로봇에이전트를 이용한 인터넷 주요 통계산출 알고리즘 설계 및 구현

김 원, 진 용 옥, *송 관 호
경희대학교 전자공학과, *한국인터넷정보센터
전화 : 02-2186-4502 / 핸드폰 : 011-342-4802

The Algorithm Design and Implementation of the Internet Statistics System for using the Robot Agent

Weon Kim, Yong Ohk Chin, *Khwan Ho Song
Dept. of Electronic Engineering, Kyunghee University
*Korea Network Information Center
E-mail : wkim@nic.or.kr

Abstract

This thesis proposes the design method of intelligent robot agent system and deals with the implementation of the system which is able to produce key internet statistics. It is believed that the statistics lead to effective investment from internet industry on its development. The system consists of robot agent process module, statistics production module and management module, and has an algorithm that can produce periodically the number of domestic homepages, active domain using .kr or gTLD and intenet hosts. It provides the result of the implementation and performance of the system as well.

I. 서론

일반적으로 인터넷 통계는 크게 인터넷 이용자 수, 이용실태 현황 등의 이용자에 관한 통계 부문①과 홈페이지 개수, 호스트 개수 등의 인터넷 환경에 관한 통계 부문②으로 구분할 수 있다.

① 인터넷 이용자에 관한 통계는 국내의 경우 한국인터넷정보센터(KRNIC)의 인터넷 이용자 통계조사, 국외의 경우 GVU's WWW User Survey, American Internet User Survey, Nua Internet Survey 등을 통해 주기적으로 발표되고 있다.

② 인터넷 환경에 관한 통계는 국내외적으로 국가인터넷레지스트리(NIR : National Internet Registry)가 도메인 개수, 호스트 수 등에 관한 자료를 산출하여 간헐적으로 발표하는 경향이 있다.

본 고의 2장에서는 인터넷 환경 통계산출 현황과 문제점, 3장에서는 인터넷 환경에 관한 통계 즉, 홈페이지 개수, 활성도메인의 국내 보유개수, 국내 호스트 개수 등을 정기적으로 산출할 수 있는 로봇에이전트를 설계하고 구현한 결과에 대해 기술하고, 4장에서는 로봇에이전트에서 수집한 HTML DB를 이용하여 주요 통계를 산출할 수 있는 기법을 제시하고 구현하였으며, 마지막으로 5장에서는 시험 및 산출결과를 통하여 향후 추가적인 연구 대상을 제시하였다.

II. 인터넷 환경통계산출 현황과 문제점

본 장에서는 인터넷환경 통계 산출에 관한 현황과 문제점을 도출함으로써 본 논문에서 연구·제시하고자 하는 중요성을 기술한다.

① [홈페이지 개수 산출] 인터넷 환경에 관한 통계 중에서 홈페이지 수의 산출은 미국의 경우 전문검색업체인

Inktomi사에서 자사의 로봇을 통하여 수집한 HTML 문서를 분석하여 그 개수를 간헐적으로 산출하여 발표하고 있으나, 아래와 같이 국내의 경우 그러한 로봇에이전트와 수집한 HTML를 분석하기 위한 알고리즘 및 기법이 개발되어 있지 않으므로 현재 자료제공이 가능하지 않다.

② [활성도메인(.kr, gTLD)의 국내 보유개수 산출] 인터넷 환경에 관한 통계 중에서 국제도메인 개수의 경우 미국의 Network Solutions사에서 등록된 전세계의 도메인을 분석하여 간헐적으로 국가별 순위 통계만을 제공하고 있으나, 국가도메인(.kr)과 gTLD(.com, .net,. org 등)의 국제도메인중에서 웹, Ftp, E-Mail 등의 서비스를 하는 활성화 도메인(Active Domain)에 대한 산출 방법은 존재하고 있지 않다.

③ [호스트 개수 산출] 인터넷 환경에 관한 통계 중에서 호스트 수의 산출을 위하여 기존의 방법으로 주로 DDT (Domain Debug Tool) 프로그램을 이용하여 도메인 사용기관의 DNS(Domain Name Server)에 등록된 호스트 개수를 산출하여 왔다. 그러나 이러한 방법은 DDT 프로그램에 대한 보안강화 등의 문제점으로 인하여 현재 공신력 있는 자료 제공이 가능하지 않다.

Ⅲ. 로봇에이전트의 필수기능

본 장에서는 로봇에이전트 설계와 구현된 구조의 필수기능에 대해 기술한다. 인터넷 환경통계 산출 작업은 그 결과가 인터넷 환경 변화의 진행 정도를 정확하게 제시할 수 있어야 하므로 주기적으로 수행되어야 한다. 그러므로 인터넷 환경 통계 산출을 위한 로봇에이전트는 기본적으로 아래와 같은 필수기능이 구현되어야 한다. 그림 1.은 로봇 에이전트의 기본구성도로서 3가지의 필수기능이 구현되어야 한다.

