The 7th International Symposium on Operations Research and Its Applications (ISORA'08)
Lijiang, China, October 31–Novemver 3, 2008


# Reputation-based Contents Crawling in Web Archiving System

Hiroyuki Kawano*

Nanzan University, Aichi 4890863

**Abstract** The size of the web archive is increasing exponentially, many national libraries are making efforts to preserve born-digital scientific, artistic and cultural contents. However, in order to crawl and store huge volume of digital information, it is very hard to resolve various problems from the social, legal and technical view points. In this paper, from the view points of long-term preserving digital contents with good reputation of trustiness, uniqueness and valuation, we discuss strategies to preserve monotonously increasing digital contents on web servers. According to experimental results of our reputation model, it makes possible to crawl socially valuable contents for archiving.

**Keywords** Web Archive, Web Crawling, Reputation Management

## 1 Introduction

Recent years, the size of the web systems is increasing exponentially, so it is becoming hard to keep the quality and social structure of web contents and to preserve valuable web resources. For example, in 2001, there exist 1 billion pages on surface web and 550 billion pages in deep web [1] . In 2003, the volume of web data is 167TB of surface web and 92PB of deep web [2] .

Furthermore, the number of pages published on the web servers is appearing and disappearing. Many public organizations such as "National Libraries" and IIPC (International Internet Preservation Consortium, www.netpreserve.org), are making efforts to preserve these contents[2] in order to preserve the huge volume of born-digital information in the internet, including scientific, artistic and cultural contents provided by various web systems. Many researchers discuss various technical problems in order to develop better web archives.

Therefore, in order to archive monotonously increasing digital contents, we also discuss many crawling and preserving problems from various technical aspects[10]. For instance, there are optimizing problems of hardware and network costs for operation of archiving service and execution of web crawling from various web services and systems[4,

*From November in 2002, the author investigates web archiving strategies as a part-time researcher of Digital Library Section in National Diet Library in Japan.

[1] http://www.brightplanet.com/technology/deepweb.asp

[2] http://www2.sims.berkeley.edu/research/projects/how-much-info-2003

11]. Further difficult problem is how to gather digital contents from surface and deep/hidden webs selectively or entirely.

Moreover, in order to estimate the quality and value of web contents, definitions of metadata formats like URI/RDF/MODS (Metadata Object Description Schema) are also problems. We have to 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.

In this paper, we focus on strategies to preserve digital contents provided by web systems from the view points of good reputation of trustiness, uniqueness and valuation[7]. Mainly, we discuss policies of contents crawling programs using reputation models. Firstly, we introduce applying reputation models which we proposed in P2P contents distribution systems[6]. We discuss the behaviors of contents distribution as a reputation model, and we make clear that the proposed scheme achieves one of suitable strategies. We also show simulation experiments and present dynamic characteristics of reputation rates of contents holders, we discuss the strategy of crawling contents depending on reputation values.

# 2 Web archive systems

As we stated in Section 1, many organizations are making efforts to build archive systems and to preserve huge volume of born-digital contents. For example, well-known web archive is Internet Archive (www.archive.org), and we have WARP (Web Archiving Project, http://warp.ndl.go.jp/)[5] by National Diet Library (NDL) in Japan. There are many other organizations and projects, such as MINERVA, Kulturarw3, netarchive.dk, PANDORA, AOLA and so on.

In this section, we have short summary of the architecture and technology of web archive systems, firstly we present the typical different characters of search engines and web archive systems shown in Table 1. It is possible to extend many technologies of web search engines for development of web archive systems. Basically web search engines, including our developed Mondou[8, 9], consist of three modules, *web robots*, *database systems*, *search programs*, we can implement various advanced technologies which are based on the research results of database system, information retrieval, data mining, text/web mining, information visualization and so on[3].

The first module, *web robots*, is the program which crawl web contents from web services and replicate contents into database systems. Typical web robots parses HTML/XML documents and choose important metadata and keywords by using natural language processing techniques of morphological analysis and other heuristic functions. There are many crawling programs, such as heritrix (http://crawler.archive.org/), wget and others.

Furthermore, in search engines, web robots have the fast gathering function for more popular pages like authority pages, by analyzing the structures of web hyper links and directories. In web archiving systems, crawling quality is more important issue in order to preserve the consistency of web histories on web servers. We also developed the cooperative distributed web robots[12].

