The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)
Chengdu-Jiuzhaigou, China, August 19–23, 2010

Management of Storage Devices and File Formats in Web Archive Systems

Hiroyuki Kawano

Nanzan University, Aichi 4890863

Abstract Many national libraries are making efforts to crawl and store various born-digital information, there are many difficult problems of the social, legal and technical aspects. In this paper, from the view points of long-term preservation of digital contents, we focus on the the urgent task of storage system, since the size of the web archive is increasing exponentially. In order to archive monotonously increasing contents, we discuss management of storage devices and file formats in web archive systems. Firstly, we propose an architecture of hierarchical storage system based on characteristics of storage devices and file compression formats. Next, we modify the file moving algorithm by using file access frequency. We also evaluate the performance of our proposed algorithm with predicted data based on actual statistics of a web archive system.

Keywords Web Archive, Storage Management, File Moving Algorithm, Hierarchical Storage Systems

1 Introduction

Recent years, in order to preserve and archive contents during very long term, many public organizations such as IIPC (International Internet Preservation Consortium, `www.netpreserve.org`) and national libraries are making efforts to crawl and store the huge volume of born-digital contents, including scientific, artistic and cultural information [7]. Many researchers discuss various technical problems in order to develop web archives. For example, we have a problem that the number of pages published on the web servers is appearing and disappearing. The volume of the web contents is increasing exponentially, so it is becoming hard to crawl and keep hyperlink structures of entire web contents.

We discuss many crawling and preserving problems from various technical aspects[5, 4, 6] in order to archive monotonously increasing digital contents. Optimizing operation costs of web archiving is also important from various web services and implementing advanced web services. Technically, more difficult problem is how to preserve important digital contents from surface and deep/hidden webs selectively or entirely in order to reduce the number of duplicate contents effectively.

Moreover, in order to estimate the quality and value of web contents, we improve the technologies of information retrieval techniques for multimedia contents, and have to further consider technologies of emulation and migration for contents described by various applications and intellectual properties of copyright/copyleft/creative commons.

After considering these problems discussed in previous researches, we focus on strategies to preserve digital contents from the view points of hierarchical storage systems again.

2 Web Archive Systems

As we stated in Section 1, many organizations are making efforts to operate archive systems and to preserve huge volume of born-digital contents. For example, well-known web archive is Internet Archive (www.archive.org), there are many other organizations and projects, such as MINERVA, Kulturarw3, netarchive.dk, PANDORA, AOLA and so on. In Japan, WARP (Web Archiving Project,
http://warp.ndl.go.jp/)[3] is operated by National Diet Library (NDL).

We have an introduction of the architecture and technology of web archive systems, firstly we pay an attention to the typical different characters of search engines and web archive systems. In search engines, web robots have the fast gathering function for more popular pages like authority pages, by analyzing the structures of web hyperlinks and directories. On the other hand, in web archiving systems, crawling quality is more important issue in order to preserve the consistency of web histories on web servers. However, it is possible to extend many technologies and programs of web search engines for development of effective web archive systems.

For example, *web robots* are program which crawls web contents from web servers and replicates contents into database systems. There are various crawling programs for archiving, such as heritrix (http://crawler.archive.org/) and others. Typical web robots parse HTML/XML documents and choose important metadata and keywords by using natural language processing techniques of morphological analysis and other heuristic functions. Furthermore, the metadata standards, such as MARC 21, MARCXML, MODS (Metadata Object Description Schema), MADS, EAD, METS, MIX and PREMIS, are useful and helpful for keeping the quality of archiving contents.

Furthermore, *database systems* stores the huge volume of web texts and multimedia contents not only of original files, but also of keywords, creation date, frequency of updation, number of hyperlinks and many other attributes. We make several tables with various attributes, such as URLs, keywords, date, connections of hyper links, types of http servers, IP addresses, and various control/management tables for operating web archive systems.

In order to archive the web publishing the entire contents in storage systems, we also consider long-term preservation carefully.

3 Management of Storage Devices and File Formats

In this section, in order to preserve huge volume of data, we consider an advanced architecture of hierarchical storage systems and moving algorithm presented in Fig.1.