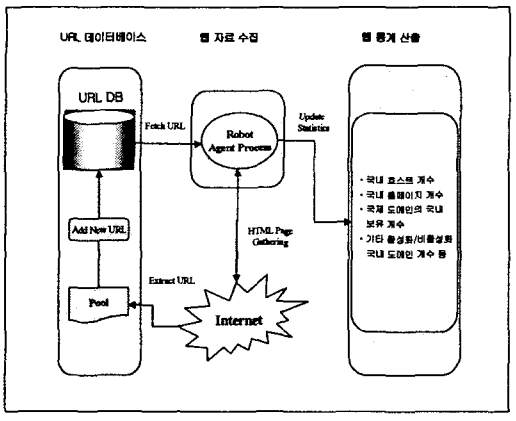

그림 1. 로봇에이전트 기본 구성도

① Incremental Search 기능 : 웹 문서 량의 증가에도 불구하고 주어진 기간 내에 신속하게 HTML 문서를 수집할 수 있는 능력

② Continuous Acting 기능 : 주기적으로 문서 자동 수집이 가능한 편리한 관리 기능

③ Fault Tolerant 기능 : 에러복구(Error Recovery)를 할 수 있는 능력

Ⅳ. 산출 기법(Methodology)

본 장에서는 로봇에이전트를 이용하여 산출하고자 하는 인터넷 환경통계 산출기법에 대해서 기술한다.

4.1 국내 홈페이지 개수

일반적으로 홈페이지 개수는 www.nic.or.kr과 같은 URL로 홈페이지 서비스가 되는 경우와 웹서버에 계정이 있는 이용자가 그 계정과 같은 이름의 디렉토리를 만들어 홈페이지를 만들어 서비스하는 경우로 구분된다. 홈페이지를 작성할 때에도 이 디렉토리에 특별한 서브디렉토리(예: html)를 두어 그 안에 필요한 .html 이나 .htm과 같은 파일을 작성할 수 있게 된다. 실제로 URL에서도 Tilda(~)가 붙은 첫 번째 depth를 하나의 홈페이지로 본다. 즉 다음은 홈페이지의 하나로 정의된다.

예) http://nic.or.kr/~wkim, http://202.30.64.22/~wkim

그러나, 다음과 같은 URL은 홈페이지가 아닌 것으로 정의한다.

예) http://nic.or.kr/~wkim/intro
http://nic.or.kr/wkim

그러므로 로봇이 수집한 각 HTML 문서에 대하여 도메인 이름에서 .kr로 끝나는 HTML 문서와 호스트 이름(예: nic.or.kr)이나 IP주소로만 된 URL일 경우에는 그 해당 호스트의 IP가 국내의 IP범위에 있는 문서들 중에서 홈페이지 개수를 산출하였다.

<국내 홈페이지 개수 산출 기법>

① 국내 HTML 문서 수집
- ①-1 .kr 국가도메인으로만 구성된 HTML 문서 수집
- ①-2 .com, .net, .org 등에서 국내 IP주소 배정리스트에 속하는 HTML 문서 수집
- ①-3 호스트이름 또는 IP주소로만 된 URL 수집

② ①번에서 추출된 HTML문서의 국내 IP주소범위 점검

③ 호스트시작페이지 개수 + 계정을 가진 홈페이지 개수 + 홈페이지 제공업체 홈페이지 개수

4.2 활성도메인(.kr, .gTLD)의 국내 보유개수

.com과 같은 국제도메인은 국제인터넷주소관리기구(ICANN)에서 관장하고 있으며 VeriSign사에서 위임받아 등록 및 제반 관리를 수행하고 있으며, .kr 도메인(ccTLD)는 한국인터넷정보센터에서 등록 및 제반 관리를 수행하고

있다. 그러나 등록한 모든 도메인이 실제로 활성화(Active)되고 있지는 않다. 국내 인터넷 사용자가 등록하여 활용하고 있는 활성 국제도메인 개수, 활성 국내도메인 개수로 구성되는 2 종류의 통계를 산출하였다. 각 종류별 통계의 정확한 정의는 아래와 같다.

(1) 활성국제도메인 개수

국내 IP 영역에 속하면서 1차 도메인명이 COM, NET, ORG, EDU 등의 도메인 개수를 의미한다. 주어진 도메인에 대하여 해당 도메인을 이용하여 웹사이트나 홈페이지를 구축한 경우, 전자메일 서버 또는 FTP 사이트를 구축한 경우 중에서 한가지 이상에 해당하면 그 도메인은 활성화되고 있다고 정의한다.

