로봇에이전트를 이용한 인터넷 주요 통계산출 알고리즘 설계 및 구현

정회원 김 원*, 진 용 옥**

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

Weon Kim*, Yong Ohk Chin** *Regular Members*

요 약

인터넷 시장규모 확대 및 이용환경의 급속한 변화로 인하여 인터넷 이용자에 관한 통계와 인터넷 환경에 관한 통계정보 요구는 증대되고 있다. 그러나 인터넷 환경에 관한 통계 정보중에서 국내 호스트개수, 홈페이지 개수, 국제도메인의 국내 보유 개수 산출 등은 인터넷에 연결하는 이용기관의 보안 강화의 문제점과 전문 지능형 로봇에이전트 시스템의 부재 및 국제도메인 등록기관의 통계 비공개 등으로 국내에서 주기적으로 산출하는 데 문제점이 부각되고 있다. 본 논문에서는 인터넷 관련 주요 통계정보의 정확한 산출·제시로 민간의 인터넷 산업에 대한 효과적인 투자 유도를 가능케 하기 위해서 인터넷 주요 통계 산출이 가능한 로봇에이전트 설계 기법을 제안하고 구현한다. 모듈은 로봇에이전트 프로세스 모듈, 통계산출 모듈, 관리 모듈 등으로 구성되었으며, 국내의 호스트 개수, 홈페이지 개수, .com 등 국제도메인의 국내 보유 개수 등을 정기적으로 산출되기 위한 알고리즘과 그 구현결과를 제시한다.

ABSTRACT

The demand for statistics on internet users and environment is increasing as the internet changes rapidly and the size of its market grows explosively. But there are such several obstacles as security of the organization logging on to the internet, and non-existence of intelligent robot agent system to measure and no open of gTLD registry's statistics that it is hard to obtain statistics periodically on the number of domestic internet hosts, homepages and generic top level domains in korea. This thesis proposes the design method of 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 internet hosts, homepages and .com domains. It provides the result of the implementation as well.

Ⅰ. 서 론

2000년 12월 한국인터넷정보센터[1,2,3]의 발표에 따르면 현재 국내 인터넷 이용자 인구는 약2,000만 명에 근접하고, 2000년 12월 Nua사 발표자료에 의하면 전세계적으로 4억명 이상으로 추산되고 있다[4]. 일반적으로 인터넷 통계는 크게 인터넷 이용자수, 이용실태 현황 등의 이용자에 관한 통계 부문(1)과 홈페이지 개수, 호스트 개수 등의 인터넷 환경에 관한 통계 부문(2)으로 구분할 수 있다.

* 경희대학교 전자공학과, ** 경희대학교 전파공학과 교수
논문번호 : 010042-0320, 접수일자 : 2001년 3월 20일

(1) 인터넷 이용자에 관한 통계는 국내의 경우 한국인터넷정보센터(KRNIC)의 인터넷 이용자 통계 조사, 국외의 경우 GVU's WWW User Survey, American Internet User Survey, Nua Internet Survey 등을 통해 주기적으로 발표되고 있다[5,6].

(2) 인터넷 환경에 관한 통계는 국내외적으로 국가인터넷레지스트리(NIR : National Internet Registry)[7]가 도메인 개수, 호스트 수 등에 관한 자료를 산출하여 주기적으로 발표하고 있다.

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

Ⅱ. 국내외 인터넷통계산출 현황과 문제점

본 장에서는 인터넷관련 통계 산출에 관한 현황과 문제점을 도출함으로써 본 논문에서 연구·제시하고자 하는 중요성을 기술한다. 그러나 인터넷 환경에 관한 통계 부문에서는 국내외적으로 정확한 통계 산출에 어려움에 직면하고 있다. 그 주된 이유를 살펴보면 다음과 같다.

① [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 그림 1.과 같이 DDT 프로그램에 대한 보안강화 등 아래와 같은 문제점으로 인하여 현재 공신력 있는 자료 제공이 가능하지 않다.

