A New Data Storage and Service Model of China Web InfoMall[1]

Hongfei Yan, Lianen Huang, Chong Chen and Zhengmao Xie

Computer Networks and Distributed Systems Laboratory,
Department of Computer Science and Technology, Peking University, P.R. China
{yhf, hle,cc,xzm}@net.pku.edu.cn

Abstract The Web consists of enormous pages which is easier vanishing than traditional media such as newspaper, journals. To preserve the web resources, we began the China Web archiving project, named Web InfoMall, from 2001. The paper describes the data storage and service model of Web InfoMall 2.0 to meet the goals of collecting the stuff broadly, storing them perennially, and locating requests efficiently. Currently the Web InfoMall holds 0.7 billion pages (10.6 terabyte) together with 5 terabyte of digital web resources other than web pages, having the ability of collecting more than 1 million pages per day, a storage capacity to hold more than 10 billion pages (about 150 terabyte), and a scheme to manage large numbers of pages.

Keywords: *web resources, Web InfoMall, distributed crawling, Tianwang storage format, service model*

Introduction

The Internet was introduced to China in 1994, and the first web sites emerged from 1995. From then on, the China Web got an unprecedented development. At the end of 2003, the registered sites amounted to 595,550 with ".cn" domain suffix [CNNIC,2004]. During the nine years, the China Web enjoys exponential growth, with the same rate to the whole Web. Tianwang[Tianwang,2004] crawled more than 60 million web pages up to the July of 2002, and 105 million pages in January of 2003. According to a quantitative analysis, Chinese pages is about double every year[Li,2003]. To store as integrality as possible, it is necessary to have an effective repository that is capable of holding 1 billion pages, as well as supporting efficient access. From the beginning of 2001, a reaction to this state of want was initiated under the leadership of Xiaoming Li at Peking University. Such an effort leads to the born of the Web InfoMall (http://www.infomall.cn). The goal of the Web InfoMall is

[1] This work is supported by the China 973 Grant (G1999032706) and the MOE project (20030001076).

to construct a platform that supports Web information storage, exhibition, research, development, and application. Its main tasks are permanently collecting web pages into a repository, providing free access to the archived data, and offering an open testbed to encourage multidisciplinary study. In June 2004, the Web InfoMall[2] has been keeping about 0.7 billion web pages and attends to over 10 thousand query requests per day.

In this paper we begin with an introduction of the design, implementation and application of the Web InfoMall 2.0. We follow this with ongoing research. Finally we present related work and conclusion.

The distributed crawling architecture

The reasons why we chose the distributed crawling architecture come from the need of the information scale and the cost of deployment. The number of Web pages is tremendous. Connected by hyperlinks, they reside at thousands of independent sites. Single processor system do not have the ability to manage huge volumes of data due to process ability and disk capacity, no matter to say keeping up with the rapidly growth of the Web. It is a natural choice to adopt parallel processing technologies. The parallel processing schemes of symmetric multiprocessing, massively parallel processing, and not uniform memory access, are high performance but hardly enjoy strength by contrast with the clustering scheme. Clustering is the most suitable architecture for applications as Web archiving, which is not only a cost-effective architecture, but also with high scalability. With this scheme, we can meet the objective of collecting pages - simple operations, low communications between processes, high requirement to disk capacity and speed of I/O.

The basic task of a crawling subsystem is to acquire web pages. Its functions include: picking a url from the *task pool*, resolving IP address through a DNS, creating a connection to the web server, sending a request, and receiving the response from the server, closing the connection, parsing the captured web page, adding new links to the task pool, and storing pages to hard disks. Task pool maintains unvisited URLs list.

Figure 1 shows the architecture of the parallel and distributed crawling subsystem of the Web InfoMall. Every two *coordinator processes* establish connections and form a logical strong connected graph. Coordinate process is the process that manages crawlers to fetch pages on the Web. Its main function is to assign c*andidate URLs* to crawlers, store web pages returned by its crawlers and communicate with other nodes. It is the core part of the crawling subsystem and each node runs a coordinate process in our distributed subsystem. Candidate URL is the URL of a web page that has not yet crawled but linked to by one of the crawled pages. They are usually maintained in an unvisited URL list.

[2] The Web InfoMall 1.0 was launched in the beginning of 2002, and was updated into version 2.0 at the end of 2002. The Web InfoMall 2.0 rewritten is very different from the version 1.0.

