Distributed Location Aware Web Crawling

Odysseas Papapetrou
cspapap@cs.ucy.ac.cy

George Samaras
cssamara@cs.ucy.ac.cy

Department of Computer Science, University of Cyprus
75 Kallipoleos str., P.O. Box 20537, Nicosia, Cyprus

ABSTRACT

Distributed crawling has shown that it can overcome important limitations of the today's crawling paradigm. However, the optimal benefits of this approach are usually limited to the sites hosting the crawler. In this work, we propose a location-aware method, called IPMicra, that utilizes an IP address hierarchy, and allows crawling of links in a near optimal location aware manner.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Distributed Systems

General Terms

Performance

Keywords

location aware web crawling, distributed web crawling

1. INTRODUCTION

Due to the changing rate and the size of the web, the current crawling systems appear inadequate to keep a significant web mirror for searching purposes. Realizing these limitations, we recently proposed UCYMicra [1, 2], a distributed web crawling system that utilizes mobile crawlers which are send to locally crawl the collaborating web servers. The migrating crawlers crawl (the web-pages hosted in the web-server or the LAN of the web-server), process, and transmit the results back to the search engine, and finally monitor the local web pages for changes. However, this approach performed optimized crawling only of the collaborating sites. The migrating crawlers were unable to optimally crawl non-local URLs. In this work, we suggest IPMicra that facilitates crawling of each URL from the most near crawler (nearness in terms of network latency) without creating excessive load to the Internet infrastructure. We use data from the four Regional Internet Registries (RIRs) to build a hierarchical clustering of IP addresses, which assists us to perform an efficient URL delegation to the migrating crawlers.

2. REGIONAL INTERNET REGISTRIES

Regional Internet Registries are non-profit organizations that are delegated the task of handling IP addresses to the clients. Currently, there are four regional Internet Registries covering in the world: APNIC, ARIN, LACNIC, and RIPE NCC. All the sub-networks (i.e. the companies' and the universities' sub-networks) are registered in their regional registries (through their Local Internet Registries) with their IP address ranges. Via the RIRs a hierarchy of IP ranges can be created. Consider the IP range starting from the complete range of IP addresses (from 0.0.0.0 to 255.255.255.255). The IP addresses are delegated to RIRs in large address blocks, which are then sub-divided to the LIRs (Local Internet Registries); lastly they are sub-divided to organizations, as IP ranges, called subnets.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

3. LOCATION AWARE WEB-CRAWLING

Location aware web crawling is distributed web crawling that facilitates the delegation of the web pages to the 'nearest' crawler (i.e. the crawler that would download the page the fastest). **Nearness** and **locality** are always in terms of network distance (latency) and not in terms of physical (geographical) distance. In order to find the nearest crawler to a web server we use **probing**. Experiments showed that the traditional ICMP-ping tool, or the time that takes for a HTTP/HEAD request to be completed, are very suitable for probing. In the majority of our experiments, the crawler with the smallest probing time was the one that could download the web page the fastest. Thus, the migrating crawler having the smallest probing result to a web server is possibly the crawler most near to that web server. Evaluating location **aware** web crawling, and comparing it with distributed location **unaware** web crawling (e.g. UCYMicra) was actually simple. UCYMicra was enhanced and, via probing, the URLs were optimally delegated to the available migrating crawlers. More specifically, each URL was probed from all the crawlers, and then delegated to the 'nearest' one. Location aware web crawling outperformed its opponent UCYMicra, which delegated the various URL randomly, by requiring **one order of magnitude less time (1/10th)** to download the same set of pages, with the same set of migrating crawlers and under approximately the same network load.

4. THE IPMICRA SYSTEM

While location-aware web crawling has impressive results, it requires each URL to be probed by all the migrating crawlers. IPMicra specifically targets the reduction of the required probes by delegating a URL to the nearest crawler. We designed and built an efficient self-maintaining algorithm for domain delegation (not just a URL) with minimal network overhead by utilizing information collected from the Regional Internet Registries (RIRs).

4.1 The IP-address Hierarchy and Crawlers

The basic idea is the organizing of the IP addresses, and subsequently the URLs, in a hierarchical fashion. We use the `WHOIS` data collected from the RIRs to build and maintain a hierarchy with all the IP ranges (IP subnets) currently assigned to organizations (e.g., see figure 1). The data, apart from the IP subnets, contains the company that registers each subnet. Our experience shows that the expected maximum height of our hierarchy is 8. The required

time for building the hierarchy is small, and it can be easily loaded in main memory in any average system. While the IP addresses hierarchy does not remain constant over time, we found out that it is sufficient to rebuild it every three months, and easy populate it with the old hierarchy's data.