- o 데이터의 정확성 측면의 문제점
  - DNS에 등록되지 않은 Host 파악 불가
  - 프로그램 실행시 DNS 서버의 일시적인 오동작으로 인한 오차 발생
  - Dial-up ISP업체의 dynamic IP 주소 사용으로 인한 실제 ISP에 접속해 있는 Host 개수가 아닌 Modem/ISDN의 포트 개수로서 호스트 개수 산정
- o 보안강화에 따른 DNS 네임서버상의 문제점
  - Firewall 사용에 따른 Zone File Transfer (AXFR)명령을 허용치 않음에 따른 산정 불가

② [홈페이지 개수 산출] 인터넷 환경에 관한 통계 중에서 홈페이지 수의 산출은 미국의 경우 전문 검색업체인 Inktomi사[8] 에서 자사의 강력한 로봇을 통하여 수집한 HTML 문서를 분석하여 그 개수를 산출하여 발표하고 있으나, 아래와 같이 국내의 경우 그러한 로봇에이전트와 수집한 HTML를 분석하기 위한 알고리즘 및 기법이 개발되어 있지 않으므로 현재 자료제공이 가능하지 않다.

- o 홈페이지 산출을 위한 전용 로봇에이전트의 부재
  - 다이내믹 방식의 HTML 문서 수집을 위한 로봇 기술 부재
- o 홈페이지 개수 분석을 위한 소프트웨어 부재

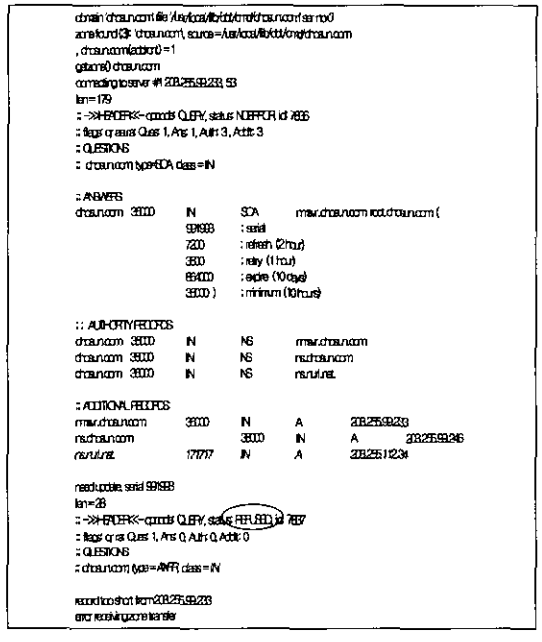


그림 1. DDT 프로그램을 이용한 존 파일전송의 실패 사례

③ [국제도메인의 국내 보유 개수 산출] 인터넷 환경에 관한 통계 중에서 국제도메인 개수의 경우 미국의 Network Solutions사[6] 에서 등록된 전세계의 도메인을 분석하여 간헐적으로 국가별 순위 통계만을 제공하고 있다. 따라서, 국내에서 보유하고 있는 국제도메인 개수를 주기적으로 산출할 수 있는 방법이 존재하지 않는다.

Ⅲ. 로봇에이전트 설계 및 구현

본 장에서는 본 고에서 제안하는 로봇에이전트[9,10] 설계와 구현된 구조에 대해서 기술한다.

1. 필수기능

인터넷 주요 통계 산출 작업은 그 결과가 인터넷 환경 변화의 진행 정도를 정확하게 제시할 수 있어야 하므로 주기적으로 수행되어야 한다. 그러므로 인터넷 주요 통계 산출을 위한 로봇에이전트는 기본적으로 아래와 같은 필수 기능이 구현되어야 한다. 그림 2.는 로봇 에이전트의 기본구성도로서 3가지의 필수기능이 구현되어야 한다.

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

② 주기적으로 문서 자동 수집이 가능하고 편리한 관리 기능

③ 에러복구(Error Recovery)를 할 수 있는 능력

①항의 기능을 충족하기 위하여 필수조건

o Incremental Search 기능

기 수집한 HTML 문서 중에서 삭제되었거나 신규로 인터넷에 올려진 HTML 문서에 대한 정보를 추출한 후 신규 문서만을 수집할 수 있는 기능

o Distributed Search 기능

인터넷 환경의 지속적인 확대로 인하여 문서 수집에 요구되는 컴퓨팅 파워가 단일 시스템이 제공하는 컴퓨팅 능력을 초과할 경우 여러 대의 시스템을 활용하여 문서수집이 가능한 기능

