Title: A Proxy Server Supporting Document Categorization

Authors: Cheng-Kang Wen, Shang-Rong Tsai, Feng-Shiuh Song

Address: Department of Electrical Engineering
National Cheng Kung University
No.1 University Road
Tainan, Taiwan

Correspondence:
Cheng-Kang Wen

Room #92571
Department of Electrical Engineering
National Cheng Kung University
No.1 University Road
Tainan, Taiwan

Email: ckwen@mail.ncku.edu.tw
TEL: 886-06-275-7575 ext. 62431 # 771
FAX: 886-6-2748678

A Proxy Server Supporting Document Categorization

Abstract

The rapid popularity of the World Wide Web makes the Internet full of information. Users spend much time in searching for their favor web pages and waiting for slow response due to the crowded Internet. Web proxy server is used to reduce the network latency. It is convenient for users to access specific interested web information if document categorization can be done on proxy server. In this paper, we describe the design and implementation of a proxy system that collects web objects and categorizes them according to information subjects.

Keywords: WWW , Document categorization, Proxy server, Search Engine.

1. Motivations

The explosive growth of the World Wide Web makes the Internet full of information. This paper intends to improve the usage of web proxy on the Internet based on the following motivations.

- To provide active web proxy service. For current web proxy design[9], proxy caches web pages under user's demand. However, proxy manager can assign proxy server to look up many search engines in advance by giving some keywords as hints on the Internet and then proxy server keeps specific web pages in its cache storage from web sites. Users can then easily and quickly get information from the proxy with similar subjects as those pages the proxy collected earlier. This will potentially enhance proxy cache hit ratio a lot.
- To categorize web information. Document categorization can be achieved automatically when a proxy collects and stores web pages only on specific topics. Different proxies keep web pages on different topics. Users can access their favor web information with specific subjects by connecting to appropriate proxy server.
- To improve proxy cache hit ratio. Cache-hit ratio of the current web proxy is not too high due to wide range of topics. Higher hit ratio makes proxy servers more useful. Categorizing information can group those pages with same interest together on a web proxy. Users can then easily retrieve information with related topics from those special-purpose proxies.
- To reduce network latency. Users usually browse web pages via web proxy

on the Internet in order to reduce possible latency. With the increase of available information, users must spend much time in searching for their favor web pages and waiting for slow response due to the crowded Internet. Users may waste much time on getting useless information. Getting useless information not only costs lots of user's time, but also increases load of web servers that provide the information. Consequently, how to get information accurately and effectively are important issues in designing web software.

- Personalized web proxy. General-purpose web proxy can not provide information instantly that is focused on specific topics to individuals or small research group. However, users hardly set up a web proxy for themselves to have special proxy service. Because building and maintaining a web proxy are too difficult for most of users. A proxy server that offers some specific topics could meet users' expectation. For instance, a proxy keeps web information only for the topic of entertainment. Users who are interested in subjects concerning movie stars or singers can easily access related web pages by connecting to this proxy server. Thus, individual user uses the proxy as if he/she owns the proxy server.
- Public web courses storage. Web courses offer a good way to self-education. Especially when distance learning is more acceptable to people on the Internet. However, to keep copies of same web courses at personal computer costs lots of users' storage space and network bandwidth. On the other hand, users have to connect to the right proxy to access the web courses that are cached at general-purpose web proxies. A proxy that searches and stores web courses on particular topics, for instance, "Computer Network" and "Computer Graphics," offers higher cache hit ratio to those users who are interested in learning lessons with such subjects.

In this paper we design and implement a proxy to collect web objects according to their subjects. This proxy provides higher cache-hit ratio to users that are interested in some specific subjects. The rest of this paper is organized as following: section 2 depicts the system overview of our proxy design while section 3 describes the implementation of system architecture. Section 4 discusses some issues in current design, and section 5 concludes.

2. System Overview

2.1 Design Concepts

More and more information makes users hard to search those documents they need. There is no standard method to judge the content of web page except browsing

the page directly. This is the most serious problem. Within information all over the Internet, Usenet's News has the best mechanism for data classification. Why do web pages not have a stand mechanism for data classification like News? If web data can categorize according to subjects, users can only search the subjects that they are interested. It can save user's time to get useful documents quickly.

