describe the low-level features and high-level semantic features of video because MPEG-7 file is a XML file. So based on MPEG-7 standard, extracting and analyzing MPEG-7 semantic description information of the video file to establish an effective semantic index for massive video, is not only meaningful to solve the automatic computer understanding for the semantic video, and to establish an effective video index after the breakthrough on indexing problem, but also commercially valuable for solving current video indexing technology based on artificial tagging.

However, how to represent video semantic information and build up an effective video index using the framework described in the MPEG-7 standard is still a difficult problem. In order to achieve the semantic characterization of video, it is proposed to make use of semantic entities and semantic relationships between entities to describe the semantic information of the video shot, which includes semantic entities that like the semantic objects in MPEG-7 standard, such as characters, events (behavior), things, time and place. Semantic relations are defined by MPEG-7 semantic relations, such as location, source, and destination. Users can annotate the video shot in MPEG-7 file by visual semantic graphic annotation tool. Demonstrate an event in a video shot by a directed semantic graph. Then following the path length of 0, 1, 2 etc. in the semantic graph to characterize semantic graph information, and converting them into strings to build index based on keywords. Similarly, semantic graph queries submitted by users follow the same procedure, and Users retrieve keywords.

As the labor of semantic annotation of video shot is too huge, we propose using the existing information in MPEG-7 description files to analyze and process the natural language. Then analyzing the grammatical structure of the video descriptions and extracting the relationship between the semantic entities, in order to automatically match semantic entities and their relationships to entities information and relationship of MPEG-7 standard. Finally, generating semantic graph and converting into keywords to build inverted index. This can greatly reduce the labor of manual annotation, and realize automatic extraction of semantic information and establishment of indexing. Similarly, the natural sentence query submitted by users can also be converted into the same keywords set to search.

In this paper, the section 2 describes the related work, the section 3 describes the architecture of XUNET system, the section 4 introduces the design and implementation of subsystems, and the section 5 give the final conclusion.

II. Related Work

A. MPEG-7 standard

MPEG-7 is formally called “Multimedia Content Description Interface”, it is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) in 1998, and it is designed to describe the content of multimedia, which contains low-level and high-level features[8] .Those files described by MPEG-7 are kinds of information based on XML, and theses XML files can be generated by some analysis software.

MPEG-7 mainly contains two elements:

1) MPEG-7 offers a comprehensive set of audiovisual description tools in the form of Descriptors (D) and Description Schemes (DS) that describe the multimedia data, forming a common basis for applications. Descriptors are used to define and express the specific characteristics of a particular aspect of syntax or grammar. The common descriptions associated with the video are divided into color, texture, shape and motion, each category contains several kinds of descriptors, so as to describe the visual information in different structures and ways. Description scheme is described by one or more of the D and DS form, and DS provides the relationship between their structure and syntax.

2) DDL (Description Definition Language) is a language which is used to specify a description of the description scheme. It is a pattern language, and it can be a characterization which is the result of modeling the audio and video data. The DDL provides a MPEG-7 description tool, including descriptors and description schemes. It also provides some rules building descriptors to the description scheme.

The purpose of describing the semantic content of the video in the MPEG-7 description file can be achieved by using annotation tools which contains text marked by free text, keywords, structured text, and the relationship between semantic entities, and the semantic entities generally include objects, events, concept, location and time, then semantic relationships include the agent, agentOf, patient, patientOf, similar, opposite, user, userOf, location, locationOf, time, timeOf[9], which are MPEG-7 standards[9]. In addition, in the MPEG-7 description file, the video descriptions can be divided into many shots by time, and the shot is divided into a number of key frames. In this system, some free descriptors, semantic graph descriptors and semantic text descriptors are added to each shot in the MPEG-7 description file.

B. Lucene

Lucene is a full-text search engine tool which is open source, it was developed originally by Doug Cutting, but now lucene is a subproject of the Apache Software Foundation project team. Lucene was developed by java, and it is a fully object-oriented framework for full-text search engines, and it can be easily embedded in the target program in order to achieve full-text search functionality, at the same time, developers can also extend Lucene to customize their search function easily. The system can include nodes of Indexer, Searcher, Manager and Collector. The purpose of Indexer is to set up the index, Searcher can search the index, Manage can manage the whole information, and Collector can collect the search results.

