한국경영과학회/대한산업공학회 2003 춘계공동학술대회
2003년 5월 16일-17일 한동대학교(포항)

업종별 포탈 사이트의 효율적 정보제공을 위한 웹로봇 시스템 개발에 관한 연구

Development of the web robot system for an efficient information delivery of portal sites classified by the business types

김광명[1], 백상규[2], 김선호[3]

[1]스타닷컴, kmkim@e-ventree.com
[2]한국과학기술정보연구원, vack@kisti.re.kr
[3]명지대산업공학과, shk@mju.ac.kr

Abstract

While the impact of internet on our everyday life keeps increasing, the internet users are turning toward the value added information only that suits their purpose instead of just wandering over a sea (pool) of information. Although it is easy to locate the information using various search engines, we still finds it time and cost consuming to update the existing information or to add similar information.

There are several portal sites managed by each industrial types or associations/organizations for the purpose of strengthening the competitiveness of small and medium-sized enterprises. And currently, tremendous manpower and expenses are put into the selection and quality improvement of information to satisfy the needs toward the high quality information by the specialized users in their fields. Collection and updating of the information is partially available manually during the period of budget allocation by the government. It becomes problematic, however, to continue this service when the project expires.

This report presents the system for the information collection, analysis and classification of portal sites operated for the special purposes and automatic upload to the proper sites. In addition, expression of web robots and application of robots in several information types are suggested which will eventually accomplish currency of information and automatic updates and addition of documents with the least expenses of maintenance.

1. 서 론

인터넷상의 정보는 매일 방대한 양이 생성과 소멸을 반복하고, 수시로 변경된다. 이러한 인터넷 정보의 유동성 때문에 사용자에게 효율적인 정보제공을 위한 웹로봇 개발 및 관련연구는 WWW(World Wide Web)이 보편화된 1990년대 중반부터 지속적으로 이루어지고 있다. 인터넷 사용자들은 자율적인 Web Navigation 시대를 지나 검색엔진을 이용한 정보 위치파악 시대를 거쳐 자신의 요구나 키워드에 의한 부가가치정보만을 수동적으로 받거나, 분야별 포탈사이트에서 관련정보를 얻는 추세로 바뀌어가고 있다. 이에 따라 웹로봇은 크게 두 가지 형태로 개발되고 있다[4]. 하나

는 정보검색엔진에 필요한 것으로 사용자가 원하는 정보를 찾아주는 역할을 위한 것이고, 또 다른 하나는 사용자에게 맞춤정보 제공을 목적으로 한 것으로 수집된 정보들 가운데서 사용자의 요구나 성향에 따라 가장 적절한 정보만을 골라 정형화된 데이터를 제공한다. 전자는 불특정 다수의 웹사이트가 대상으로 속도나 성능이 핵심인 반면, 후자는 한정된 또는 지정된 수십 혹은 수백개의 웹 사이트에서 비정형 데이터를 정형화시켜 축적시키는 것으로 정확성과 유연성이 강조된다.

근래에 들어 사용자들에게 맞춤정보를 제공하고자 하는 업종/분야별 포탈사이트들이 생겨나고 운영되고 있지만, 웹로봇의 정확성과 데이터 처리의 유연성이 포탈사이트가 요구하는 수준에 미치지 못하여 아직까지 인력에 의한 수동작업으로 컨텐트 구축 및 유지관리가 이루어지고 있는 것이 현실이다. 이러한 현실은 기업이나 단체로 하여금 많은 인력과 시간 투자를 요구하게 되어 정보화 수준을 떨어뜨리고 정보화 욕구를 감소시킨다[1, 3].
산업 및 업종별 경쟁력 강화와 중소기업의 정보화 인프라 구축 및 전자상거래 활성화를 위하여 정부에서는 많은 예산을 투입하여 "업종별 B2B 마켓플레이스 구축", "조합 정보화 기반 구축사업"등을 추진해 왔으며 많은 파급효과를 거두고 있다. 그러나 일정기간의 예산 지원 후 독자적인 운영을 위한 수익모델을 추구해야 되는 부담 있으며, 아직도 유료 컨텐츠 제공을 통해 이익을 실현하기 힘든 현실 상황에서 일부는 사업 기간이 만료됨에 따라 사업비 지원이 중단되어 시스템 운영 및 데이터 갱신등 유지관리가 제대로 되지 않아 서비스가 중단되거나 애써 구축한 사이트의 생명력을 소실할 우려가 발생할 수 있다[7].