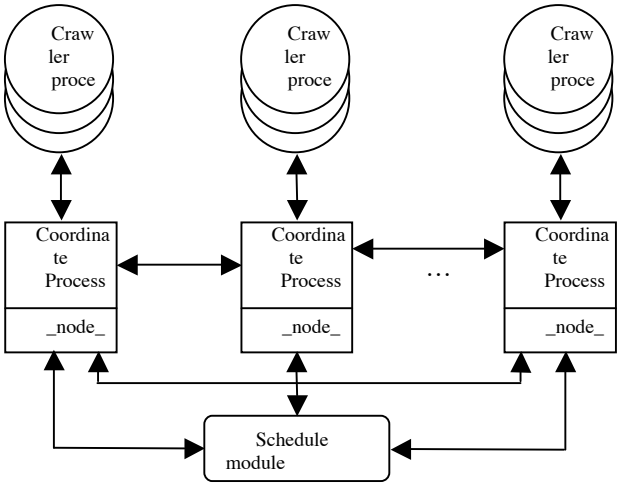

Figure 1 The architecture of the parallel and distributed crawling subsystem

The schedule module stores information including IPs and ports of all registered coordinate processes in the system. When the state of any coordinate process changes, the schedule module will deliver the updated information to other coordinate processes. Every coordinate process sets a few hundred crawlers to collect web pages in a certain domain scope. The crawler receives URLs from its coordinate process, fetches the web pages pointed by the URLs and passes the content of pages to the associated coordinate process. To get full-duplex communication, every two coordinate processes has two connections. When any coordinate process finds URLs not belonging to its scope, it will send them to the responsible coordinate processes. To minimize traffic communication, each coordinate process only deliver URL strings.

[Yan, et al.,2001a] gives a detail description of this distributed crawling architecture, and [Yan, et al.,2001b] presents an efficient dynamic reconfiguration model that can be used in the architecture to bring the whole system in a coordinated state when the nodes are added to or removed.

The storage strategy and service model

To deal with large volumes of data, the low cost storage and efficient access are two fundamental issues to be considered. We designed Tianwang format for raw web pages storage. This format stores the raw pages sequentially without the storage fragment waste. Also it has fault tolerance to an unexpected hardware or software failure. The archived data in terms of the format can be opened for sequential and random read. They are respectively used for batch-pages acquiring and single-page

browsing. To implement these access operations, we designed the efficient service model as is detailed in the following.

3.1 Tianwang storage format for raw web pages

The goal of web pages storage is long-time preservation and multi-application oriented. So the format should be simple enough and convenient enough to use.

Yet we face the following challenges, 1) the size of raw web pages is not regular – it may range from lKB to several MB. Individual files will usually require more space to store than the true size of the file, because the block size is more than one byte, and a block is never divided between multiple files. Suppose the size of a page file is 6KB and the block size is 4KB, the file will occupy 2 blocks. If the space waste in storing a single file will be 2KB, how much will cost in storing numerous files? 2) the life of storage device is not unlimited and the system software is not absolutely robust, so the storage format should have recovery property to resist the unexpected hardware or software failure and minimize data losing. If partial data lost, the remain data should still be available. In short, we should save storage space as much as possible and enhance the fault tolerance in our storage format.

Definition of Tianwang storage format

According to the previous considerations, the definition of Tianwang format is:

1) every record includes a raw data of a page, records are stored sequentially, without delimitation in between.
2) a record consists of two parts: header (HEAD), data(DATA), and ends with a line feed ('\n'), i.e., HEAD + a blank line + DATA + '\n'.
3) the header part consists of some attributes. Each attribute is a non blank line. Blank line is forbidden in the header.
4) an attribute consists of a <name,value> pair, with delimitation ":".
5) the first attribute of the header must be the version attribute, i.e., version: 1.0.
6) the last attribute of the header must be the length attribute, i.e., length: 1800
7) all names of attributes are in lowercase for the sake of simplicity.

Note: A blank line is only composed of one line feed ('\n' in C language). In Microsoft Windows system, a new line is composed of one carriage return and one line feed ('\r\n' in C language). However in most Unix system, a new line means only one line feed. We prefer to Unix's new line schema.

Version 1 of Tianwang storage format

Storage format may have many version to meet future extending requirements. Currently Tianwang storage format is version 1. Below is shown a typical record for a given web page. Comments are behind double splash token.

```
version: 1.0                                    // version number
url: http://www.pku.edu.cn/                     // URL
origin: http://www.somewhere.cn/                // original URL
date: Tue, 15 Apr 2003 08:13:06 GMT             // time of harvest
ip: 162.105.129.12                              // IP address
unzip-length:          // If included, the data must be compressed
length: 18133                                   // data length
                                                // a blank line
XXXXXXXX                                        // the followings are data part
XXXXXXXX
….
XXXXXXXX                                        // data end
                                                // insert a new line
```

Semantics of all attributes of the head part:

Version, version number. The following explanation applies to version 1.0.