In our system, based on Lucene, by saving and updating the real-time global document frequency df and the overall maximum number of documents maxDoc in Manager which is the central control node, besides each searcher server updates its document frequency df and the maximum number of documents for the real-time global data maxDoc before querying through the synchronization of Manager, which guarantees the rating points coming from a number of Searchers are comparable, Collector can get the result by merging and sorting different searchers' results. Therefore, the modified Lucene search framework can be adapt to the distributed search architecture.

C. LTP and semantic knowledge library

Language Technology Platform (LTP) is a language processing system framework from Harbin Institute of Technology. It defines the XML-based text that provides a set of bottom-up language processing module, provides the results of the visualization tools, and shares library dependency tree, The synonym expanded version of the word forest resources. LTP has used lexical, semantic, syntactic and other six Chinese processing functions.

Natural language processing system eventually needs the support of the powerful knowledge library, first we must know what knowledge is and what knowledge the computer can understand is, Knowledge is a system and the relationship between various concepts or between the attributes of concepts[10], we can describe the general concepts and generate the relationship between concepts. At present, the most important knowledge libraries are as follows: HowNet and Chinese WordNet developed by Southeast University.

D. Calculating the word Similarity Based on How-Net

When compared to traditional semantic dictionary, How-net does not match each concept with a node in hierarchy of the concept tree, but proposes using a knowledge description language to describe a concept, which organized the hypernym and hyponym relationship into the tree hierarchy. Therefore, we can calculate the similarity between two semantic expressions which use the knowledge description language[10].

- Calculating Words Similarity

For the two Chinese words W1 and W2, if W1 has n meanings or concepts: S11,S12,.....,S1n; W2 has m meanings or concepts: S21,S22,...,S2m; we can define the maximum of all the concept similarities as the similarity of W1 and W2. so we can convert the similarity of words into the similarity of concepts[10].

$$Sim(W_1, W_2) = \max_{i=1\cdots n, j=1\cdots m} Sim(W_{1i}, W_{2j}) \quad (1)$$

- Calculating Sememe Similarity

We can use the sememe to describe all the concepts, so calculating the sememe similarity is the basis of computing the concepts similarity, all the sememes construct the tree system based on the hypernym and hyponym relationship, in the tree, we can view the length of the path as the actual distance between two sememes. So we can calculate the semantic distance similarity. Suppose the distance of the path between two sememes in the tree system is Dis(P1, P2), we can calculate similarity as follows:

1) When the two sememes are in the same tree, the length of path equals to the sememe distance. According to the following formula, we can get the semantic similarity between the two sememes. Where p1 and p2 can represent the two sememes, Dis(P1, P2) is the length of path between p1 and p2 and a positive integer, and α is an adjustable parameter.

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad (2)$$

2) In the second case, the two sememes are in different sememes trees, we can define the length as the unlimited constant, That is to say, the semantic similarity is 0.

III. XUNET SYSTEM DESCRIPTION

Fig. 1 shows that XUNET is a video retrieval system based on MPEG-7 descriptions; by analyzing the MPEG-7 descriptions of the content of the videos, XUNET indexes them. XUNET supports not only retrieving video and clips using keywords and their semantic expansion, but also the retrieval based on semantic map and the semantics of natural language. What's more, we can quickly locate the related video segments in massive videos with the rapid retrieval responsive time and the precise results through this system.

A. System Architecture

XuNet video retrieval system consists of five subsystems just like the description of Fig. 1. These five subsystems are video pre-processing subsystem, video semantic processing subsystem, video distributed retrieval subsystem, video distributed storage subsystem and the global information management subsystem. System provides three functions: distributed indexing of video descriptions, distributed video storage, distributed retrieval and VOD[11].