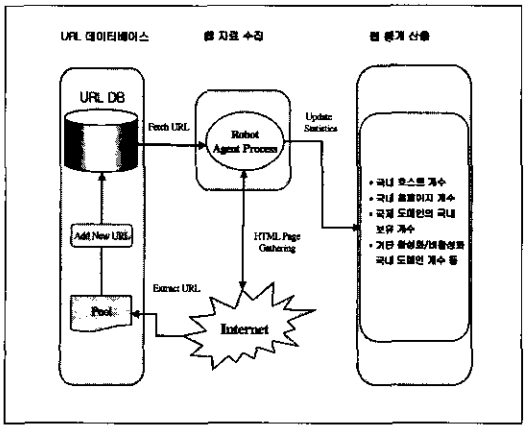


그림 2. 로봇에이전트 기본 구성도

②항의 기능을 충족하기 위한 필수조건

o Decoupled Architecture 기능

하위 HTML 수집 로봇에이전트 모듈과 통계 산출을 위한 상위 HTML 분석 모듈이 상호 독립적으로 동작할 수 있는 기능

o Continuous Acting 기능

웹 문서 수집 로봇에이전트는 그림 2.에서 보여지는 것과 같이 관리자가 프로세스를 종료시키거나, 신규 및 변경된 URL이 새로이 수집된 HTML 문서에서 더 이상 나타나지 않을 때까지 지속적으로 자료를 수집하게 된다.

③항의 기능을 충족하기 위한 필수조건

o Fault Tolerancy 기능

로봇이 구동되고 있는 서버가 다운되거나 네트워크의 장애 등으로 인하여 로봇 프로그램의 수행이 중단되었을 경우, 시스템 장애가 복구된 후 로봇 프로그램을 재실행하게 되더라도 이전까지의 수집 상태를 유지한 상태에서 계속적으로 HTML을 수집할 수 있는 에러 복구 기능이 제공되어야 한다.

2. 입·출력 파일구조 및 생성

본 절에서는 로봇에이전트에서 공통적으로 처리되는 기본적인 입·출력파일의 종류와 구현된 그 구조에 대해서 기술한다. 마스터 파일은 각 모듈에서 공통적 으로 사용되는 입·출력 파일로서 세가지의 대분류로 구분된다.

2.1 URL 파일 (*.URL)

URL 리스트는 웹 상에 존재하는 HTML, 일반문서, 이미지, 사운드, 비디오 파일 등의 각 유형 또는 주어진 역할에 따라 생성되는 파일의 종류로 구분된다. 그 파일구조는 가변길이 URL 레코드들의 집합으로 구성되어 있으며 그림 3.과 같다.

00000001	00000001	00000001	00000001	22	nic.com/~wk/intro.html
:	:	:	:	:	:
①	②	③	④	⑤	⑥

그림 3. URL 레코드 필드구조

① INFO(INFO) 필드는 로봇 및 링크추출 프로세스상태(1 byte), 사이트 현황(1 byte), 예약 영역(1 byte), 해당 URL에 위치한 파일 유형 및 타입 (1

byte)을 나태는 필드로서 4 byte 구조로 되어 있다.
② ID(HDID/DOCID/IMGID/SNDID/VIDID) 필드는 HTML, 일반문서, 이미지, 사운드, 비디오 등 문서종류의 ID를 나타내는 필드(type unsigned int)로서 각 유형별로 고유한 값을 갖는다.
③ FileStatus(STATUS) 필드는 각 문서의 생성 날짜 또는 크기를 나타내는 필드(unsigned int형)이다.
④ SourceHTMLID(HTID) 필드는 NonHTML 문서를 위한 필드로서 NonHTML 문서가 in-link된 HTML 문서의 ID를 표시한다.
⑤ URLByteSize 필드는 URL 문자열의 Byte 단위 크기를 나타낸다.
⑥ URL 필드는 해당문서의 URL을 저장하는 필드(type char)이며, 크기는 URL의 Byte 수와 동일하다.

URL파일은 구현된 load_URL() 함수를 이용하여 각 유형별로 URL 파일 구조체에 저장되는데 저장되는 URL 파일 런타임 구조는 그림 4.와 같다.

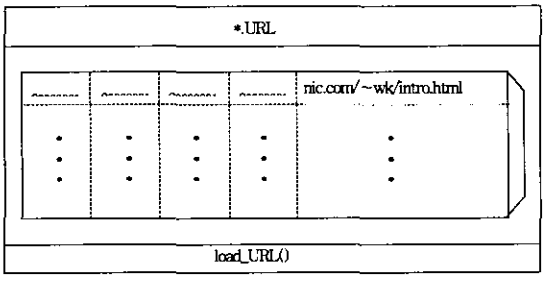


그림 4. URL 파일 런타임 구조

마스터 URL 파일(*.URL) 이름은 문서 유형에 따라 (HTML, DOC, IMG, SND, VID, LINK) + ".URL")로 구성된다. 예로서 http://nic.com/~wk/intro.hwp 의 경우는 "DOC.URL" 파일에 위치한다.