The second module, *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 need several tables with

Table 1: Differences between Search Engine and Web Archive System

| | Web Search Engine | Web Archive System |
|---|---|---|
| Crawling | Freshness by time stamps and informative file types: html, text, pdf, doc and others | Accurate crawling of entire web pages stored in target web sites, as rapid as possible |
| Quality | Focusing on special attributes and descriptions: title, meta, hyperlink tags | Quality control is strongly required (original/master copies, archiving shots management) |
| Search | Recall and Precision (results influenced by commerciality, simple and easy query input) | Difficulties of document searches (historical change and heterogeneous keywords, evolution of hyperlink structures) |
| Preservation | Short time: several months (popular and fresh web pages by users preference) | Long time: several centuries as paper, micro film etc. migration, transformation |

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 addition to these typical attributes, we have to consider time attributes carefully, in order to preserve the web publishing sequences in the entire web archive systems. The metadata standards, such as MARC 21, MARCXML, MODS, MADS, EAD, METS, MIX and PREMIS, are useful and helpful for describing the quality of digital contents.

The final module, *search programs*, requires the most complex technologies, which are executed by queries of database systems. Moreover most of users independently require personal search results and patterns, trends, knowledge derived by using social filtering and advanced data mining techniques. Actually, in the steps of web documents retrieval, it is difficult to choose suitable combination of keywords in order to discover the meaningful results of documents from search results. So, statistics, frequencies, topics, trends, experience and other support techniques are utilized for this purpose.

# 3 Reputation model of contents distribution systems

Reputation model is emerging research fields, several techniques of reputation models play important roles in order to improve web archiving systems.

## 3.1 Reputation model of web services and systems

In order to handle monotonously increasing digital information, we have to consider many difficult problems of long-term preservation from various technical aspects. Here, we discuss trusted resource optimization problems of archiving contents. Several techniques of data mining, such as machine learning, inductive learning, knowledge representation, statistics and information visualization, with considering characteristic features of databases, play important roles in order to compute values of importantness in the network

systems.

In the network systems without the central organizations, it is very hard to evaluate and decide selection of digital contents with valuation, uniqueness, trustiness and importance. In our previous researches[6], we proposed our reputation computation algorithms based on trust chains of server connections, and also discuss the properties of the maximum utility function $u_{ij}(t)$ from server i to client $j$ at time $t$. In the following subsections, we try to apply straightforwardly our reputation model to the measurement of contents preservation.

$$u_{ij}(t) = \alpha d_{ij}(t) + \beta U_{max} + \gamma f(R_{ij}(t)) \tag{1}$$

In the above function $u_{ij}(t)$, we use the following parameters:

- $d_{ij}(t)$: at time $t$, total access counts of providing contents from server $i$ to client $j$ and reward parameter $\alpha$
- $U_{max}$: maximum service resources and voluntary contribution rate $\beta$ without any requirement
- $R_{ij}$: at time $t$, reputation value from server $i$ to client $j$ and reputation-based contribution rate $\gamma$
- $f()$ is a monotonously increasing function, with 0 at $R_{ij}(t) = 0$ and $U_{max}$ at $R_{ij}(t) = 1$

Depending on combination of the values $\alpha$, $\beta$ and $\gamma$, we classify 6 different behaviors of various servers presented in Table 2. Some servers and clients cause free-riding problems in the internet[1].

Table 2: Parameter variations of servers with different behaviors

| Type of servers | $\alpha$ | $\beta$ | $\gamma$ |
|---|---|---|---|
| $T_{free}$ (free riders) | 0 | 0 | 0 |
| $T_{rec}$ (balanced service) | 0 | $0 < \beta < 1$ | 0 |
| $T_{rep}$ (reputation-based service) | $0 < \alpha < 1$ | 0 | $0 < \gamma < 1$ |
| $T_{pos}$ (with positively contribution) | $0 < \alpha < 1$ | $0 < \beta < 1$ | $0 < \gamma < 1$ |
| $T_{cont}$ (with voluntary contribution) | — | 1 | — |
| $T_{mal}$ (malicious service) | — | 1 | — |

## 3.2 Web crawling algorithms

Web robots crawl web resources by using http protocols, and they analyze web services and contents and store them into database systems. When making request for a document retrieval, usual robots check the "`robots.txt` file on web servers. Therefore, web administrators are able to manage the behavior of web robots by using this file. Moreover, web robots use some scores to evaluate importantness of web pages, based on analysis of "title, headings and sub-headings, anchor strings, and so on". Typical web robot visits web servers sequentially with the breadth-first manner.