(2) 활성국내도메인 개수

현재 80 포트를 통해서 웹 서비스를 하는 있는 도메인 개수를 의미한다.

<활성 도메인의 국내 보유개수>

① 국내 HTML 문서 수집
② HTML을 파싱하여 URL, Email주소, FTP주소 정보추출
③ 도메인네임을 추출하여 국제도메인과 국내도메인 분류
④ 국내 IP주소범위 점검
⑤ 활성 국제도메인 개수 및 활성 국내도메인개수산출

4.3 국내 호스트 개수

일반적으로 Anonymous FTP 서버, 메일서버, 전자결재서버, 웹서버, 인트라넷서버 등 DNS zone화일에 등록된 서버를 호스트라고 정의하며, WWW 호스트는 DNS zone화일에 등록되어 있고 웹 서버로서 구동되는 것으로 정의한다. 기존에 활용하여 왔던 DDT 등의 방법으로는 더 이상 정확한 국내 호스트 개수의 산출이 가능하지 않다. 그러므로 본 고에서의 존 파일 Transfer 기능이 정상적으로 동작하는 호스트의 경우에는 존 파일을 분석하여 호스트를 산출하고, 또한 웹 상의 HTML 문서에서 링크나 Email 주소 등에서 서브 도메인을 추출하여 호스트를 산출하여 이 두 결과를 조합하여 국내 호스트 개수를 산출하였다.

<국내 호스트 개수 산출 기법>

① 국내에 배정된 IP주소 리스트에서 IP추출 · 생성
② gethostbyaddr()기능의 library function으로 해당 IP주소의 Name server 등록 확인
③ 등록된 IP의 이름(HOSTIP.LIST) 출력
④ 위의 2과정에서 Timeout초과시 미등록 처리

<WWW 서비스 호스트 개수 산출 기법>

① 국내에 배정된 IP주소 리스트에서 IP추출 · 생성
② 해당 IP의 80포트로 소켓(Socket) 연결 시도
③ 연결성공시 HTTP 문서요청 Query 시도
④ 일정한 크기의 문서 수신시 WWW 서비스 호스트로 간주하고 해당 IP(WWWHOSTIP.LIST)를 저장
⑤ 위의 ②, ③, ④ 과정에서 Timeout 초과시 1로 리턴

V. 시험 및 산출결과

5.1 시뮬레이션환경

로봇에이전트 시스템은 그림2.와 같이 2개의 하드웨어로 구성되는데, Back-End 로봇 서버는 주로 HTML 문서를 수집하는 HTML 로봇 서브시스템과 호스트 로봇 서브시스템 및 신규/삭제 HTML 처리 서브시스템이 구동되며, Front-End 서버는 주로 통계 산출 서브시스템이 구동되어 관리자 홈페이지를 통하여 각종의 통계 기법에 의하여 산출된다.

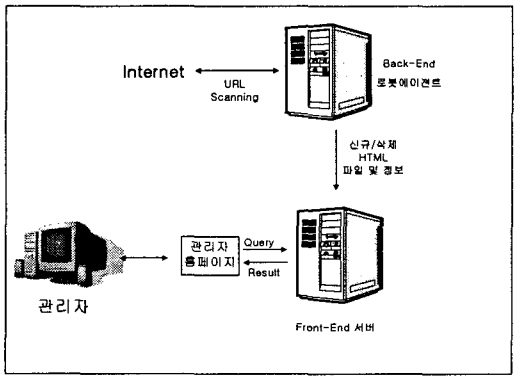

그림 2. 로봇에이전트 시뮬레이션 구성도

5.2 성능분석 및 산출결과

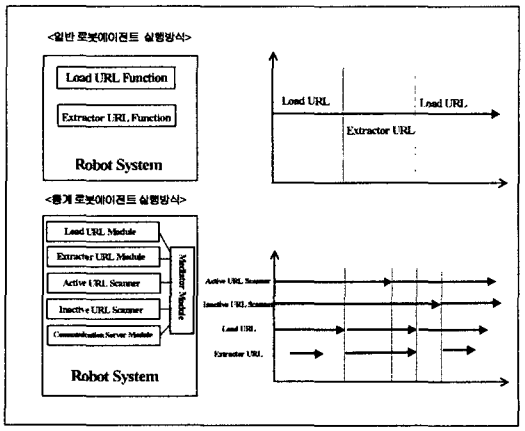

그림 3. 개선된 로봇에이전트 실행방식

인터넷 주요 통계산출을 위한 로봇에이전트의 모듈별 성능을 보면 HTML 문서를 로드하는 모듈의 경우 시간당 4만개의 HTML 문서를 수집할 수 있는 성능을 보였다. 특히 신규로 생성 또는 소멸되는 HTML 문서를 수집 및 처리하는 모듈도 24시간당 96만개의 HTML 문서를 처리할 수 있는 성능을 보였다. 기존 검색엔지 로봇 에이전트는 일반적으로 단일시스템의 경우 24시간에 20만건의 HTML을 수집할 수 있다. 자료 수집속도 측면에서 기본 로봇 에이전트의 수집속도 제한성을 극복하기 위하여 그림 3.와 같이 구성 모듈간의 의존성을 최소화하여 병렬·분산 수집 및 처리가 가능하도록 하였다.