2.2 HTML FreeID 파일 (*.FreeID)

지속적인 웹문서(HTID), 일반문서(DOCID), 이미지 파일(IMGID), 사운드 파일(SNDID), 비디오 파일(VIDID)의 변경에도 동일한 ID 한계값을 유지하기 위해서 사용되는 파일이다. URL 파일을 생성하기 위해서 신규 HTML 파일을 파싱할 때 새로운 URL이 나타나면 해당 FreeID파일(*.FreeID)에서 재사용 가능한, 즉 Free가 된 ID 존재여부를 파악하여 존재하면 그 값을 부여하고 그렇지 않을 경우는 새로운 HTID, DOCID, IMGID, SNDID, VIDID 값을 부여한다.

① 파일 종류 : HTML.FreeID, DOC.FreeID, IMG.FreeID, SND.FreeID, VID.FreeID로서 구분된다.
② 파일 구조는 그림 5.와 같다.

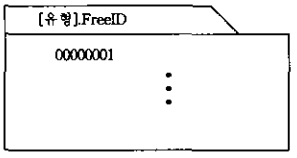


그림 5. FreeID 파일구조

③ 필드 설명 : 유형 ID(HTID, DOCID, IMGID, SNDID, 또는 VIDID)는 각 URL 유형별 고유한 문서의 ID를 나타내는 필드이다.
④ 런타임(Runtime) 구조 : load_FreeID() 함수를 이용해서 그림 6.과 같은 구조로 메모리에 로딩된다.

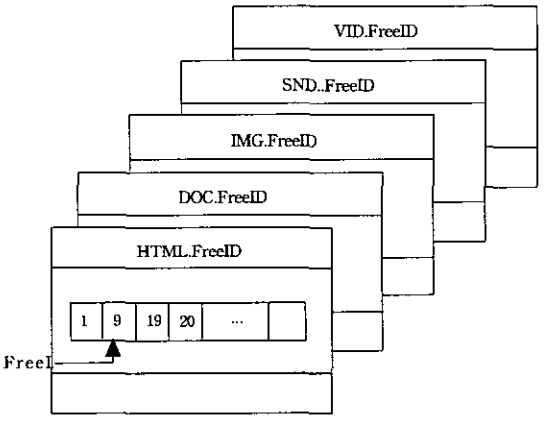


그림 6. FreeID 런타임 구조

2.3 로봇 수집 문서의 관리

수집된 HTML 문서는 URL 파일(ToLoad.URL)로부터 URL set을 위임받아 로봇이 방문한 후 해당 URL 파일의 필드(INFO, STATUS) 정보를 비

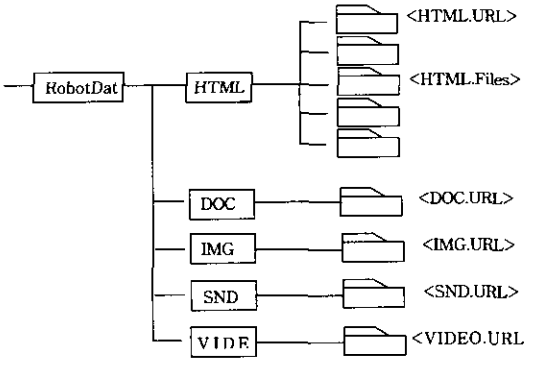
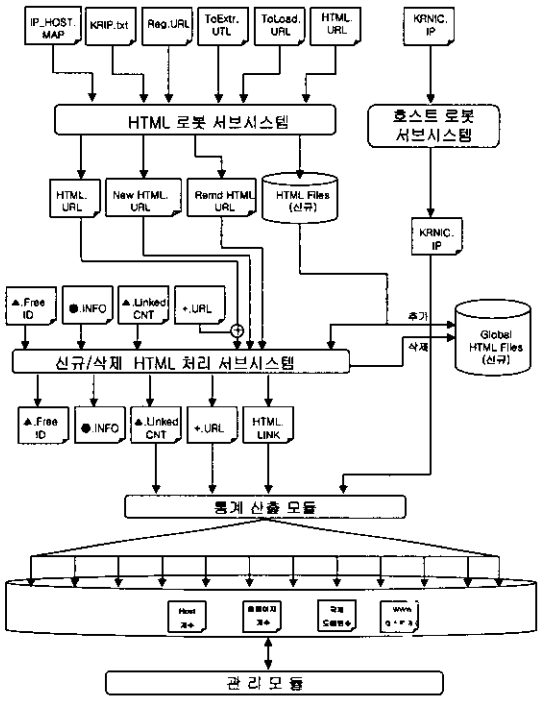