**Algorithm 1.**
(Breadth-first traversal per server)

1. Define $S_0$, $S_1$ sets of all the servers that have been found.
2. **foreach** $s \in S_1$ **do**
3. Get information about documents $d$ on $s$ from *Database*.
4. Robot evaluates and analyzes $d$, and stores data to *Database*. At the point, if he discovers unknown servers, he adds them to $S_0$. if **not** then **stop**.
5. $S_1 = S_0$ and **continue** to 2.

In order to operate web robot programs for web archive systems suitably, it is very important to gather web pages selectively and entirely. Here, in order to gather much more meaningful web contents deeply, the robot administrators have to decide the appropriate selection policies. Then, according to the defined policy using $u_{ij}(t)$, web robot programs gather trusted web contents automatically.

## 3.3 Crawling policies by reputation model

When we browse web pages, we also recognize and evaluate the characteristics and importantness of web pages. Here, we try to characterize the importantness and popularity of web contents, in order to improve crawling techniques of important servers of web services for preservation of web sites. Furthermore, in order to preserve the web contents, we have to keep the consistency of web pages, such as updating sequences of web contents, connectivities of hyperlinks to other web servers and navigational consistency.

Therefore, we evaluate the reputation of web contents as loosely connected social networks with time attributes from first published (or appearing) time to expiry (or disappearing) time $(t_p, t_e)$ on the web server, and we try to discover the characteristics of the popular web pages based on the following types of hyperlinks.

- **Inner link:** the link to other pages on the same web server.
- **Outer link:** the link to other pages on the other web servers.

We estimate the quality of hyperlinks in the web by the values $u_{ij}(t)$ of ($D_p$, $T_p$, $S_p$; $I_p$, $O_p$), which is the combination of existent duration, modified times and file size of a web page, the number of inner links, and the number of outer links for the parent document $p$. Then, each child document $C_k$ is referred from the parent $p$, we also calculate $(D_k, T_k, S_k; I_k, O_k)$ for $C_k$.

Then, in order to evaluate the quality of contents $C_k$, we proposed the following traversing algorithm, based on the assumption that the important web pages are referred many times from other servers. We also define the score function $p_k$ based on the above attribute values related web pages.

**Algorithm 2.**
(Selecting a URL on a server)

1. Define the following sets about server $s$:

   $G_s$**:** Set of URLs on $s$ to be obtained

$K_s$**:** Set of URLs on $s$ that has already been obtained
$O_s$**:** Set of URLs on $s$ to be obtained referred from other server

2. **if** $G_s \neq \phi$, **then**
3. Get a URL $d$.
4. **else**
5. **if** $O_s \neq \phi$, **then** $G_s = O_s, O_s = \phi$, and **go to** 2.
6. **if** $K_s \neq \phi$ **do**
7. Find $k$ that has largest score in $K_s$.
8. $G_s$={ $h|k$ refers $h\}$, $K_s = K_s - \{k\}$, and **go to** 2.
9. **else** exit since there are no URLs to obtain.
10. $G_s = G_s - \{d\}$.
11. Let Robot get $d$.
12. $K_s = K_s \cup \{d\}$.

# 4 Experimental Results and Discussion

In [6], we had simulation experiments using the following system parameters:

- Servers: 1000
- Maximum chains of trustiness: 6
- Lower bound of threshold value: 0.97
- (several parameters are omitted in this paper)

In Figure 1, we present dynamic behaviors of reputation rate in the system with various servers, here we present two simulation results with the different incorrect service ratio. The incorrect service ratio is the rate of services results which do not match request conditions.

In figures, $T_{pos}$ and $T_{rep}$ keeps better reputation rate than $T_{rec}$ and $T_{free}$. Both of reputation rates of $T_{rec}$ and $T_{free}$ decrease rapidly, it becomes difficult to utilize the system resources. We omit properties of $T_{cont}$ and $T_{mal}$, since they provide only voluntary contribution and malicious service consuming their own maximum resources. $T_{cont}$ and $T_{mal}$ servers behave specially without any evaluation of reputation from other servers.

By using the value of reputation rates, we can crawl digital contents with strong effects in the systems.

# 5 Conclusions

The size of digital world in the information society is increasing exponentially, it makes difficult to preserve the informative or valuable contents including rich information and knowledge for future generations. Web archive is one of dominant information infrastructure in digital information society.