Url, the URL of a web page. If web server's response message has a location field, the Url means the actual URL location. This attribute is necessary.

Origin, the original URL. When web server's response message has a location filed, the "origin" means the derivation URL.

Date, the time of storing, its format is conformed to date and time specification in RFC822. This attribute is necessary.

Ip, the IP address of web server holding the page.

Unzip-length, the original size of the data part. If included, the data must be compressed.

Length, the length of the data part. If the raw page is not stored in compress mode, the value equals to the page length.

Finding correct records from remnants

If data have been partly destroyed, the following measures could be taken to find remnant records.

1) Find the flag string "version" and record the position of the string.
2) Judge the subsequence data. If they are not satisfied with the head part of Tianwang storage format, go back to 1) for next string "version", otherwise, suppose this is a correct record, go on checking the next two records.
3) If we can find three continuous correct records, we might as well consider the record of the first step is correct and we can extract all correct web pages.

Because raw web page format is random and Tianwang format is strict, it is very little likelihood of finding error records. So all exacted records are acceptable.

3.2 The service model

Reasonable service model can quickly drive the mass data from disorder to regularity with good integrity. And thus the users can access them conveniently and quickly. The service model in Web InfoMall mainly includes data indexing and service providing. To illustrate our view, we use the architecture in Figure 2.

Since the raw pages are continuously collected from the Web, they are stored in Tianwang format in crawling time order. This order is natural and significance for both successive research and process. We tokenize the raw data batch in month, e.g., "200405" representing the web pages collected in May 2004. Each batch of data is independently stored and backup before being indexed for service.

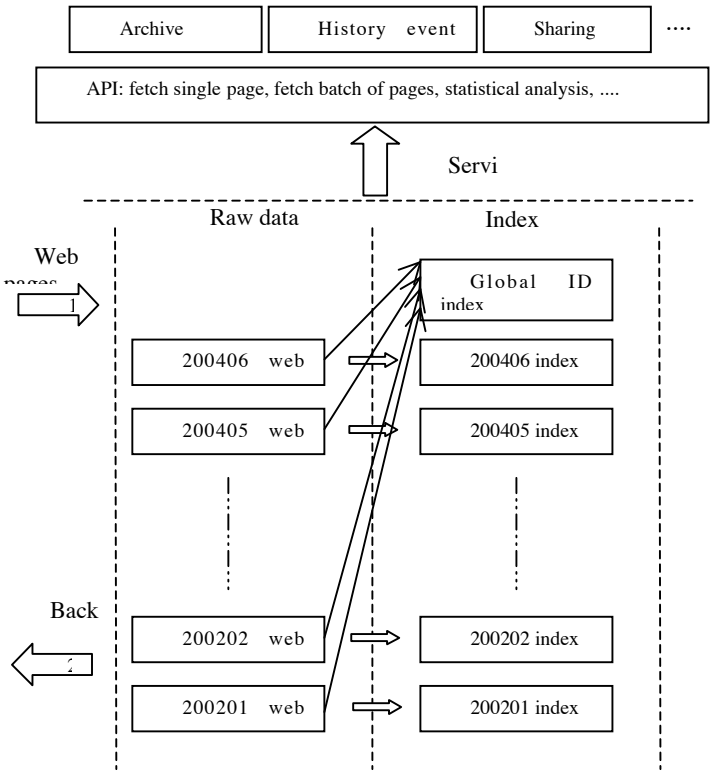

Figure 2 The architecture of the service model

Each raw page is given an ID which is stored into *Global ID index*. The ID is composed of the batch number (using year-month as token) and the unique serial

number (a 32-bit integer) in its batch. To locate a page, we need to construct two level indexes, one for batch numbers, and the other for pages in a batch of data.

The indexed data can meet the requirements from the diverse APIs, such as fetching single page or batch of pages, acquiring a collection of pages relating to certain subjects from different batches, offering statistical analysis on the archived data. The current Web InfoMall APIs are able to provide services including archive browsing, historical event focus and sharing archive.

Applications

Up to now, the Web InfoMall contributes to the China Web archiving and research mainly in three aspects, 1) entirely preserving China Web data from 2001 to now, and offering open access for research. 2) constructing all-dimensional, and wide-ranging miniature of web social evolution by multiple service APIs. 3) promoting multidisciplinary study, helping the humanities use mass science data in their research.