그림 7. 로봇 수집자료 및 URL 파일 저장경로

교 또는 기록하고 그림 7.과 같은 디렉토리 구조로 다운로드한 파일들을 저장한다.

3. 구현된 모듈 기능

본 절에서는 국내 호스트 개수, 국내 홈페이지 개수, .com과 같은 국제도메인의 국내보유 개수 등에 대한 최적화된 인터넷 주요 통계 산출을 위한 수집 모듈과 입·출력 파일의 흐름은 그림 8.과 같다.



주 ▲ : HTML, DOC, IMG, SND, VID
- : DOC, IMG, SND, VID
+ : DOC, IMG, SND, VID, LINK
* : HTML, DOC, IMG, SND, VID, LINK

그림 8. 수집모듈과 입·출력 파일의 흐름도

3.1 로봇에이전트 프로세스 모듈

o HTML 로봇서브시스템 : HTML 문서수집 기능 수행

- Load URL 생성모듈 : 문서 Load 모듈이 수집할 HTML의 URL 리스트를 생성하는 기능으로 신규 HTML 페이지로부터 In-link HTML URL을 추출하여 HTML.URL과 비교항 ExtrNew.URL을 작성한다. 또한 관리 등록 URL 리스트(Reg.URL)를 HTML.URL과 비교하여 신규 URL 리스트인 ExtrNew.HRL에 추가한다.
- HTML loader 모듈 : HTML 문서 및 URL 적재(Load)를 위한 프로세스로서 그 자식 프로세스들인 Bot 프로세스들에게 ToLoad.URL의 URL 셋(set)을 위임하여 웹상의 문서를 수집한다.
- Active URL scanner 모듈 : HTML.URL의 URL중에서 Active한 URL에 대하여 내용이 변경되었거나 비활성화된 HTML 페이지의 URL 리스트를 작성한다.
- Inactive URL scanner 모듈 : HTML.URL의 URL중에서 Inactive한 URL에 대하여 재활성화되었거나, 죽은 HTML 페이지의 URL 리스트를 작성한다.
- Mediator 모듈 : 각 모듈의 출력파일들을 활용하여 그 서브시스템의 재구동시의 입력 파일들과 인덱싱 시스템의 입력으로 사용될 신규 및 삭제 URL 파일을 생성한다.

o Host 로봇서브시스템 : 국내에 배정된 IP주소 리스트를 입력으로 하여 HTML 로봇서브시스템에서 수집된 HTML의 URL리스트와 비교하여 각 IP별로 일반적으로 메일서버, FTP서버인 호스트, 국내 WWW 서비스를 하는 호스트 여부를 결정하여 보관한다. 세부모듈로서 IP scanner 모듈이 존재한다.

o 신규/삭제 HTML 처리 서브시스템

- 신규 HTML 처리 모듈
- 삭제 HTML 처리 모듈

3.2 통계산출 모듈

o HTML 통계처리 모듈 : 홈페이지 개수, HTML 페이지 개수, 월별 HTML 페이지의 변경 및 삭제 비율, 주요 HTML 페이지별 피참조(Linked) 수를 산출한다.

o 도메인 통계산출 모듈 : 국제 도메인 개수, 도메인별 HTML 페이지수, 활성(Active) 도메인 및 비활성(Inactive) 도메인 개수를 산출한다.

o 호스트 통계산출 처리 모듈 : 일반적으로 인터넷에 연결된 서버역할을 하는 호스트 개수와 Web 서비스하는 호스트 수를 산출한다.

3.3 관리 모듈

o 통계 Query 처리 모듈 : 웹 인터페이스를 통하여 인터넷 통계산출 결과를 조회할 수 있는 기능을 수행한다.

o 시스템 관리 모듈 : 웹 인터페이스를 통하여 각

서브 시스템의 구동 상황 조회 등의 관리기능을 수행한다.