In this paper, we try to extend the mathematical model of reputation in [6], we discuss the strategy of long-term preservation based on reputation regarding importance, fairness, trustiness, uniqueness and valuation. By using evaluation of our reputation model, it is possible to select trusted services and contents for archiving. Especially, we also discuss that the evaluation function provides the flexible and dynamic contents crawling mechanism based on the reputation model.

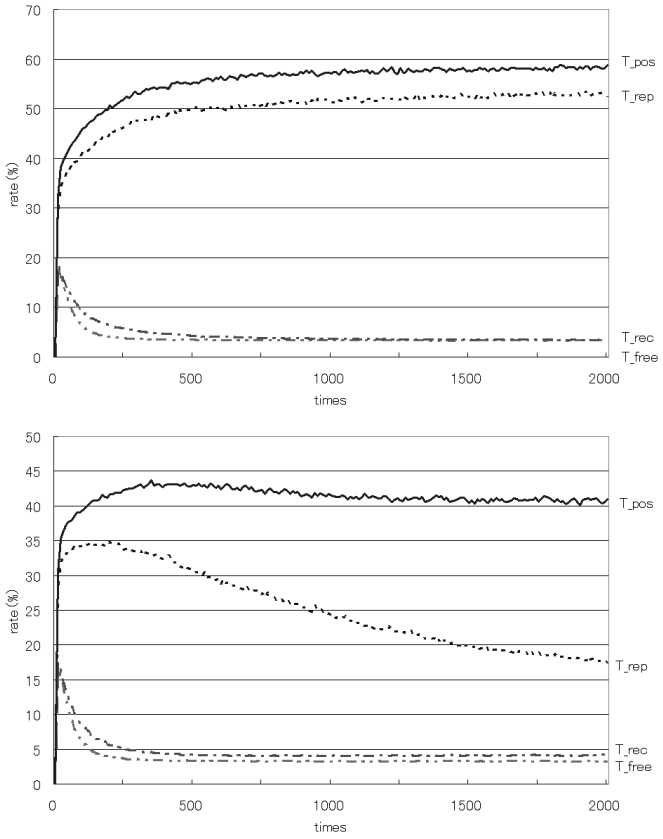


Figure 1: Obtaining service rate for various peer types (upper : percentage of false service = 0, lower : percentage = 0.05)

## Acknowledges

A part of this work is supported by “the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 19500098, 2008” and “2008 Nanzan University Pache Research Subsidy I-A-2”.

# References

[1] E. Adar and B. Huberman, “Free Riding on Gnutella,” First Monday, Vol. 5, No.10, Oct. 2000.

[2] S. Abiteboul, G. Cobéna, J. Masanes and G. Sedrati, “A First Experience in Archiving the French Web,” Proc. of ECDL 2002, Lecture Notes in Computer Science, No. 2458, pp.1–15, 2002.

[3] S. Chakrabarti, “Mining the Web: Analysis of Hypertext and Semi Structured Data,” Morgan Kaufmann, 2002.

[4] D. Geels and J. Kubiatowicz, “Replica Management Should be a Game,” Proc. of the 10th Workshop on ACM SIGOPS European Workshop, pp. 235–238, 2002.

[5] 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)

[6] Y. Ito and H. Kawano, “Reputation model for evaluation of reputation in P2P environment," IEICE Transactions on Information and Systems, Vol.J91-D, No.3, pp.628–638, 2008. (in Japanese)

[7] A. Josang, R. Ismail and C. Boyd, “A Survey of Trust and Reputation Systems for Online Service Provision,” Decision Support Systems, Vol.43, No.2, pp.618-644, Mar. 2007.

[8] H. Kawano, “Mondou: Web Search Engine with Textual Data Mining,” Proc. of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp.402–405, 1997.

[9] H. Kawano and M. Kawahara, “Mondou: Information Navigator with Visual Interface,” Data Warehousing and Knowledge Discovery, Second International Conference, DaWaK 2000, pp.425–430, 2000.

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

[11] K. Ranganathan, M. Ripeanu, A. Sarin and I. Foster, “To Share or not to Share: an Analysis of Incentives to Contribute in Collaborative File Sharing Environments,” Proc. of the 1st International Workshop on Economics of Peer-to-Peer Systems, 2003.

[12] H. Yamana, K. Tamura, H. Kawano, S. Kamei et al., “Experiments of Collecting WWW Information using Distributed WWW Robots,” Proc. of SIGIR’98, Melbourne, Australia, pp.379–380, 1998.