Experiments of Collecting WWW Information using Distributed WWW Robots

Hayato YAMANA
Electrotechnical Laboratory
www.etl.go.jp/~yamana/

Kent TAMURA
Tokyo Research Laboratory
IBM Corporation

Hiroyuki KAWANO
Kyoto University
www.kuamp.kyoto-u.ac.jp/~kawano/

Satoshi KAMEI
Kyoto University
www.kuamp.kyoto-u.ac.jp/~skame/

Masanori HARADA
The University of Tokyo
www.graco.c.u-tokyo.ac.jp/~harada/

Hideki NISHIMURA
Corporate R&D Group
Sharp Corporation

Isao ASAI
Osaka Prefecture University
www.bekkoame.or.jp/~asaisan/

Hiroyuki KUSUMOTO
Keio University
www.sfc.wide.ad.jp/~kusumoto/

Yoichi SHINODA
School of Information Science
JAIST

Yoichi MURAOKA
Waseda University
www.info.waseda.ac.jp/muraoka/

Abstract This paper presents the experiments of collecting the documents on the WWW using distributed WWW robots. We propose distributed WWW robots to collect the documents quickly. Our final goal is to collect all of the documents on the WWW in Japan within one day. Currently, eight distributed WWW Robots are running in Japan. The experimental results show that we are able to gain 5.8 to 9.7 times speedup when four distributed WWW robots are placed at different places in comparison with when only one WWW robot is used.

1 Introduction

The world-wide web, in short the web, is a large distributed digital information space. It is the most popular internet service and is now indispensable for us. The web itself, however, has no protocols for searching the web documents. Thus, we need to develop an information retrieval system called *a WWW search engine*.

The WWW search engines can be divided by their databases into directory-types and robot-types. *Yahoo*![9] is one of the popular directory-type search engines. However, the number of the web documents indexed in directory-type search engines are usually smaller than that in robot-type search engines, because the web documents of directory-type search engines are collected and classified into categories by having a full-time staff. In contrast with this, the web documents of robot-type search engines are collected automatically by the software called "*a WWW Robot*", which goes around the internet and collect the web documents. Therefore, the robot-type search engines such as AltaVista[1] and HotBot[2] have many indexed web documents.

Currently, there exist more than 160 WWW robots and more than twenty WWW robots go around the same WWW servers. Many WWW robots result in large traffic both on the internet and on the WWW servers. Besides that, it becomes difficult to collect all of the web documents in a short period, because the number of WWW servers increases exponentially. For example, the robot used with Senrigan[7] takes about one month to collect half a million documents that are less than 10% of all the web documents in Japan.

To solve these problems, we propose distributed WWW robots to collect the web documents quickly. Currently, eight distributed WWW Robots are running in Japan.

In Section 2, we propose the distributed WWW robots. This is followed by the experimental results in Section 3. Section 4 indicates future work and our conclusions.

2 Distributed WWW Robots

The distributed WWW robots consist of one "Public Robot Server Manager (**PRSM**)" and several "Public Robot Servers (**PRS**)" as shown in Fig.1. The PRSM controls all of the PRSs that collect the web documents.

First, the PRSM sends the control signal to PRSs to measure the data transfer speed $\{t(S_i, R_j)\}$ between WWW servers $\{S_i\}$ and PRSs $\{R_j\}$. After measuring $\{t(S_i, R_j)\}$, PRSs send back them to the PRSM. Then, the PRSM decides the assignment of the WWW servers for every PRS. The assignment is led by the algorithm described in Section **2.1** ,which divides PRSs into exclusive sets of PRSs. After receiving the assignment, all PRSs initiate to collect the web documents. When a PRS finds a new WWW server, the PRS reports it to the PRSM that decides its assignment in the same way. [3]

Secondly, the collected documents in each PRS are distributed to "Search Service Servers (**SSS**)" to be indexed and to be prepared for search services. At this time, the collected web documents in a PRS are compressed to shorten the distribution time. We assume that we may have many SSSs to provide search services.

Permission to make digital/hard copy of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or fee. SIGIR'98, Melbourne, Australia © 1998 ACM 1-58113-015-5 8/98 $5.00.

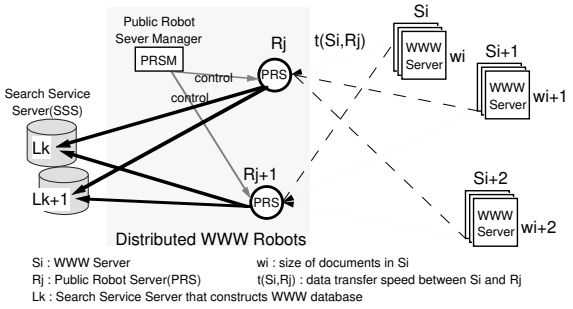

Figure 1: The Distributed WWW Robots

Scheduling Algorithm The scheduling procedure has the following two steps.