- Distributed indexing of video descriptions

XUNET decompresses the compressed packages which consist of MPEG-7 descriptions, key frames and they are annotated by the administrator based on semantic information. After that, it parses the XML documents of MPEG-7 descriptions under the control of global information management subsystem; then stores the description information of videos and clips in global information management subsystem and InfoDB database; then submits the free text annotation information of description

information to video semantic processing subsystem and stores

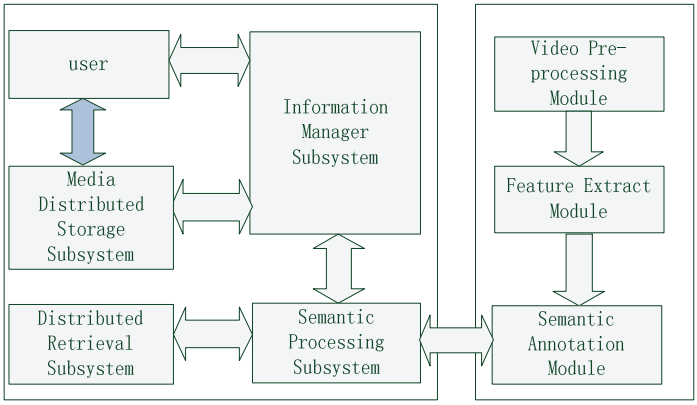

Figure 1. System Architecture

the video semantic information which is gotten from last procedure in database; all the description information of videos and video clips will finally be transmitted to video distributed retrieval subsystem for distributed indexing.

- Video distributed storage

After the format conversion, video cutting, semantic analysis of video pre-processing subsystem and video semantic processing subsystem, the video (.Mpg and other formats) can be submitted and deployed to video distributed storage subsystem through ftp.

- Distributed retrieval and VOD

The semantic query is first submitted by users through web interface, and then submitted to video semantic processing subsystem by XUNET. After the processing, we can get a set of keywords that contains the semantic information, when users submit it to video distributed retrieval subsystem to retrieve; the search results could be returned to web application servers and showed on the website. Through these procedures, users can get videos and clips eventually.

B. User Interface

The user interface in XUNET video retrieval system includes three parts: management interface, query interface and result interface.

- Management Interface

Management interface includes the interface to submit video MPEG-7 description, key frame and original video and the interface to view, delete and update the video description in database.

- Query Interface

We provide three kinds of query interfaces: keywords search interface, semantic graph search interface and natural language search interface. We have designed simple input frames for keywords and natural language search and a visual query interface for semantic graph search. Each query interface includes the following two functions: a. synonym extension and hypernym and hyponym extension; b. retrieving the video or video segment. Moreover, there are some other advanced search functions.

- Result Interface

Result interface consists of the interface to show key frames and the interface to show video descriptions. We highlight the keywords in the video descriptions to help users browse the search results more comfortable. In addition, users can play the videos and video segments listed in the result.

IV. THE DESIGN AND IMPLEMENTATION OF ALL THE SUBSYSTEMS

A. Video Pre-processing Subsystem

Fig. 2 shows that the purpose of video pre-processing subsystem is to analyze the video so as to get its descriptive information, To begin with, in pre-processing module a video is converted to the format of MPEG-1 and cut into segments. Extract the feature of MPEG-1 video files and get the MPEG-7 video description and key frame, we can analyze the key information in the MPEG-7 description which includes low level features just like the beginning time, the ending time and the length of shot. In addition, by the semantic annotation module we can manually and semi-automatically label semantic annotation to shots in the MPEG-7 video description. Finally the semantic information in the MPEG-7 description is analyzed and stored into database in the information management subsystem with key frame.

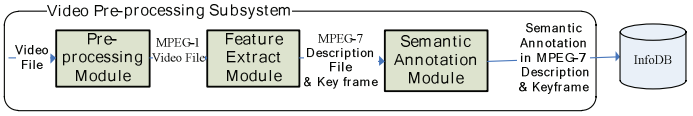

Figure 2. Video Pre-processing Subsystem Architecture

Manual annotation includes text annotation and semantic graph annotation in video shots. a. use text to describe the story in the shots. b. use semantic entities such as people, events, objects, place and time and semantic relationship between entities to describe semantic information in shots. We can represent entity as a node and use vectors to express the relationship between entities. All these procedures could be realized by the annotation tool our team exploited. Then the semantic information of shots can be described by the vector graph whose length scales from 0,1,2 to N. The graph will then be converted into strings which could be finally taken into the sub-tag in the MPEG-7 shots semantic graph to describe the semantic graph.