본 논문에서는 업종이나 분야별로 특화되어 있는 포탈사이트들이 최소의 비용과 노력으로 지속적으로 사용자의 요구에 부응하는 효율적인 정보를 제공하기 위한 컨텐트를 구축하고, 유지관리하기 위한 웹로봇으로서 Infovider(Information+Provider)의 설계 및 구현에 관한 방법을 연구하였다.

2. 관련연구

2.1 웹로봇의 개념

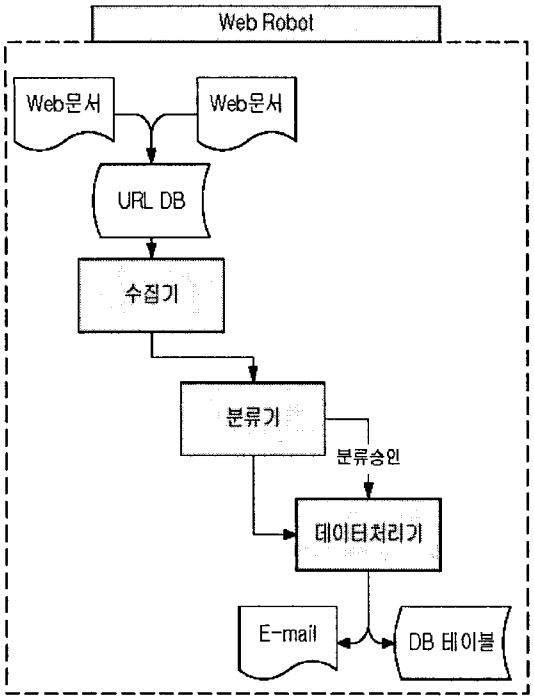
1993년 전세계 웹서버의 숫자를 파악하기 위해 만들어진 "World-Wide Web Wanderer"를 시작으로 초기의 웹로봇은 정보검색을 위해 연구되었다. 당시의 웹로봇은 인터넷 정보를 보다 빠르고 정확히 찾을 수 있는 검색엔진의 한 부분으로서 '관리자의 개입 없이 중복되지 않는 URL을 자동으로 찾아가 인터넷 페이지 정보를 축적하여 사용자가 검색할 수 있도록 하는 소프트웨어'라 정의한다[5].

근래의 폭발적인 인터넷 정보의 증가는 이러한 검색엔진의 중요성을 더욱 부각시켰다. 하지만 검색엔진을 통해 정보의 소재를 파악한 사용자는 해당 사이트에서 새로운 관련 정보의 출현과 변경을 모니터링하기 위해서는 매번 해당 사이트를 방문해야 하는 번거로움이 있다.

따라서 본 논문에서 다루고자 하는 웹로봇인 Infovider는 '한정된 웹사이트에서 정해진 요구에 의해 정보를 수집하고 재가공하여 분류한 후 사용자가 원하는 맞춤 정보를 메일링 서비스하거나 데이터베이스에 적재하는 소프트웨어'로 정의하고, 기존 웹로봇에 비해 수집항목의 정확성과 데이터 처리의 유연성 개선에 초점을 맞추었다.

2.2 웹로봇의 구성요소

일반적으로 웹로봇은 <그림 1>에서 같이 수집기, 분류기와 데이터처리기로 나눌 수 있다. 수집기는 정해진 웹페이지에서 정보를 수집하며, 중복 URL을 방지하기 위한 데이터베이스를 갖고 경우에 따라 분류를 위한 인덱스를 생성하거나 요약문을 생성하는 모듈을 포함한다.