There are researches of file moving algorithms among different levels of memory devices. For instance, in paper[1] using the combination of parameters "Lower Layer Retrieval Rate" R and "Retrieval Consumption Rate" C, they derive a performance measure PCR "Production Consumption Ratio". They propose LRU (Least Recently Used) algorithm with access frequency f_{age} and average file size X in [2], MV "Migration Value" is a performance measure of file moving policy. The update latency of MV is also an important parameter.

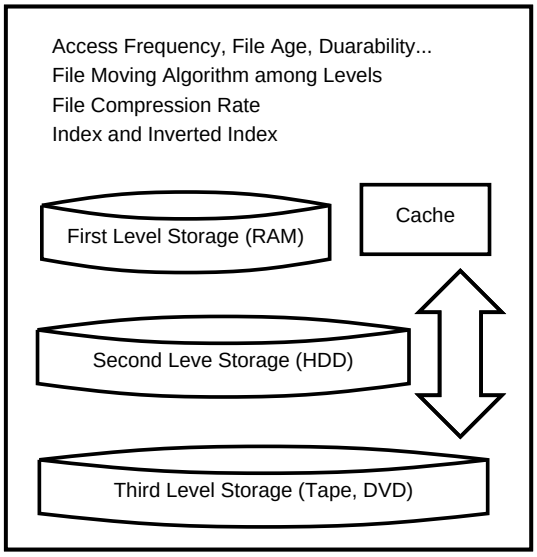

Figure 1: Hierarchical Storage

In the above hierarchical storage systems, it is important to keep the balance of file access frequency and memory device properties. Because the property of memory devices is quite different from the view points of access speed, volume, durability and others. Many researchers discuss the architecture of hierarchical storage system using tape devices, SSD, HDD and other cache memory devices. In Table1, we present characteristics of momory devices shortly.

Furthermore, in Table2, the size of files strongly depends on types of the file, such as plain and various compressed formats, moreover files with very low frequency access have to be tar+gzipped with chunks of files.

Memory Device	Size	Transmission Speed	Durability (years)
RAM	10GB	over 100Mbps	8
HDD	2TB	over 10Mbps	5
DVD-ROM	4GB	Over 1Mbps	20
Magnetic Tape	35TB	Over 1Mbps	30
Blu-Ray Disc	25GB	Over 10Mbps	20
Holographic Memory	over 1TB	100Mbps	20
Microfilm	-	-	500-900
Paper	-	-	1000-

Table 1: Characteristics of Memory Devices

File Formats		txt	html	pdf	jpg
compress of a file	lha	33.14	26.84	82.99	98.04
	gz	30.94	26.08	73.31	97.85
compress of groupe files	lha	32.88	27.57	96.53	98.74
	tar+gz	30.70	25.31	73.26	97.89

Table 2: Compressed Rate of File Formats (%)[4]

4 Hierarchical Storage in Archive Systems

There are major three formats of files, "Text (txt,html), Documents (pdf,doc), Multimwedia (jpg,gif,wmv)", we consider the hierarchical architecture of storage systems with various parameters, such as total volume, file size and compression rate presented in Table3. We estimate the total volume of storage systems. If all files are stored without compression, the total volume of storage size S_{all} is $S_{all} = S_t + S_p + S_m$. With compression depending on the file access rate, S_{all} in the i-th level storage is given by the equation with S_{ti}, S_{pi}, S_{mi}.

File Types	Size	Compression Rate/Time one file	archive file
plain text	S_t	Cs_t/Ds_t	Cm_t/Dm_t
XML/PDF documents	S_p	Cs_p/Ds_p	Cm_p/Dm_p
multimedia	S_m	Cs_m/Ds_m	Cm_m/Dm_m

Table 3: Property of Preserving File Format[4]

Depending on file access frequencies, our storage system moves those files among different storage levels satisfying with various restrictions. Therefore, we extend File Aging Algorithm proposed in [1].

Modified File Moving Algorithm

1. Firstly, all files are stored into the first level of storage system.
2. Value MV is updating and Moving cycle is defined by $Date$ daily.
   (a) A new file has the folloing value
   $MV = (X/Size) \cdot f_{age}$
   (b) Value of archived files is updated,
      i. If the file is accessed today
      $MV = MV + (X/Size) \cdot f_{age}$
      ii. If file access does not happen
      $MV = MV \cdot f_{age}$
3. Depending on $Date$, Values of MV is sorted, files are compressed and moved higher/lower levels.