There are currently two methods for document categorization of web pages:

- Adding new tags to HTML[3] to indicate what category the web page belongs to.
- Using PICS [4] [5] rating system that W3C develops to define a rule of web pages categorization.

However, the two methods described above have difficulties in implementation. First, it will need all documents providers cooperate with one another. Second, to process so many existing web documents is a huge work. Therefore, the methods are hard to implement in current web environment.

In this paper, we propose another alternative. By searching and fetching web documents related to specific subjects in advance via web proxy, documents with similar topics can be collected on the proxy. Different proxies collect different groups of documents. Thus, documents categorized are done automatically over different web proxies. Users can access their favorite web documents with higher hit ratio by connecting to appropriate proxy.

This design is more useful if users who are in a small group have the same interested topics. For instance, members of same laboratory or employees of the same company usually access related information. The access patterns will raise proxy hit ratio apparently. The original idea is to provide a special proxy that collects some specific subjects of web objects. When users use this proxy to access the data the proxy collects, it can save network bandwidth and shorten user's waiting time. If users are interested in different subjects of data, they will use another proxy that collects the other subjects of data. Figure 1 is the system overview of our design.

How does our proxy collect web objects according to specific subjects without any modification to original HTML pages? The proxy server supporting document categorization will send queries with specific keywords to many search engines. The proxy will prefetch[2] those web pages according to those URLs returned from search engines as search results. Administrators of the proxy also can assign specific URLs as parameters for proxy to prefetch in advance. Thus, the proxy server will be full of information with some specific subjects. Users can then retrieve web objects with higher hit ratio with their favorite subjects by connecting to the proxy.

2.2 Design Goals

The proxy server we design in this paper can support:

- Document categorization by using proxy server and search engines.
- Specific information subjects gathering. To provide higher cache hit ratio, the proxy server will collect web objects according to subjects. If users are interested in some specific subjects of WWW data, they can access web objects with higher cache hit ratio by connecting to the proxy.
- Easily control of proxy server. Administrators can control the Proxy Manager from web browser. It provides friendly interface and remote control functions to web proxy.
- Multi-level prefetch. The result URLs from search engines can be prefetched in multi levels. We expect all documents concerning the subjects or keywords to be prefetched by proxy server. Users access those documents with similar subjects via the proxy will get higher hit ratio comparing to that connecting to original web servers directly. It saves much network bandwidth and shortens user's waiting time.
- Pages prefetch reservation. We may reserve some pages to browse. Proxy can be scheduled to retrieve some specific pages in advance when network traffic is not too busy. Users can then browse these reserved pages from proxy server without connecting to the original web servers.

3. System Architecture

3.1 System Components

As shown on figure 2, the system is consisted of the following components:

- Proxy Manager Interface:

  This component is implemented in Java applet to provide a friendly interface between the Proxy Manager component and proxy administrators. Administrators can set parameters to control the operations of Proxy Manager from remote web browser.

- Proxy Manager:

  This component is implemented in Java application that communicates with Proxy Manager Interface to get parameters and settings. It sends queries to many search engines or web servers through web proxy server.

- Web Proxy Server:

  A general proxy server caches web pages or results from search engines of some specific subjects.

The following steps show how the system operates:

1. To run the Proxy Manager up on the site that proxy server resides. However, Proxy Manager and proxy are not necessary running on the same machine.

2. Administrator uses web browser to connect to the proxy server that runs with Proxy Manager. Proxy Manager will return a web page including the Proxy Manager Interface.
3. Administrator gives keywords, prefetch levels, and special URLs as parameters to Proxy Manager via Proxy Manager Interface.
4. Proxy Manager gets the parameters from Proxy Manager Interface. It analyzes the parameters and sends queries to search engines or web servers through proxy server.
5. Proxy Manager gets the responses of search engines or web servers and analyzes the context of web pages to get hyper links. It sends queries to the links repeatedly until the prefetch levels parameter is satisfied.
6. When prefetch error occurs during network connection; Proxy Manager should automatically retry after a period.
7. When users want to access the documents of particular subjects, they can use this proxy server as his front-end proxy server.