Ⅳ. 알고리즘 개발

본 장에서는 논문에서 기술하는 로봇에이전트를 이용하여 산출하고자 하는 주요통계 산출알고리즘에 대해서 기술한다.

1. 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. 2장에서 지적한 바와 같이 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 그림 9.와 같이 국내 호스트 개수를 산출하였다.

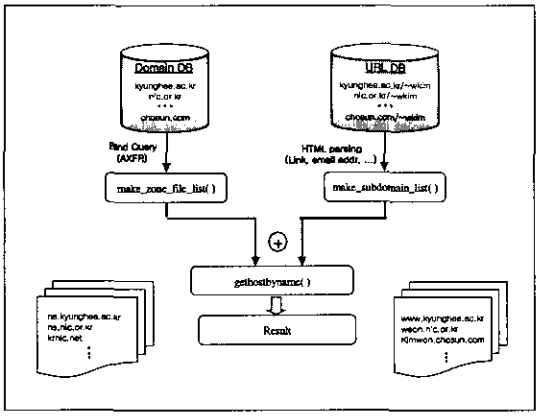


그림 9. 국내 호스트 개수 산출 알고리즘

<국내 호스트 개수 산출 알고리즘>

① 국내에 배정된 IP주소 리스트에서 IP추출 · 생성
② gethostbyaddr()기능 내장 프로그램으로 해당 IP의 네임서버 등록을 확인
③ 등록된 IP의 이름 출력
④ 위의 2과정에서 Timeout초과시 미등록 처리
⑤ 위의 ①, ②, ③ 과정 반복 수행

<WWW 서비스 호스트 개수 산출 알고리즘>

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 IP를 저장
⑤ 위의 ②, ③, ④ 과정에서 Timeout 초과시 1로 리턴
⑥ 위의 ①, ②, ③, ④ 과정 반복 수행

2. 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 첫번째 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 알고리즘>

① HTML 문서 수집
② .kr 국가도메인으로만 구성된 HTML 문서 수집
③ 호스트이름 또는 IP주소로만 된 URL 수집
④ .com, .net,. org등에서 국내 IP주소 배정리스트에 속하는 HTML 문서수집
⑤ ②, ③, ④번에서 추출된 HTML문서의 국내 IP 주소범위 점검
⑥ 위의 ①, ②, ③, ④, ⑤ 과정 반복 수행

3. 국제도메인의 국내 보유 개수

.com과 같은 국제도메인은 국제인터넷주소관리기구(ICANN)[11] 에서 관장하고 있으며 Network Solutions 사에서 위임받아 등록 및 제반 관리를 수

행하고 있다. 그러나 국내 기업 또는 개인의 국제도메인(*.com) 보유 개수를 Network Solutions 사로부터 주기적으로 얻는 것은 현실적으로 매우 어렵다. 그러므로 본 고에서는 국내 인터넷 상의 HTML 문서상에 링크(link)되어 있는 URL을 추출한 후 호스트 이름 중에 .com 등 도메인 명으로 되어 있는 모든 호스트 이름을 분석한 후 이들의 IP들을 추출하여 국내 IP 내에 있는가를 검사하여 국내 IP 내에 있는 .com 도메인을 국내 기업 또는 개인 보유의 .com 도메인으로 간주하여 국내 보유 .com 도메인의 개수를 그림 10.과 같이 산출한다.

<국제도메인의 국내 보유 개수 산출 알고리즘>

① HTML 문서 수집

② .com, .net, .org 등 국제도메인으로 구성된 HTML 문서만 수집

③ ②번에서 추출된 HTML문서의 국내 IP주소범위 점검

④ 위의 ①, ②, ③ 과정 반복 수행

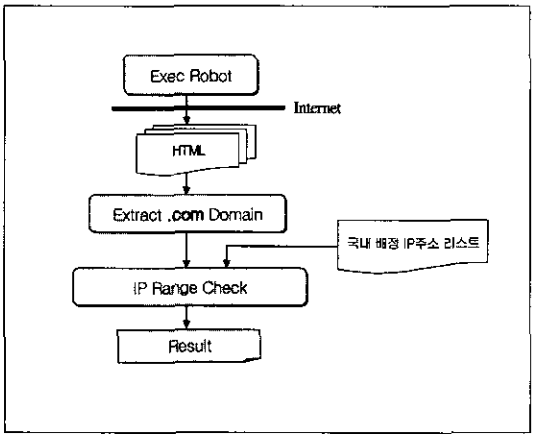


그림 10. 국제도메인의 국내 보유수 산출 알고리즘

V. 시험 및 산출결과

본 장에서는 구현된 로봇에이전트를 이용한 인터넷 주요 통계산출의 시뮬레이션 환경과 그 결과에 대해 기술한다.