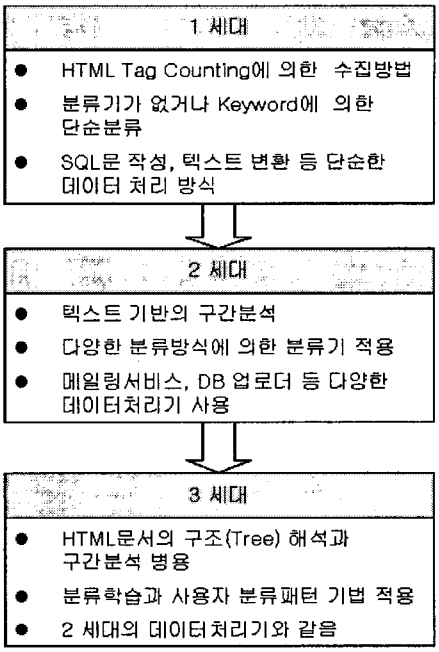
<그림 1> 웹로봇의 구성요소

분류기는 규칙, 확률 또는 학습 기반으로 문서를 분류하며, 좀 더 정확한 분류를 위해 관리자 또는 전문가가 개입할 수 있도록 분류 승인 모듈을 갖는다[2].

데이터처리기는 분류된 정보를 사용자의 요구에 따라 메일링 서비스하거나 DB 테이블에 업로드하여 서비스하는 기능이다.

2.3 웹로봇의 기술동향

웹로봇은 수집기와 분류기의 형태와 기능에 따라 <그림 2>과 같이 1세대, 2세대, 3세대로 나누었다.



<그림 2> 웹로봇의 기술추이

1세대 웹로봇

1세대의 수집기는 HTML 문서내의 태그들을 계산하여 시작위치와 끝위치 사이의 String을 항목별로 가져오는 방법으로 Script, DHTML이 웹문서에 적용된 이후 쓰이지 않고 있다. 1세대의 분류기는 사용자가 지정한 키워드를 수집문서 데이터베이스에서 Query에 의해 분류하는 방법으로 기술분류의 경우 해당 분야 전문가가 아니라면 키워드의 도출이 어렵고, 키워드 누락에 의한 분류되지 않는 문서가 많았다.

2세대 웹로봇

2세대의 수집기는 현재의 범용 웹로봇에 적용되고 있는 방식으로 수집하고자 하는 항목영역 전후의 텍스트를 분석하여 영역을 설정한다. 이러한 방식은 대상 웹페이지의 변화 또는 항목 변화에 대응하는 유연성이 좋지만 일반 사용자들이 텍스트를 분석하여 영역을 설정하는 것에 어려움이 있다. 2세대의 분류기는 지금까지 연구되어진 다양한 분류기법이 적용되었고, 데이터처리기는 다양한 DBMS와

사용자 요구에 맞춰 개발되었다.

3세대 웹로봇

3세대의 수집기는 2세대의 수집기의 기능에 HTML 문서를 Tree구조로 해석하여 수집항목에 대한 영역을 설정하는 것으로 사용자는 웹브라우져 기반의 인터페이스에 의해 자유로이 영역을 설정할 수 있다. 3세대의 분류기는 2세대의 분류방법에 사용자의 분류패턴을 해석하는 방법을 더하여 보다 정확한 분류가 가능토록 개발되었다. 또한 3세대의 분류기는 같은 키워드 또는 분류방법을 수집된 문서의 항목 또는 필드별로 가중치를 달리하여 전문화된 기술이나 업종분류가 가능토록 하였다.