For example, according to the previous report[8] about WARP in National Diet Library, plain text and HTML/XML files are 10% of 5TB. it is possible to compress these

files. Furthermore, 20% of web contents are updated per a month and 60% of contents are updated per three months[9]. Therefore, it is possible to edit and modify hyperlinks in these files, intelligent navigation systems can reduce more volume of storage area in order to compress files except documents, pdf and multimedia files. It is possible to reduce storage volume of 90% of duplicate files which are crawled by web robots.

Furthermore, 44 web sites are crawled every month and 2,145 web sites are crawled per three months, 40% of archive files are duplicated in storage systems, therefore it is possible to reduce at most 30% of storage volume except 10% of text/HTML/XML files with modified hyperlinks. As a result, it is possible to reduce storage cost from 15% to 30% by using intelligent crawling robots and navigation systems.

We assume the following model for performance evaluation, the volume of third level storage is 50 times volume of second level storage, the volume of second level storage is also 50 times volume of first level storage. 10,000 files have normal distribution, the distribution of file access is Zipf with 100(access/day).

Fig. 2 presents the change of the migration value, *MV* of file with highest access frequency increases and *MV* of file with lowest access frequency decreases rapidly, *MV* is converging to a constant value gradually. Therefore, we can realize rather stable storage system by using our proposed "File Aging Algorithm".

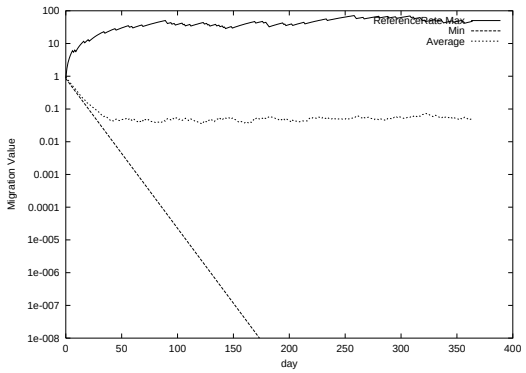

Figure 2: Change of MigrationValue[4]

5 Conclusion

In this paper, we discuss the architecture of long-term hierarchical storage systems in order to preserve digital contents, and propose modified file moving algorithm with access frequency and format translations. WARC format is standardized at present, file compression format is also important in long-term storage systems.

Acknowledgements

A part of this work is supported by “the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 19500098, 2010” and “2010 Nanzan University Pache Research Subsidy I-A-2”.

References

[1] A.E. Dashti and S. Ghandeharizadeh, “On Configuring Hierarchical Storage Structures,” (online), available from http://citeseer.nj.nec.com/119042.html.

[2] T. Gibson and E. L. Miller, “An Improved Long-Term File-Usage Prediction Algorithm,” (online), available from http://citeseer.nj.nec.com/310544.html.

[3] N. Hirose, “Practice and Challenges on Web Archiving at the National Diet Library, Japan: The Internet to be a Stable Intellectual Infrastructure,” IPSJ SIG Notes (Information Processing Society of Japan), No.DBS-130-12 and No.FI-71-12, 2003. (In Japanese)

[4] M. Kojo, N. Hirose and H. Kawano, “Web Archive: Proposal of Storage Systems for Long-term Preservation,” DBSJ Letters, Vol.3, No.4, pp.9-12, 2005. (In Japanese)

[5] H. Kawano, “Web archiving strategies based on web log mining patterns,” 2004 CORS/INFORMS International Meeting INFORMS, TC18, 2004.

[6] H. Kawano, “Strategy of Digital Contents Archive Based on Reputation Model," Proc. of 19th International Conference on Systems Engineering, IEEE ICSENG ’08, pp.288-293, 2008.

[7] Lyman, P., “Archiving the World Wide Web", Building a National Strategy for Preservation: Issues in Digital Media Archiving,” 2002.4, (online), available from http://www.clir.org/pubs/reports/pub106/web.html.

[8] NDL, "Survey on Comprehensive Collection, Storage, and Archiving of Japanese Web Sites Outline," 2005 (online), available from http://www.ndl.go.jp/en/aboutus/bulkresearch2005summary_e.html

[9] J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres and S. Decker, "Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources," Proc. of LDOW2010, 2010(online), available from http://events.linkeddata.org/ldow2010/papers/ldow2010_paper12.pdf