1. 시뮬레이션환경

그림 11.과 같이 로봇에이전트 시스템은 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 알고리즘에 의하여 산출된다.

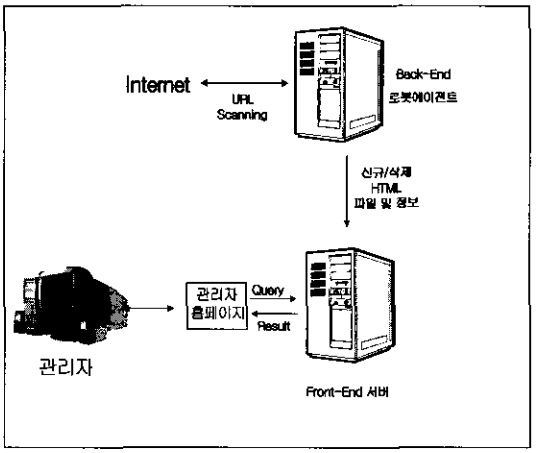


그림 11. 로봇에이전트 시뮬레이션 구성도

2. 산출결과

인터넷 주요 통계산출을 위한 로봇에이전트의 모듈별 성능은 표 1.과 같다. 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 시간당 4만개의 HTML 문서를 처리할 수 있는 성능을 보였다.

표 1. 로봇에이전트 모듈의 성능

구분	세부모듈	처리시간	비고
HTML 로봇 서브시스템	HTML Loader 모듈	22H/100만 페이지	4만개/h
	HTML Extractor 모듈	40H/100만 페이지	2만5천개/h
호스트 로봇 서브시스템	IP Scanner 모듈	200H/국내 IP	5만 IP/h
신규/삭제 HTML 처리 서브 시스템	신규 HTML 처리 서브시스템	27H/100만 페이지	4만/h

이번 시험을 통한 호스트와 WWW 호스트 개수 산출결과는 표 2, 3과 같은데 최대 예상 소요시간이 300여시간 소요될 것으로 전망하였으나, 본 고에서 구현된 로봇에이전트를 통하여 시험을 한 결과 34

시간으로 충분하였다.

2001년 2월 10일부터 2월 25일까지 16일간에 걸쳐 HTML Loader와 IP scanner 모듈의 동작으로 수집된 우리나라의 국내 Active한 HTML 문서 개수는 14,035,439개(.html, .htm, .asp, .php 포함)이며 .kr 도메인을 이용하여 서비스하고 있는 HTML 문서는 약830만개, .com 등 국제도메인을 이용하여 서비스하고 있는 HTML 문서는 약570만개로 집계되었다.

표 2. 국내 호스트 예비산출 결과

IP개수	산출HOST 개수	실험환경		
		Process개수	최대예상 소요시간	실제소요 시간
13,810,176	548,088	400	300	34

표 3. 국내 WWW 호스트 예비산출 결과

IP개수	산출WWW HOST 개수	실험환경		
		Process 개수	최대예상 소요시간	실제소요 시간
13,810,176	115,078	800	300	29

또한 표 4.와 같이 국제 도메인별 국내의 HTML 문서 개수를 산출한 결과 .com을 이용한 HTML 문서는 3,871,889개이었다. 그러나 이러한 통계결과는 2월 현재의 통계이며,로봇시스템이 연속 동작됨에 따라 증가할 것이고 최종적으로 산출된 통계결과는 아니다라는 것을 밝힌다.