The Web InfoMall reposits about 1 million web pages every day, besides offering its own daily queries respectively, supports the historical pages cache for Tianwang search engine[Tianwang,2004]. We advocate the sharing of China scientific data. The Web InfoMall acts by publishing data storage format, and providing data to social circles. In China, the beneficiary research units of the Web InfoMall include Peking University, Tsinghua University, Chinese Academy of Sciences, Shanghai Jiaotong University, Renmin University of China, Harbin Institute of Technology, etc.

Based on the huge volumes of the Web InfoMall data, we developed some applications to produce the macro panorama of the China Web in time and spacial dimension. The results reflect the China social evolution trend. For example, we built a model to trace web information on certain topics[Li, et al.,2003]. Using this model, it is possible to obtain, for a specific topic, the strength of presence on the Web from a variety of angles, having worth for social scientific research. Based on this model, an experiment was conducted using “16th Congress of China Communist Party” as the topic from October 22nd to November 24th, 2002. The experiment results show that the amount of information relating to the 16th Congress is 7.3% among all the information, and the amount of topical information exhibites a strong taking off from November 2nd, and reached its peak on November 20th. Another experiment was done in April of 2003, we did a local analysis on web activities in Guangdong province[Yan,2003]. It was the time that SARS was epidemic and American launched Iraq war. We specially choose government web sites for representative. The experiment results show that the sites mentioned SARS amount to 56.4% among all sites, and Iraq war 38.5%. The phenomenon of web information has close relation to historical events during the given period of time.

Based on the last modify time of pages, we estimate change frequency of Web and life age of pages. The experiment result shows that life duration of pages obeys exponent probability distribution, and the average update interval of 50% pages is less than two months. Using the same method to various types of sites, the experiment

result shows that half-life of commercial sites is less than two month, and government or education sites is about four months.

To evaluate the quality of Chinese information retrieval system, promote the research of this field, we built a 100 gigabyte test collection based on the Web InfoMall in June 2004, named CWT100g[CWT,2004]. All the samples were carefully selected from more than one million sites with the consideration of representativeness, integrity, size, and availability. CWT100g is now freely available to social circles and has been designated as a test collection of SEWM2004[SEWM,2004] for evaluating all participant IR systems.

The archived data of the Web InfoMall, the relevant tools and models constructed from it benefit to social science research. For example, collaborating with Guanghua school of management of Peking University, we tapped into the archive for the economic monitoring, the marketing supply and demand analysis, and the human resource status inspect.

Ongoing research

Web resources include not only pages, but many other types of media files. Besides the collection of web pages, we have been collecting these kinds of resource since 2003. We build a separate system, named CDAL (Chinese Digital Assets Library)[Chen, et al.,2004], to reposit these resources and classify them to more than 400 categories. These resources have been capable of accessing based on their category information. We are on the way to combine the information process of web pages with these multimedia resources.

The newly statistic in June 2004 show the scale of the multimedia web resources in CDAL is 5 terabyte, including more than 3 thousand videos, about 40 thousand audios, over 80 thousand pictures, and more than 10 thousand digital books, etc.

People usually question that collecting all data without distinguishing good or bad will impair the usefulness of the whole data due to the chaos of the web information. At the present, we consider the Web as a kind of mineral resource, and the Web InfoMall an information repository waiting to be exploited. The content-based selection to extract pith and remove draffy can be the successive objectives for the Web InfoMall.

Related work

The present work benefits from past research in the areas of web archiving, search engines, and information retrieval technologies.

Many countries, e.g.,[Kahle,1997],[Torsteinn Hallgrimsson and Bang,2003], [Zabicka,2003], also have launched the projects of Web archiving. However, as web archiving projects, they pay more attention to the storage architecture and have less attention on how to collect and use web pages efficiently.

As the the precursor of the Web InfoMall 2.0, version 1.0[Li, et al.,2004] gave us many design and running experience. Through one year's online running, we

realized the importance of a standard format for perennial storage, and the requirement of indexing each batch of pages independently.

Search engines, e.g., [Google,2004], [Tianwang,2004], also archive web pages for retrieval and snapshot services. However, they only keep the latest copy of a url and ignore the older versions.

Many information retrieval models and tools have been brought forward and used[Baeza-Yates and Ribeiro-Neto,1999],[Witten, et al.,1994].

In this paper, we combine the advantages of the above three aspects to construct the China Web InfoMall 2.0.

