December 3, 2002

Nippon Telegraph and Telephone Corporation
NTT-X

NTT and NTT-X Begin Verification Test on a "New-Information Search Engine" Covering all Web Pages in Japan

--Enables Retrieval of Information Updated as Recently as 15 Minutes Ago using Ultra-high-speed Information Gathering Technology--

Nippon Telegraph and Telephone Corporation (NTT; Head Office: Chiyoda-ku, Tokyo; President: Norio Wada) in cooperation with NTT-X (Head Office: Chiyoda-ku, Tokyo; President: Takao Nakajima) will begin a verification test on December 4, 2002 on a "New-Information Search Engine" developed by NTT Cyber Solutions Laboratories for application to "Next-generation Internet Search Services." These tests will take place on the "goo" (*1) Internet portal and will run for about four months.

This search engine covers the approximately 80 million Web pages in Japan and enables retrieval of information that has been placed on a Web server as recently as 15 minutes ago. It can be applied to various types of news sources that are continuously being updated as well as to new-product information, sports bulletins, event information, etc.

○Purpose

Web pages on the Internet are being updated with large volumes of new information day-by-day if not hour-by-hour. The fact is, however, that the contents of Web pages that can be searched by conventional Internet search services corresponds to information that is from several days to several weeks old. In addition, the several search services that can find updated information target a limited number of sites. In fact, the "Up-to-date News Articles Experimental Search Service" (*2) provided by NTT and NTT-X on "goo" since August 7 of this year targets only specific sites. This is because of the time required for converting all information on the Internet into searchable form at the performance levels of current commercial search engines. Solving the conflict between target-data update frequency and the scale (quantity) of pages to be searched has been a major problem.

The continuous-connection environment is helping to make the Internet a widely used means of collecting information, and there is much anticipation for "Next-generation Internet Search Services" that can provide updated information even when targeting the entire Internet that continues to expand rapidly.

Against this background, NTT has been working on the development of a "New-Information Search Engine" with the aim of providing an information search service that can provide the latest updates of constantly desired information without limiting target data. This search engine is based on "ultra-high-speed information gathering technology" that can collect all Web pages in Japan once or more per day. The "New-Information Search Engine" developed in Japan is viewed as a genuine contender to the mainstream search engines from overseas that have been used up to now by search sites within Japan.

○Test Overview and Objectives

To demonstrate the effectiveness and usefulness of the "New-Information Search Engine" developed by NTT as an Internet search service for the broadband era, a link to "Test Service for New Web Searching" will be pasted on the top page of the "goo" Internet portal operated by NTT-X and on the search-results page of this portal s "Web Page Search Service." The development of this new search engine will be announced to all "goo" users in this way.

NTT regards its "New-Information Search Engine" as a next-generation search engine and. wishes to evaluate its performance in an actual environment through this test. At the same time, NTT-X would like to examine and evaluate user needs to make "goo" a more extensive and powerful portal and to determine what type of Internet search services are needed by Internet users of the broadband era.

○Technological Features

This "New-Information Search Engine" uses the following technologies to collect updated information in a high-speed and efficient manner. These technologies make it possible to collect more than 100-million pages of information in one day and convert them for searching.

1. Ultra-multiplexed information-gathering control technology using Web-space automatic learning(attachment 1)
This technology performs efficient control of multiple information-gathering robots (crawlers) by learning the structure of Web space consisting, for example, of virtual domains (*3) and mirror servers (*4), and achieves intelligent multiplexed information-gathering control that prevents duplicate accessing of the same domain and IP addresses (*5). The speed at which this technology can gather information is about twice that of conventional technologies and it can be scaled up to cover several hundred million pages.

2. Updated-page learning and gathering control technology (attachment 2)
This technology controls the learning and gathering of updated pages. It can determine whether the main body of a Web page has been updated and changes the index (*6) only for pages that have been updated.

To immediately reflect Web pages that have been collected at high speeds in search results, high-speed index updating must also be performed. This test uses the following technology as used in the "Up-to-date News Articles Experimental Search Service" in addition to the above technologies.

3. Real-time indexing technology with compression
This technology performs high-speed extraction of keywords from each collected page and updates the index in real time. It also compresses the index table to reduce the amount of transmitted data.

○Future Developments

With the goal of making portal sites all the more powerful in the broadband era of continuous connections, NTT plans to continue its efforts in developing even faster and more accurate Internet search services. NTT-X plans to use the data obtained in this test to aim at applying this engine to "goo."

Glossary

(*1) goo (http://www.goo.ne.jp/)
A portal site having about 13 million unique users (see note below). It provides "communities" having about 3 million members as well as news and other types of "content" all centered about various search services with the "Internet Search Service" being the most popular.

Note: Calculated on the basis of access ratings from Internet access-rating surveys conducted by the Nippon Research Center, Ltd..

(*2) Up-to-date News Articles Experimental Search Service
With the aim of evaluating and testing "real-time indexing technology with compression" for immediately reflecting collected Web pages in search results (developed by NTT Cyber Solutions Laboratories), this experimental search service has been undergoing a public trial on "goo" since August 7, 2002. The service targets mainly information that is continuously being updated from news-oriented Web sites it can perform fast and efficient gathering of new information.

(*3) Virtual domain
A virtual domain allocates multiple domain names on one physical server and provides multiple services simultaneously through different domain names. In the case of rental servers, for example, multiple users of one server can publicly appear as separate servers. Note that a "domain name" is an identifier assigned to a server on the Internet and can be thought of as an Internet address. Because IP addresses consisting of numeric strings are difficult for people to deal with, domain names are represented by a combination of alphabetical characters, numerals, and some symbols.

(*4) Mirror server
A mirror server has the same function as another server and is set up to reduce the load on that server caused by concentrated access. For example, when a certain server on the Internet provides highly popular content, concentrated access to that server may cause its processing ability to be exceeded forcing the server into an inaccessible state. Mirror servers are deployed to distribute load and prevent this phenomenon from occurring.

(*5) IP address(Internet protocol address)
An IP address is an identifying number allocated to each and every computer, server, etc. connected to an IP network such as the Internet.

(*6) Indexing
The process of indexing creates an index to enable high-speed searches to be performed. Specifically, it extracts words from collected Web pages to act as search keywords and creates a database that records the correspondence between each keyword and the Web pages that it appears in. A search engine can achieve high-speed search processing by referring to this index.

- (Attachment 1) Ultra-multiplexed Information-Gathering Control Technology using Web-space Automatic Learning
- (Attachment 2) Updated-Page Learning and Gathering Control Technology

For more information, please contact:

Nippon Telegraph and Telephone Corporation
NTT Cyber Communications Laboratory Group
Planning Division, PR Section: Mr. Yamashita and Mr. Hagino
Tel: 0468-59-2032
E-mail: ckoho@lab.ntt.co.jp

NTT-X
PR Section: Mr. Suzuki, Mr. Tabata, Mr. Kuriyama
Tel: 03-5224-5500
E-mail: pr@nttx.co.jp

NTT NEWS RELEASE