표 4. 국제 도메인별 국내 HTML 문서 개수

도메인 종류	HTML 개수
.com	3,871,889
.net	1,508,909
.org + .edu	313,178
계	5,693,976

Ⅵ. 결 론

지금까지 로봇에이전트를 이용한 국내 호스트 개수, 홈페이지 개수 및 국제도메인의 국내 보유 개수 등을 정기적으로 산출하기 위한 알고리즘 설계 및 구현에 관하여 기술하였다. 향후에는 HTML 상 In-link된 일반문서, 이미지, 사운드, 비디오 파일 개수를 산출할 수 있는 알고리즘에 대한 개발·구현을 할 예정이며, 주요 HTML 문서별 피참조(Linked)수도 산출할 수 있는 알고리즘 개발 및 구현이 가능할 것이다. 이처럼 통계산출 로봇에이전트는 다양한 분야의 참조지수가 될 수 있는 인터넷 주요 통계를 산출할 수 있을 것이다.

2000년 후반기부터 우리나라의 벤처는 한글도메인의 등록 및 운영을 20개 업체가 다양하게 틈새시장을 노려 서비스하고 있으며, WAP 폰을 이용한 키워드 방식의 무선도메인 등도 등장하고 있다. 따라서 본 지능형 로봇에이전트로 mDN(Multilingual Domain)[12,13]과 키워드방식의 한글도메인, WAP 폰을 이용한 홈페이지 개수 및 도메인 개수도 산출할 수 있는 알고리즘 개발과 성능의 개선을 추진할 것이다. 또한 국제적으로 IETF 등 국제회의 참석 등으로 통계를 전문적으로 산출할 수 있는 로봇에이전트와 알고리즘에 대한 논문과 동향 등을 분석하고 앞서 나가야 할 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] 김원, "NIC와 국가망 인터넷서비스 현황 및 전망", 제2회 수원지역 대학교수 및 전파·정보통신 전문가 초청세미나, pp 185-215, 1998. 7. 3, 경희대

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

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

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

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

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

[9] 신동욱, "인터네트 환경에서의 분산정보 검색시스템의 설계 및 구현", 한국과학재단연구과제 961-0911-060-2, 1998, 충남대

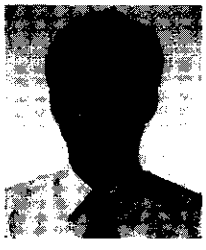
[10] 김판구, 조유근, "상호정보에 기반한 한국어 텍스트의 복합어 자동색인", 정보과학회 논문지 21권 6호, 1994

[11] http://www.icann.org

[12] 김원, 진용옥, "인터넷의 한글도메인 체계 구현

에 관한 연구", 전자공학회 하계종합학술대회 제21권 제2호, pp. 301-304, 1998. 6. 27, 경산대(발표장소)
[13] 김원, 진용옥, "한글도메인이름을 지원하기 위한 Proxy HDNS 구현", 전자공학회논문지 제36권 C편 제12호, pp1~9, 1999. 12

김 원(Weon Kim) 정회원



1984년 2월 : 한양대학교 전자공학과 졸업.
1989년 2월 : 한양대학교 대학원 전자공학과 공학석사
1998년 3월~현재 : 경희대학교 전자공학과 박사과정
1984년 11월~1987년 2월 : 국방과학연구소(연구원)
1989년 1월~1992년 6월 : (주)데이콤(주임연구원)
1992년 7월~1999년 6월 : 한국전산원(선임연구원)
1999년 7월~현재 : 한국인터넷정보센터(기술지원부장)
<주관심 분야> 차세대인터넷, 로봇에이전트, 컴퓨터 네트워킹

진 용 옥(Yong Ohk Chin) 종신회원
1968년 2월 : 연세대학교 전기공학 졸업
1975년 2월 : 연세대학교 전자공학과 공학석사
1981년 2월 : 연세대학교 전자공학과 공학박사
1975년~1978년 : 광운전자공과대학 통신공학과 부교수
1979년~1994년 : 경희대학교 전자공학과 교수
1995년~현재 : 경희대학교 전파공학과 교수
1980년 7월 : 통신기술사(전기통신 부문)
1992년 1월~1995년 12월 : 한국음향학회 회장
1991년~현재 : 경희대 부설 정보통신공학연구소 소장
1993년 9월~현재 : 국어정보학회 부회장
1991년 3월~1996년 : 한국통신학회 협동이사
1996년~현재 : 통.우.연 협동이사
1997년~현재 : 한국방송공사 경영평가 위원회 위원장
1996년~현재 : 사단법인 미래사회정보생활 기획이사
1998년~현재 : 경희대학교 부설 정보통신 창업지원센터 소장
<주관심 분야> 통신시스템 및 네트워크구성, 한의정보공학, 국어정보공학