Conclusion

In this paper, we present the implementation strategies of the China web archive system, the Web InfoMall 2.0, on three aspects, i.e, the approach to get the resources from the Web, the strategy to store the web pages, and the service model to make the history web pages available freely to the public.

With the accumulation on data and the successive development on related application tools, the Web InfoMall has become the platform for both social research and scientific research. We showed several macro parameters like the average life of web pages and the general update interval of China web sites. Also, our experiments on the web social response to significant events demonstrated the relationship between the virtual and the physical society.

We also briefly discuss our ongoing work on other forms of web resources besides web pages. With this part of work, the Web are preserved integrallty.

Different with other web archive projects, the China Web InfoMall comes from the research of computer science instead of the information management. We employ our experience on large scale distributing system and mass data processing to design and implement the project. The scalability, the access efficiency and the robust of the system are proved to be satisfied in the past three years' practice. But we are lack of the knowledge on the information character itself, as a result, the cataloging and the content-base information structure research have not been pushed.

With the recognition importance on historical web information, we believe that more people will involve in the preservation actions, especially in the kinds of special topic collection which are concerned by a number of people. In those applications, the methods and technologies of the Web InfoMall may be used for reference.

All archived data in the Web InfoMall are freely to the public, welcome to visit the site, http://www.infomall.cn.

References

[Baeza-Yates and Ribeiro-Neto,1999] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: Addison-Wesley-Longman, 1999.

[Chen, et al.,2004] C. Chen, H. Yan, and X. Li, "CDAL: A Scalable Scheme for Digital Resource Reorganization," presented at The 3rd International Conference on Web-Based Learning (ICWL2004), Tsinghua University, Beijing, China, 2004.

[CNNIC,2004] China Internet Network Information Center. http://www.cnnic.net.cn.

[CWT,2004] Chinese Web test collection. http://net.pku.edu.cn/~webg/cwt/.

[Google,2004] Google Search Engine. http://www.google.com.

[Kahle,1997] B. Kahle, "Preserving the Internet," Scientific American, vol. 276(3), pp. 82-83, 1997. (http://www.hackvan.com/pub/stig/articles/trusted-systems/0397kahle.html).

[Li,2003] X. Li, "An estimate on Chinese historical static web pages," Transactions of Peking University (in Chinese), vol. 39, pp. 394-398, 2003. (http://162.105.80.88/crazysite/home/report/upload/1540990286.doc).

[Li, et al.,2003] X. Li, H. F. Yan, and J. J. Zhu, "A Model for Collecting and Processing Topical Information in the Web and Its Application," Journal of Computer Research and Development (in Chinese), vol. 40, pp. 1667-1671, 2003..

[Li, et al.,2004] X. Li, Z. Xie, L. Sun, and H. F. Yan, "Web InfoMall: the Concept and Design for a Mass Web Pages Storage System," accepted by the special book of China national key research grant, 2004.

[SEWM,2004] SEWM, "The second Search Engine and Web Mining Symposium," 2004. (http://www.scut.edu.cn/sewm2004).

[Tianwang,2004] Tianwang Search Engine. http://e.pku.edu.cn.

[Torsteinn Hallgrimsson and Bang,2003] Torsteinn Hallgrimsson and S. Bang, "Nordic Web Archive," presented at ECDL2003 (7th European Conference on Research and Advanced Technologies for Digital Libraries), Trondheim, Norway, 2003. (http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=hallgrimsson).

[Witten, et al.,1994] I. H. Witten, A. Moffat, and T. C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. New York, NY: Van Nostrand Reinhold, 1994.

[Yan, et al.,2001b] H. Yan, J. Wang, and X. Li, "A dynamically reconfigurable model for a distributed Web crawling system," presented at International Conference on Computer Networks and Mobile Computing, Beijing, China, 2001b.

[Yan,2003] A Snapshot and Analysis of Guangdong Province Web Information in April of 2003 (in Chinese). http://www.infomall.cn/gd/.

[Yan, et al.,2001a] H. F. Yan, J. Y. Wang, X. M. Li, and L. Guo, "Architectural design and evaluation of an efficient Web-crawling system," presented at Proceedings of 15th International Parallel and Distributed Processing Symposium, San Francisco, California, USA, 2001a. (Also published in Journal of Systems and Software, vol. 60, pp. 185-193, Feb 15, 2002.).

[Zabicka,2003] P. Zabicka, "Archiving the Czech Web: Issues and Challenges," presented at ECDL2003 (7th European Conference on Research and Advanced Technologies for Digital Libraries), Trondheim, Norway, 2003. (http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=zabicka).