Semi-automatic annotation utilizes the movie script with timestamp to annotate, the scene description and dialogue in accordance with the time of MPEG-7 shots in the sub-tag text. The process can be done automatically.

B. Information Management Subsystem

Information management subsystem includes Web server, Overall information manager and InfoDB database. Web server plays the key role in receiving the requests from users and administrators, and returning the results: a. deliver the query of searching and playing video from users into retrieval subsystem and MDN subsystem, Finally return the search results and play the video; b. the query from administrators can be carried to Overall Information Manager and return the results.

Overall Information Manager's functions are as follows: a. analyze the MPEG-7 description submitted by the administrator, then save the video description and key frame to InfoDB; b. receive the query from administrators and operate the video or video segments in InfoDB.

- InfoDB

Before adding the videos into the retrieval system, we should partition segments based on scene, extract the key frame and the feature descriptor information, meanwhile, we can add a manual annotation and convert these annotated key frames into MPEG-7 description files.

InfoDB is responsible for analyzing and storing these MPEG-7 description files and the key frames, these information is the structured data and stored in a relational database (such as MySQL), Currently the information is mainly stored in two tables: video information table and video segment information table. The video information table that views the video as a unit is capable for preserving the title, length, published time, content and IsIndexed, Table 1 shows video information table structure.

Video segment information table saves the video segment information such as Video ID, Offset, Length, content, Keyframe, EdgeHistogram and scalableColor. Table 2 shows video segment information table structure.

InfoDB has its own management interface, administrators can view and modify the video index state. When the administrator chooses to add a video index, InfoDB can convert the VideoID to the task object and send it to the Manager node, and generate the index.

C. Retrieval Subsystem

Be a fundamental service system in XUNET video retrieval system, the retrieval subsystem illustrated in Fig. 3 provides the following functions: building inverted index of video description information and searching for keywords.

- Video Indexing Process

Before providing users with the retrieval function, the retrieval subsystem has to index video and video segment description. Administrator can select the compacted package consisted of video XML description file and key frame and click the "index" button to make the system execute adding index process. Firstly, the Overall Information Manager announces the

TABLE 1. VIDEO INFORMATION TABLE STRUCTURE

Name	Data Type
VideoID	String
Title	String
Length	Time
Publish Time	Date
Content	String
IsIndexed	Boolean
……	……

TABLE2. VIDEO SEGMENT INFORMATION TABLE STRUCTURE

Name	Data Type
SegmentID	String
VideoID	String
Offset	Int
Length	Time
Content	String
KeyFrame	BLOB
EdgeHistogram	String
ScalableColor	String
……	……

Manager to index the video and video segment, Then Indexer acquires the MPEG-7 video description from Information Management subsystem, and gets the inverted index, finally Manager can deploy index on Searchers considering load balance.

- Video Retrieval Process

XUNET submits the keywords search queries to video retrieval subsystem Collector, Collector can query all of the Searcher nodes, merge the results and return to WEB server, display the results to users in WEB server.

D. Semantic Processing Subsystem

Fig. 4 shows that semantic processing subsystem has two main functions: the first one is to analyze and extend the semantic queries from users' queries and convert into keywords strings. Users can retrieve in retrieval subsystem, the other is to analyze and extend the semantic information from text annotation information in MPEG-7 video description files in Information Management Subsystem, convert into keywords strings and store in the database. Subsequently index the semantic keywords in retrieval Subsystem.

- Keywords Semantic Extended Query

The semantic extended module in the semantic processing subsystem includes two main functions: the one is a synonym expansion for searching keywords, the other is semantic

hypernym and hyponym expansion for query keywords.

The system can process the query keywords as follows: get the corresponding synonym based on synonym forest and calculate the similarity between the query keywords and synonyms. And the system analyzes the extended keywords to acquire the abstract expansion in the upper layer and the specific extension in the low layer based on WordNet. Finally the system can submit query keywords of synonym expansion and semantic expansion to the retrieval subsystem.