(1) 국내 웹문서 및 홈페이지 개수

2001년 2월과 3월 현재 HTML Loader와 IP scanner 모듈의 동작으로 수집된 우리나라의 국내 Active한 HTML 문서 개수는 2월의 경우 14,445,028개(.html, .htm, .asp, .php 포함), 3월의 경우 16,151,267개이며 .kr 도메인을 이용하여 서비스하고 있는 HTML 문서는 2월의 경우 8,519,478개, 2월의 경우 9,297,356개로 집계되었다. 또한 .com 등 국제도메인을 이용하여 서비스하고 있는 HTML 문서는 2월의 경우 6,042,965개, 3월의 경우 6,989,394개로 집계되었다. 또한, hwp, doc 등과 같은 문서파일수는 2월에 358,251개에서 3월 436,078개로 나타났다. image, sound, vid대와 같은 멀티미디어 파일수는 5,042,520개에서 8,437,883개로 증가하였다.

표 1. 국내 홈페이지 개수

구 분	2월	3월
①호스트의 시작페이지 ②계정을 가진 홈페이지 ③홈페이지 제공업체의 홈페이지 ① + ② + ③	450,236	619,794

(2) 활성도메인의 국내 보유 개수

표 2. 활성도메인의 국내 보유 개수

구분	type	활성화 개수	
		2월	3월
국제 도메인	com+edu+net+org+to	32,618	46,652
국내 도메인	ac+co+go+ne+nm+or+pe+re+지역	73,534	92,991
계		106,152	139,643

(3) 국내 호스트 개수

이번 시험을 통한 호스트와 WWW 호스트 개수 산출결과는 표 3, 4과 같은데 최대 예상 소요시간이 300여시간 소요될 것으로 전망하였으나, 본 고에서 구현된 로봇에이전트를 통하여 시험을 한 결과 30~34시간으로 충분하였다.

표 3. 국내 호스트 산출 결과 및 성능

기준일자	산출HOST 개수	실험환경		
		Process개수	최대예상 소요시간	실제소요 시간
2001. 2월	431,677	400	300	34
2001. 3월	502,762	400	300	30

표 4. 국내 WWW 호스트 산출 결과 및 성능

기준일자	산출WWW HOST 개수	실험환경		
		Process개수	최대예상 소요시간	실제소요 시간
2001. 2월	121,671	800	300	30
2001. 3월	124,087	800	300	29

VI. 결 론

지금까지 로봇에이전트를 이용한 국내 홈페이지 개수, 활성도메인의 국내 보유개수, 국내 호스트 개수 등을 정기적으로 산출하기 위한 기법 설계 및 구현에 관하여 기술하였다. 또한 HTML 상 In-link된 일반문서, 이미지, 사운드, 비디오 파일 개수를 산출할 수 있는 기법에 대한 개발하고 그 결과를 제시하였다. 통계산출 로봇에이전트는 다양한 분야의 참조지수가 될 수 있는 인터넷 주요 통계를 산출할 수 있을 것이다. 다만, 본 고에서 제시한 로봇에이전트를 이용하여 산출한 통계치는 추정치로서 적어도 수개월 정도의 검증을 통하여 지속적인 보정이 필요하고, 추진할 예정이다. 본 논문의 인터넷 환경통계 결과는 국내 인터넷 발전에 대한 기본 통계를 제시하는 이론적 토대가 될 것이다.

참 고 문 헌

[1] http://stat.nic.or.kr, http://www.nic.or.kr

[2] 김원, "국·내외 인터넷 동향", 지식정보인프라, 연구개발정보센터, pp82~87, 2000, 7월호

[3] http://www.nua.ie

[4] http://www.cc.gatech.edu/gvu/user_surveys/User_survey_Home.html

[5] http://marketingtools.com, http://www.nsi.com

[6] http://www.apnic.net

[7] http://www.inktomi.com

[8] 신동욱, "인터네트 환경에서의 분산정보 검색 시스템의 설계 및 구현", 한국과학재단연구과제 961-0911-060-2, 1998, 충남대