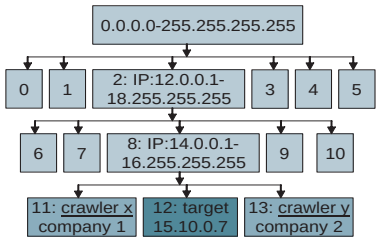

Figure 1: A sample IP hierarchy. Subnets 11 and 13 belong to company 1 and company 2 respectively. Such a hierarchy can be built from the bulk `WHOIS` information from RIRs

Once the IP hierarchy is built, the migrating crawlers are sent to affiliate organizations. Since the IP address of the machine that will host the crawler is known, we can immediately assign that subnet to the new crawler. In this way the various crawlers populate the hierarchy. The hierarchy can now be used to efficiently find the nearest crawler for every new URL, utilizing only a small number of probes. We stop probing as soon as we find a crawler that satisfies a threshold, called **probing threshold**. Probing threshold is the maximum acceptable probing time from a crawler to a page and it is set by the search engine's administrator depending on the required system accuracy.

4.2 The URL Delegation Procedure

Based on the assumption that the sub-networks belonging to the same company or organization are logically (in terms of network distance) in the same area, we use the organization's name to delegate the different domains to the migrating crawlers. In fact, instead of delegating URLs to the distributed crawlers, we delegate subnets. We first find the **smallest** subnet from the IP hierarchy that includes the IP of the new URL, and check if that subnet is already delegated to a crawler. If so, the URL is delegated to this migrating crawler. If not, we check whether there is another subnet that belongs to the same company and is already delegated to a migrating crawler (or more). If such a subnet exist, the new URL, and subsequently, the owning subnet, is delegated to this crawler. If there are subnets of the same company delegated to multiple crawlers then the new subnet is probed from these crawlers and delegated to the fastest. In fact, we stop as soon as we find a crawler that satisfies the probing threshold. Only if this search is unsuccessful, we probe the subnet with the migrating crawlers, in order to find the best one to take it over. We navigate the IP-address hierarchy bottom up, each time trying to find the most suitable crawler to take the subnet. We first discover the parent subnet and find all the subnets included in the parent subnet. Then, for all the sibling subnets that are already delegated, we sequentially ask their migrating crawlers, and the migrating crawlers of their children subnets to probe the target subnet, and if any of them has probing time less than a specific threshold (probing threshold), we delegate the target subnet to that crawler. If no probing satisfies the threshold, our search continues to higher levels of the subnets tree. In the case that none of the crawlers satisfies the probing threshold, the subnet is delegated to the crawler with the lower probing result.

4.3 An Outline of the IPMicra system

The IPMicra system is architecturally divided in the same three subsystems that were introduced in the original UCYMicra: (a) the public search engine, (b) the coordinator subsystem, and (c) the mobile agents subsystem. Only the public search engine remains unchanged. The coordinator subsystem is enhanced for building the IP hierarchy tree and coordinating the delegation of the subnets, and the migrating crawlers are enhanced for probing the sites and reporting the results back to the coordinator.

4.4 Performance and Evaluation

order to evaluate IPMicra, we only needed to test its performance in estimating the optimal (the most near) crawler for each URL, and the required probes for each delegation (since we already evaluated location aware web crawling). The experiment was conducted with 12 affiliated organizations hosting the migrating crawlers, and 1000 test sites. From the 1000 test sites, the first 650 were used in order to initialize the hierarchy, and the rest 350 were used for evaluation purposes. The experiment had very encouraging results. With a probing threshold set to 50msec, the delegation procedure resulted to 261 optimal delegations (average) from the 350 (75%), with only 1048 probes, while the brute force location aware delegation would require 12*350=4200 probes. Furthermore, the 89 sub-optimal delegations were having an average probing difference from the optimal of less than 13, which means that they were very near to the optimal. Moreover, while for practical reasons the experiment did not include many testing URLs (only 350), it was clear from the results that as the hierarchy was getting 'trained', the average number of the required probes for the delegation of a URL was decreasing. For example, while the average number of probes required for the delegation of all the URLs was 2.99 probes per URL, the average number of probes for delegation of the last 50 URLs was 2.66 probes per URL. For the first 300 URLs the average was 3.05 probes per URL.

5. CONCLUSIONS

In this work, we proposed IPMicra, an extension of UCYMicra, that allows, based on the notion of 'nearness', crawling of links in a near optimal location aware manner. The motivating power behind IPMicra is an IP address hierarchy tree, which is build using information from the four Regional Internet Registries. This hierarchy is used to delegate the web sites to near migrating crawlers in order to take advantage of the lower network latency for faster crawling. To our knowledge IPMicra is the first location aware distributed web crawler, and can offer an efficient and generic solution to today's web indexing problem. The framework can be easily applied to existing commercial approaches, like the Google Search Appliance or Grub.

6. ACKNOWLEDGEMENTS

This work is partially funded by the Information Society Technologies programme of the European Commission under the IST-2001-32645 DBGlobe project.

7. REFERENCES

[1] Odysseas Papapetrou, Stavros Papastavrou, and George Samaras. Distributed indexing of the web using migrating crawlers. In *Proceedings of the Twelfth International World Wide Web Conference (WWW)*, 2003.

[2] Odysseas Papapetrou, Stavros Papastavrou, and George Samaras. Ucymicra: Distributed indexing of the web using migrating crawlers. In *Proceedings of the 7th East-European Conference on Advanced Databases and Information Systems, Dresden, Germany*, 2003.