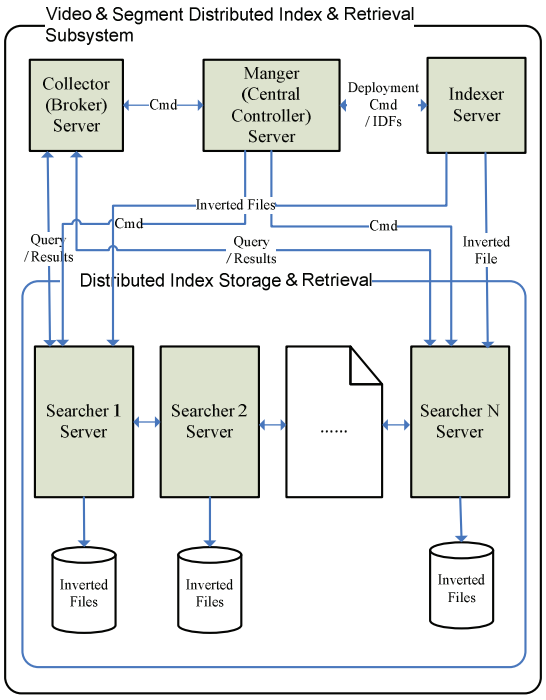

Figure3. Retrieval Subsystem Architecture

- Semantic Graph Query

Overall Information Manager Subsystem can send the semantic annotation information of video segments stored in InfoDB retrieval subsystem to establish the inverted index. Users can draw the semantic graph to describe the video and convert the semantic graph into keywords strings. Users can submit these keywords set to retrieval subsystem after synonym expansion, semantic hypernym and hyponym expansion.

- Natural Language Query

Fig. 5 shows that the functions of natural language processing module are semantic analysis and processing the simple modern Chinese sentences, extracting the semantic components and expressing the semantic content in a standardized form of semantic graph.

Because analyzing the modern Chinese sentences' structures is very complex. It is very hard for us to form the uniform standard for all the Chinese sentence structures. The research on the structure of the modern Chinese sentence are not sufficient, In XUNET system, we analyze and process parts of Chinese sentences. First declarative sentences can be analyzed, the examples of sentence structures are as follows: the structure of "subject + verb", "subject+ adjective", "subject+verb+noun1+ noun2", "subject + verb + noun+ verb+ noun" and serial verb construction.

Overall Information Manager Subsystem can first send the text annotation information stored in InfoDB in Information Management Subsystem to natural language processing module to analyze and process. We can get the information of semantic graph and convert into strings, finally after the synonym expansion and the hypernym and hyponym expansion we can get the keywords set and save into video segments in InfoDB in Information Management Subsystem.

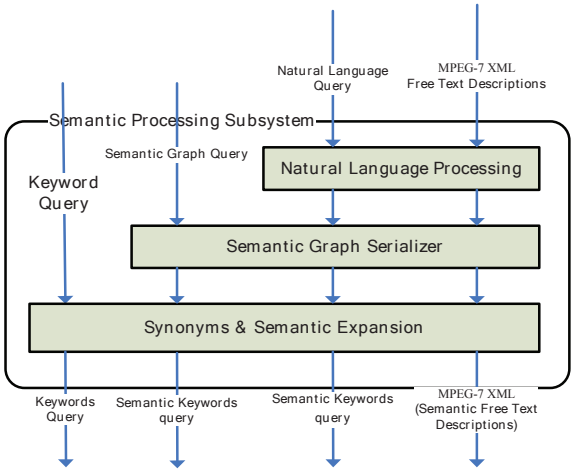

Figure 4.Semantic processing subsystem Architecture

Before the natural sentence query, Overall Information manager can first send semantic information of video segments in InfoDB by the natural sentence analysis into retrieval subsystem for generating the inverted index. Users can input the natural language query from the WEB page. And WEB Server can convert into the keywords set by processing natural sentence module and semantic graph strings module. We submit keywords to retrieval subsystem by the synonym expansion and hypernym and hyponym expansion.