3. Infovider 시스템

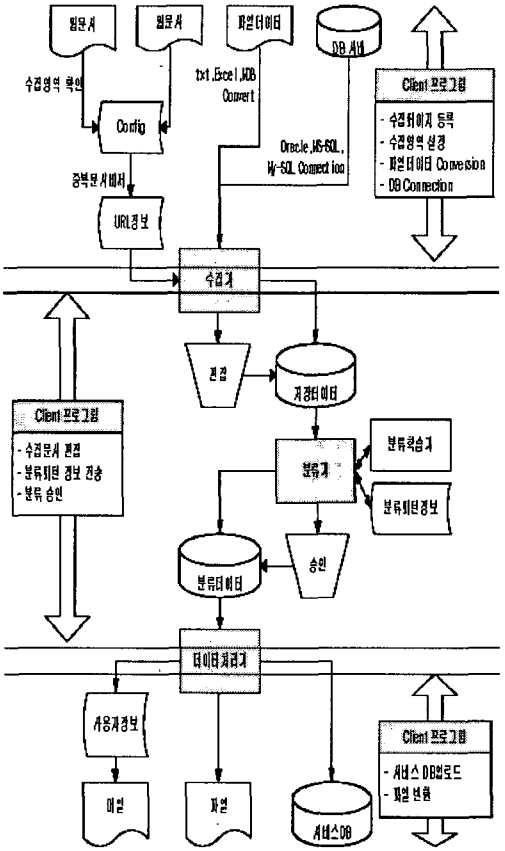
본 연구를 통해 구현된 Infovider 시스템은 2장의 개념을 적용하고 도출된 문제점을 해결하여 업종별/분야별 포탈사이트에서 사용자의 요구를 충족시키는 정보제공 서비스가 가능토록 설계, 구현하였다. 아울러 Infovider는 불특정 다수의 웹페이지에서 정보를 수집하는 목적이 아니므로 인터넷 부하에 따른 로봇 배제 조건을 고려하지 않았으며 광산업 및 부품소재 산업 포탈 서비스에 초점을 맞추었고 업종별 특성에 따른 기능 추가는 지속적으로 진행될 예정이다.

3.1 Infovider 설계

Infovider 시스템은 웹로봇의 구동을 위한 웹페이지 등록 및 수집영역 설정, 수집문서 편집 및 분류된 문서를 승인하는 Client 프로그램과 실제 작업을 수행하는 Server 프로그램으로 설계하였다. Server 프로그램은 수집기, 분류기, 데이터 처리기로 구성하였다. <그림 3>은 Infovider의 전체 시스템 구성을 나타낸다.

Client 프로그램

기존 웹로봇은 단일 사용자 또는 데이터 관리자를 대상으로 제작되어 수집대상 웹페이지를 등록하거나 분류방법 및 키워드를 조정하는 사용자 인터페이스가 빈약한 상황이다.



<그림 3> Infovider 시스템 구성도

간혹 Web-base의 사용자 인터페이스를 갖춘 웹로봇이 있지만 웹로봇 기능의 극히 일부만을 조정할 수 있어 많은 제약사항이 있다.

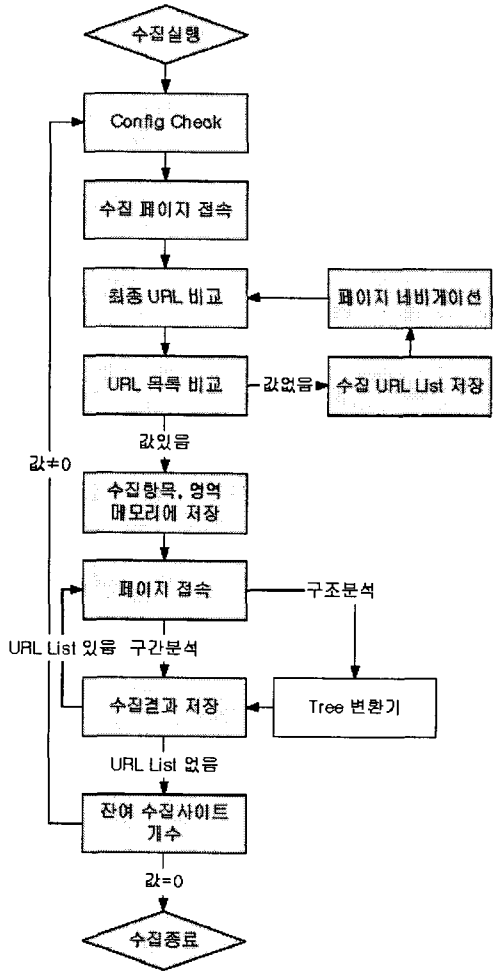
Infovider는 Windows기반의 사용자 인터페이스를 이용하여 웹로봇 전반에 걸친 작동정보 및 옵션을 설정할 수 있도록 설계하였다. 또한 단일 사용자뿐만 아니라 여러 사용자들의 웹로봇 이용목적에 맞춰 사용자 개개인에게 맞춤 정보를 적시에 제공할 수 있도록 하였다.