① Step 1: Since the bottleneck to speedup the collection of the web documents is the data transfer speed between WWW servers and PRSs, a WWW server S_i is assigned to the PRS R_j that has the highest data transfer speed $t(S_i, R_j)$.

② Step 2: After applying the Step 1, we equalize the execution time of PRSs to reduce the total execution time. The execution time of a PRS is defined as the sum of the time to collect the web documents and the time to distribute them to all SSSs.

Since the maximum execution time of PRSs bounds the total execution time, the total execution time is defined as the following function, where $w_i[byte]$ is the amount of documents in WWW server S_i, $t(S_i, R_j)$ $[s/byte]$ is the data transfer speed between WWW server S_i and PRS R_j, K is the document compression ratio to distribute the web documents form PRSs to SSSs, $t(R_j, L_k)$ $[s/byte]$ is the data transfer speed between PRS R_j and SSS L_k, and V_j is the set of WWW servers $\{S_i\}$ assigned to PRS R_j.

$$\max_{j=1,\ldots,m} \left(\sum_{S_i \in \{V_j\}} \{w_i \times t(S_i, R_j) + K \times w_i \times \sum_{k=1}^{l} t(R_j, L_k)\} \right) \quad (1)$$

The term $\{w_i \times t(S_i, R_j)\}$ shows the time to collect the web documents in S_i by PRS R_j. Besides, the term $\{K \times w_i \times t(R_j, L_k)\}$ shows the time to distribute the collected web documents from PRS R_j to SSS L_k with compression ratio K.

To minimize the (1), we adopt the *approximate algorithm* proposed in [8]. Such optimizations are known as *a multiprocessor scheduling problem* [4], which is one of famous NP-complete problems. The approximate algorithm gives near optimized scheduling.

3 Experimental Results

We calculate the total execution time by using the (1) on the assumption that we have four PRSs and four SSSs at Univ. of Tokyo, Sharp Corp., Waseda Univ., and ETL.

We measured the maximum data transfer speed of *ping command* to lead the $t(S_i, R_j)$, because we have found that the collection speed of the web documents is proportional to the speed of *ping*.

We also assumed each S_i as a set of WWW servers that have *jp* domain and have same domains in second and third level – e.g., $*.waseda.ac.jp$. The amount of documents in S_i is led by the databases that we have already collected by RCAAU – mondou[6], Senrigan[7], and ODIN[5]. The total document size is 669[MB].

The number of total S_i used in the preliminary evaluation is 6,980 which is about half of all the S_i in Japan. The compression ratio K is assumed as 0.2 that is the average compression ratio using "$tar + gzip$".

Fig.2 shows the results. As shown in Fig.2, we archive 5.8 to 9.7 times speedup when we use four PRSs in comparison with when we use one PRS.

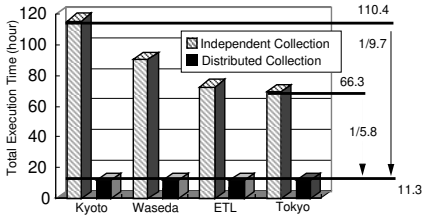

Figure 2: Experimental Results

Status The system code of both PRSM and PRS is mostly written in Java with some C. Eight PRSs are running at eight different places. We have already found 13,320 domains that have jp domain. Now, all the PRSs are still at the stage to measure the data transfer speed. We found that the data transfer speed between PRSs and a WWW server varies from 1.1 to 10.9 times. The average is 2.8. It leads the expectations that we could archive the speedup at most *(2.8× the number of PRSs)* times when the distributed WWW robots are used.

4 Conclusions

We propose the distributed WWW robots and show the current status that eight distributed WWW robots are running in Japan.

The experimental results show that we are able to gain 5.8 to 9.7 times speedup to collect the web documents when four distributed WWW robots are placed at different places.

Our final goal is to collect all the web documents in Japan within one day. We continue the experiments to archive the final goal.

References

[1] AltaVista, `http://www.altavista.digital.com/`

[2] HotBot, `http://www.hotbot.com/`

[3] Kent TAMURA, *Researches on Didtributed WWW Robots*, M.S. Thesis, Waseda University, 1997.

[4] M.R.Gray and D.S.Johnson, *A Guide to the Theory of NP-Completeness*, The Mathematical Sciences, San Francisco: Freeman, 1979.

[5] ODIN, `http://kichijiro.c.u-tokyo.ac.jp/odin/`

[6] RCAAU – mondou –, `http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/`

[7] Senrigan, `http://senrigan.ascii.co.jp/`

[8] S.Kamei,H.Kawano and T.Hasegawa, *Effectiveness of Cooperative Resource Collecting Robots for Web Search Engines*, Proc. of Pacrim'97, Vol.1, pp.410–413, 1997.

[9] Yahoo!, `http://www.yahoo.com/`