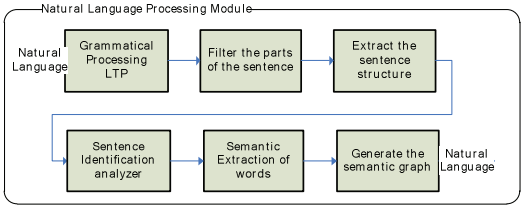

Figure 5. Processing Natural Language

- MDN Subsystem

Fig. 6 shows that MDN (Media Distributed Network) subsystem includes MDN storage subsystem, the external streaming server clusters and media content providers. And MDN storage subsystem includes the management module MM, the directory server DS, the distributed nodes PNs, the whole subsystem mainly provides with the distributed storage and the function of playing video.

V. CONCLUSION

The paper presented the architecture for an MPEG-7 compatible video retrieval system named XUNET supporting semantic queries. The system can analyze the semantic information of MPEG-7 description, build the text index and retrieve the video content effectively and efficiently. The information based on the MPEG-7 standard can describe the video content. XUNET can integrate many methods to retrieve the video and improve the retrieval accuracy. Take ever-increasing information overload into account, XUNET can design the distributed architecture, good scalability and retrieve the massive video information.

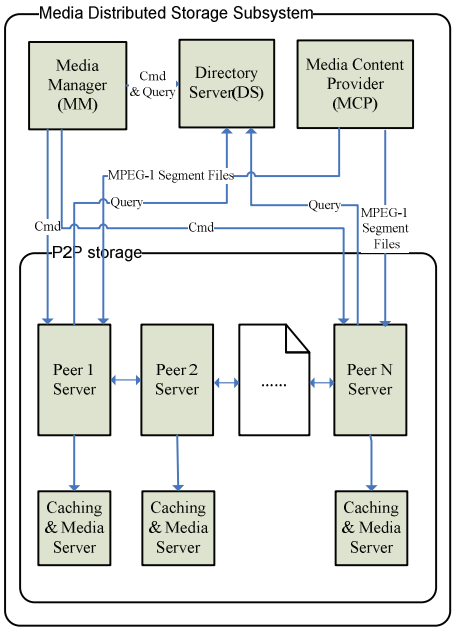

Figure 6. MDN Subsystem Architecture

In the future, we will study the different personalized video retrieval schemes, improve the accuracy of search results by users' evaluations and feedbacks, and enhance the system stability.

REFERENCES

[1] R. Yates, and B. Neto, Modern information retrieval[M]. England: Addison-Wesley Harlow, 1999.

[2] H. Lee, and A. Smeaton, "The Fischlar digital video recording, analysis and browsing system"[C]. Paris: Proceedings of the RIAO 2000-Content-based Multimedia Information Access, 2000.

[3] A. Amir, and J. Argillander, IBM Research TRECVID-2005 video retrieval system. Washington DC: TRECVID Workshop, 2005.

[4] M. Campbell, and A. Haubold, IBM research TRECVID-2006 video retrieval system. TREC Video Retrieval Evaluation Proceedings, 2006.

[5] T. Syeda-Mahmood, and S. Srinivasan, "CueVideo: a system for cross-modal search and browse of video databases"[C]. Computer Vision and Pattern Recognition, 2000.

[6] Michael G. Christel, "Carnegie Mellon University Traditional Informedia Digital Video Retrieval System"[C]. Proceedings of the 6th ACM international conference on Image and video retrieval, 2007.

[7] Worring, M.Snoek, C.G.M.de Rooij, O.Nguyen, G.P.Smeulders, A.W.M, "The Mediamill Semantic Video Search Engine"[C]. Acoustics, Speech and Signal Processing, 2007.

[8] Martinez, J.M. MPEG-7 Overview (version 10). ISO/IEC JTC1/SC29/WG11, 2002.

[9] S. MediaLab, MPEG-7 White Paper. Outubro, 2003.

[10] Z. Dong, and Q. Dong. HowNet and the Computation of Meaning[M]. Singapore: World Scientific Publishing Company, 2006.

[11] C. Zibreira, and F. Pereira. "Image Description and Retrieval Using MPEG-7 Shape Descriptors"[C]. Lecture Notes in Computer Science, 2000, Volume 1923/2000, 332-335, DOI: 10.1007/3-540-45268-0_33,2010.