사용자들은 Client 프로그램을 이용하여

- 수집하고자 하는 웹페이지를 등록하거나
- 웹페이지내에서 수집항목을 지정할 수 있고
- 수집된 문서를 편집할 수 있고
- 분류 키워드를 입력, 조정할 수 있으며
- 분류를 조절, 승인할 수 있도록 하였다.

수집기

Infovider의 수집기는 HTML 문서를 Tree 구조로 해석하는 기법과 텍스트 구간분석기법을 혼용하는 방식으로 설계하였다. Tree 구조 해석기법은 HTML 문서형태에 관계없이 사용자가 지정한 영역을 정확히 수집할 수 있는 장점이 있고, 텍스트 구간분석기법은 본문 중에 HTML Tag가 없는 일부영역에 대해 만족할만한 수집결과를 얻을 수 있다.



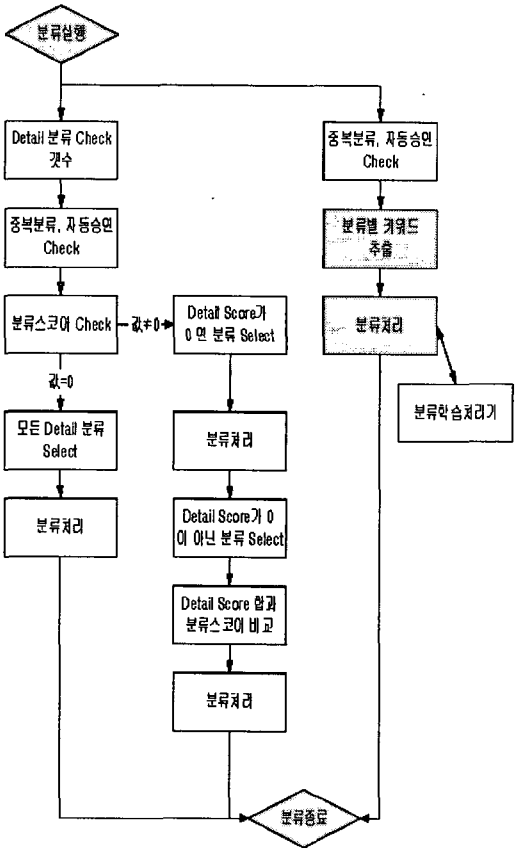
<그림 4> Infovider 수집기 흐름도

분류기

Infovider의 분류기는 기본적으로 키워드에 의한 분류방식을 채택하고 분류학습기능을 추가하여 고정 키워드에 의한 분류미처리 문서를 줄이고자 설계하였다. 또한 사용자 분류패턴분석방법을 통해 사용자의 성향을 파악하여 요구수준에 보다 근접할 수 있도록 설계하였다.

이는 현재 기술로는 완전 자동화 된 분류 시스템을 구현하기가 곤란하기 때문에 해당 업무 담당자의 검증과정을 거치면서 하나의 데이터에 다중의 분류 코드를 제공할 수 있고, 또 같거나 유사한 내용의 데이터를 서로 다른 정보원에서 추출해 올 경우에도 데이터의 중복성을 배재하기 위한 조치이기도 하다.

향후 문서요약 기능의 구현과 데이터 중복기능의 배제를 위한 기능 등이 추가 될 경우 자동으로 분류되고 업로드 되는 비율이 늘어남으로써 더욱 효율적인 시스템으로 발전 될 것이다.



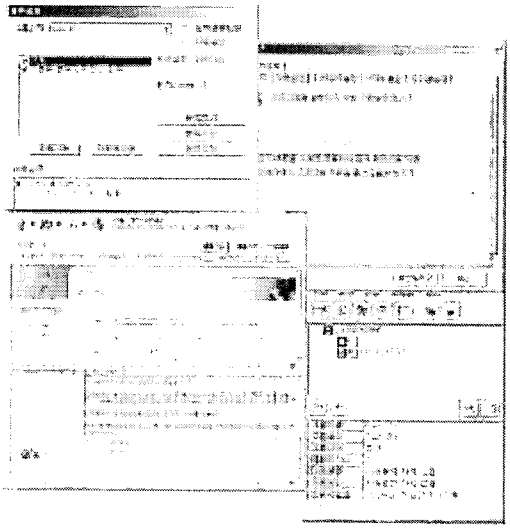
<그림 5> Infovider 분류기 흐름도

데이터처리기

Infovider의 데이터처리기는 분류를 마친 문서에 대해 사용자 정보 또는 요구항목에 대해 정형화된 메일링 서비스를 하거나, Client 프로그램을 통해 Txt, Excel, MDB 형태의 파일로 Export하는 기능을 포함하며, 또 다른 위치에 있는 웹서비스 DB Table에 업로드할 수 있는 기능을 갖도록 설계하였다.