3.2 System features

The proxy server we design to support document categorization has the following features:

- Using search engines to collect the data about some specific subjects. Administrator also can specify some well-known web servers about these subjects. Using proxy server to cache these data. Different data of subjects is collected on different proxy makes many web objects are classified approximately. If users are interested in some subjects of web data, they can access web objects with higher hit ratio by the use of this proxy. It saves much network bandwidth and user's waiting time.
- Users can specify some URLs to prefetch web pages in advance when the network is not busy. They can access the pages from proxy without from the original web servers. It can decrease user's long waiting time due to heavy load of network or web servers.
- When errors occur during connection, it can automatically retry after a period. Proxy Manager should retry several times.
- Using multithread to send queries to many search engines. It can improve performance of whole system.
- Administrators can control the Proxy Manager remotely using web browser and dynamically set the parameters of Proxy Manager without installing other applications.

3.3 Proxy Management

The main subject the proxy offers is depended on the decision of proxy manager. Proxy manager maintains the proxy via the Proxy Manager Interface that is a Java applet downloaded from proxy site. As shown in figure 3, proxy manager selects those search engines that queries messages will be sent to from search engines listing. Then proxy manager gives keywords as subject. Proxy manager can also assign prefetch levels to decide how depth a web tree should be dug.

As shown in figure 3 the Proxy Manager Interface provides ways for proxy administrator to decide what context the proxy should keep. Proxy administrator gives some keywords and chooses some search engines to search for the keywords. Once the Interface gets URL results from search engines, the proxy will fetch those web pages according to parameters of prefect level and cache them. In addition, proxy administrator gives specific URL and makes proxy cache those web objects from the URL according to the prefecth level parameter.

3.4 Experience

We set up the proxy at our laboratory. The main research topics we are interested are Internet applications and distributed systems. We use SQUID as our proxy server. The proxy server runs on a Pentium 200 MMX PC with 32 MB memory and 2 GB disk drive installed. This PC runs FreeBSD 2.2.5. Before running Proxy Manager, the average hit ratio of original proxy server is 40%~45%. As the result of running Proxy Manager, hit ration is raised up to 70%~75%.

4. Discussions

Although the proxy system we design in this paper provides higher cache hit ration to those users who are interested in specific subjects, there still are some potential issues we might meet.

How exact web pages can we get from the proxy system? It depends on the subjects or keywords given to the proxy system. This means more hints to the X-Proxy will get more information that is useful from search engines. In additions, searching more than single one “search engine” can get results that are more useful.

Since the proxy server software we use is designed for general purpose, its cache replacement policy is not suitable to our requirement completely. The cache hit ratio will go down when users use the proxy to access web information not related to dedicated subjects on the proxy. The only solution is to rewrite proxy software to suit our system.

5. Conclusions

In this paper, we propose the X-Proxy system to collect web objects with specific subjects by searching search engines. Users can access web objects with higher cache hit ratio by connecting to the proxy system to save much network bandwidth and reduce the latency. The X-Proxy system can also fetch specific URLs that proxy administrator specifies.

The methods of document categorization we use are not strictly defined. It is necessary in our future work to design a well-defined specification of data classification for the Internet. When the data all over the WWW be classified, it is easy to search for what we need.

References:

1. Azer Bestavros, Robert L. Carter, Mark E. and Crovella, "Application Level Document Caching in the Internet", ftp://ca-ftp.bu.edu/techreports/95-002-web-client-caching-ps.z, 1995
2. Venkata n. Padmanabhan, and Jeffrey C. Mogul, "Using Predictive Prefetching to Improve World-Wide Web Latency", ACM SIGCOMM Computer Communication Review p22-36
3. "A Dictionary of HTML META Tags", http://vancouver-webpages.com/META/metatags.detail.html, 1998
4. Paul Resnick, "Filtering Information on the Internet", http://www.sciam.com/0397issue/0397resnick.html, 1997
5. Paul Resnick, James Miller, "PICS: Internet Access Controls Without Censorship", http://www.w3.org/PICS/iacwcv2.htm, 1997
6. David P. Habib, and Robert L. Balliot, "How to Search the World Wide Web: A Tutorial and Guide for Beginners", http://www.ultranet.com/~egrlib/totur.htm, 1998
7. Terry A.Gray, "How to Search the Web", http://daphne.palomar.edu/TGSEARCH/, 1997
8. Jian Liu, "Guide to Meta-Search Engines", http://www.indiana.edu/~librcsd/search/meta.html, 1998
9. "Squid Internet Object Cache," http://squid.nlanr.net.