3.2 Infovider 구현

본 연구를 통해 구현된 Infovider는 Client-Server 프로그램으로서 <그림 6>은 Client 프로그램의 분류설정, 수집페이지 등록, 환경설정, 초기화면을 나타내고 있다.



<그림 6> Infovider의 Client 프로그램 화면

Client 프로그램을 이용하여 Infovider의 세부사항을 정의하고 설정하여 사용자는 원하는 업종/분야별 전문정보를 받거나 처리할 수 있다.

Infovider의 Client 프로그램은 Visual Basic 6.0으로 작성되었고, 수집기, 분류기 등의 Server 프로그램은 C++로 작성되었으며, 데이터처리기는 설정사항에 의한 반복작업을 수행하기 위해 Server의 Clon에 등록하여 수행하였다.

4. 결론 및 향후 연구과제

Infovider를 구현함으로써 독자적 수익모델 구현이 곤란한 업종별 포탈 사이트들이 사업 기간 종료 후에도 최소한의 비용과 노력으로 관련 정보들이 지속적으로 갱신되고 추가될 수 있는 기능을 확보 할 수 있을 것으로 기대된다.

또한 업종별 포탈 사이트에서 제공하고 있는 정부정책, 관련법령, 뉴스, 거래알선정보, 입찰정보들을 각 사이트에서 중복적으로 작업을 하는 것이 아니라 Infovider가 이를 해당 분야별 키워드 파일을 통해 분류하고 승인된 결과를 업로드 해줌으로써 한번 만들어진 자료가 다양한 분야에서 효율적으로 활용되어야 한다는 E-비지니스의 기본 철학인 "Make Once Use Many Times"를 실현하고 이에 따라 정보제공 사이트의 서버 부하 및 네트워크의 부하를 줄이는데 기여하게 되는 부수적 효과도 얻을 수 있을 것으로 기대된다.

향후 연구과제로는 문서 요약기능을 추가 함으로써 요약 내용에서 키워드를 추출하고 이를 기반으로 분류기능이 강화됨과 아울러 키워드별 정보제공(SDI : Selective Dissemination of Information)을 보다 효율적으로 할 수 있게 될 것이다.

또한 거래알선 사이트의 경우 멀티케스팅(Multi-Casting)을 한후 메타 검색 사이트에서는 중복된 데이터들이 많이 발생하는 모순을 나타내는데 이러한 데이터의 중복을 효율적으로 제거할 수 있는 기능도 추가로 연구되어야 할 과제이다.

참고문헌

[1] PriceWaterhouseCoopers, "Portals & Knowledge Management" , Technology Forecast 2002~2004, vol. 1, 2002

[2] 한광록, 선복근, 한상태, 임기욱, "인터넷 문서 자동 분류 시스템 개발에 관한 연구", 한국정보처리학회 논문지, 7권 9호, 2000

[3] 한국 인터넷정보센터, 인터넷 통계 Homepage, http://www.kmic.net

[4] 신진섭, "웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델", 건국대학교 대학원 박사학위논문, 2000

[5] 박규석, 이충석, 김 성, "서버 부하를 고려한 동적 로봇에이전트 시스템의 설계 및 구현", 한국정보처리학회 논문지, 7권 11호, 2000

[6] 한국과학기술원부설연구개발정보센터, "웹 로봇에 관한 연구", 연구개발보고서, 1997

[7] 한국과학기술정보원, "산업정보화 관련DB의 공동활용체계 구축방안에 관한 연구", 연구개발보고서, 2002.12

[8]정상호, "JSP를 이용한 검색엔진의 설계 및 구현", 울산대학교 대학원 석사학위 논문, 2000