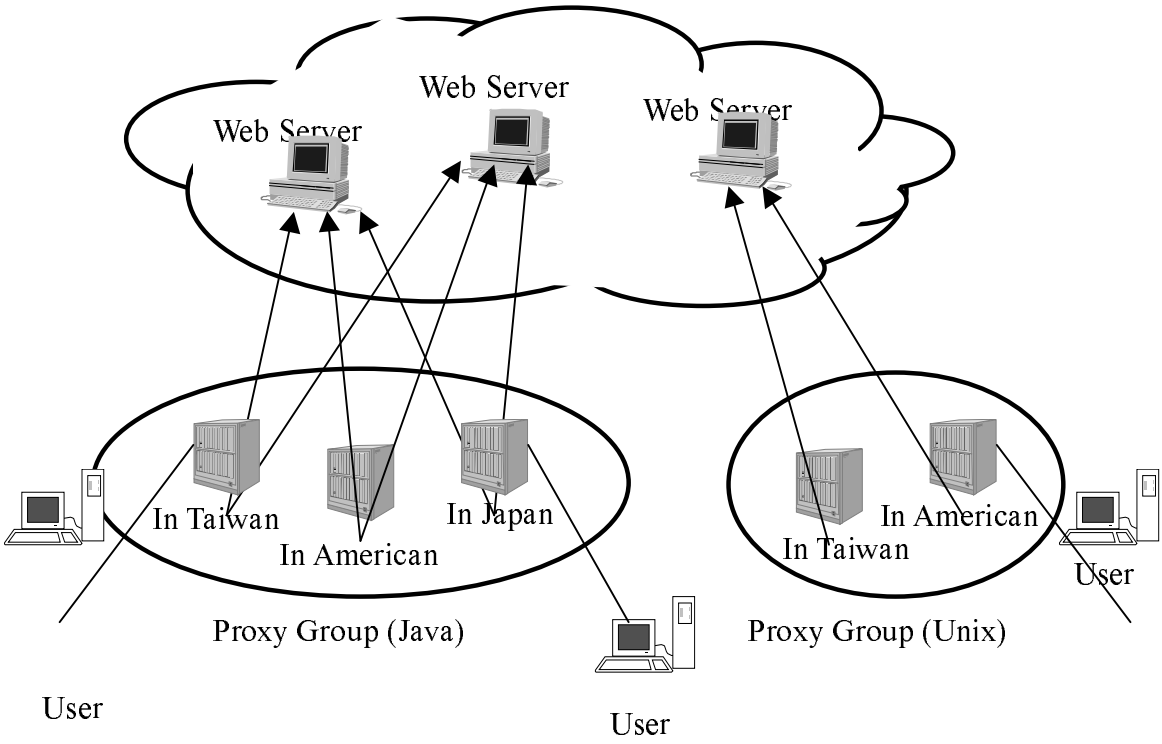

Figure 1: System Overview

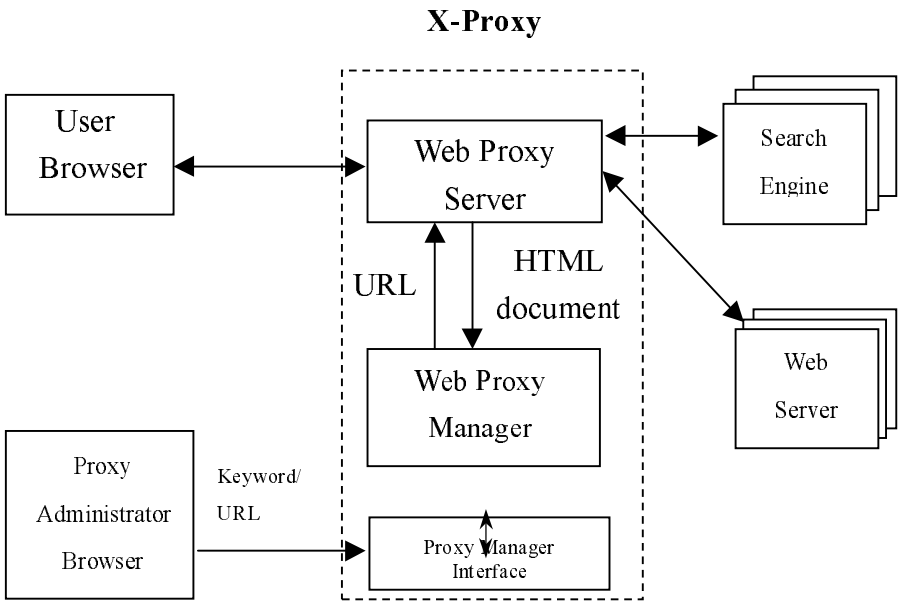

Figure 2: System Architecture

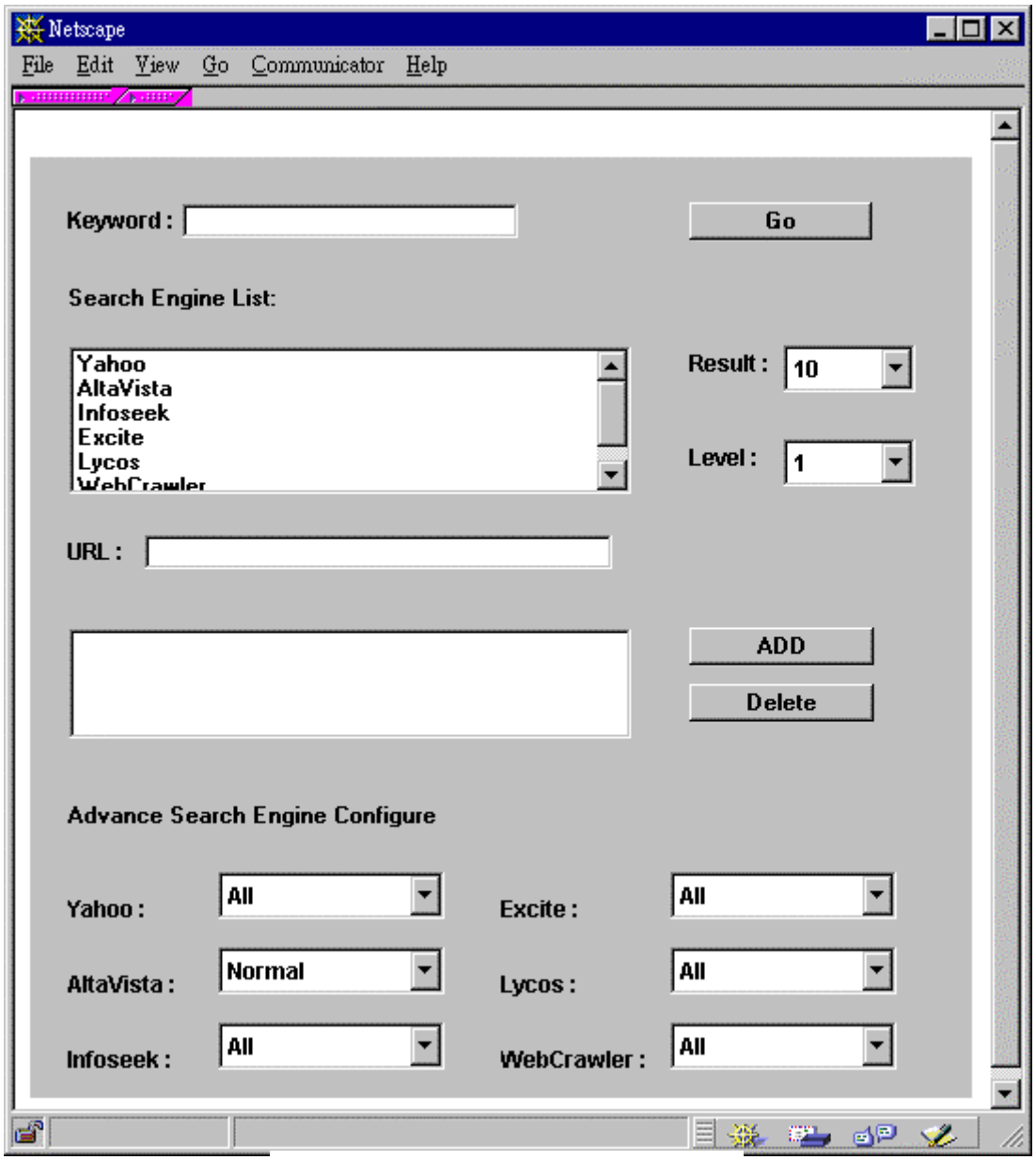

Figure 3: Proxy Manager Interface