# [솔루션 발굴]
# N-DataCrawler

## 한국소프트웨어진흥원
## 공개SW기술지원센터

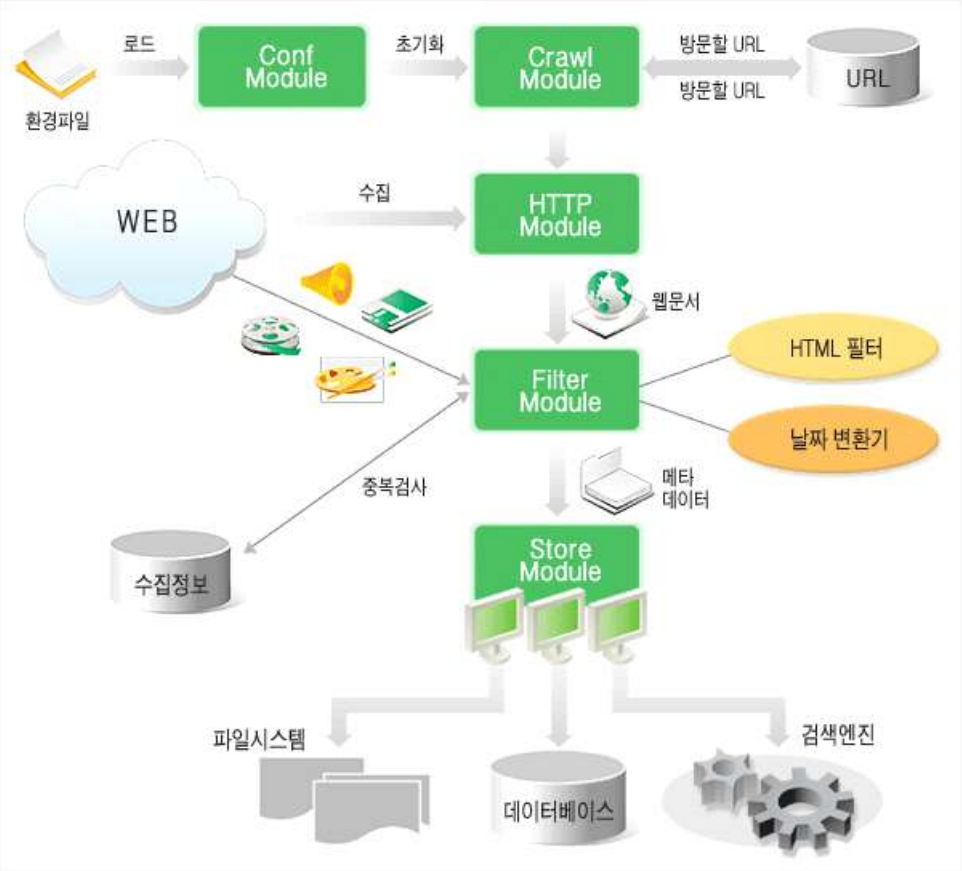
# <Revision 정보>

| 일자 | VERSION | 변경내역 | 작성자 |
|---|---|---|---|
| 2007. 7. 30. | 0.1 | 초기 작성 | 김용규 |

# 기업 / 제품 정보

| 회사명* | (주)다음기술 | 웹사이트* | http://www.ntechs.com/ |
|---|---|---|---|
| 주소* | 서울시 강남구 논현동 86-6 | | |
| 연락처* | 02-541-0840 | E-MAIL* | webmaster@ntechs.com |
| 솔루션 및 서비스명 | N-DataCrawler | | |
| 솔루션 및 서비스 설명 | O 제품 소개<br><br>-. N-DataCrawler는 (주)다음기술에서 개발한 웹 로봇으로 규칙기반의 환경 파일을 정의하여 수집한 문서에서 필요한 메타 정보만을 추출해 주는 지능형 에이전트 시스템이다.<br><br>-. 690417-1018011 N-DataCrawler는 메타 정보 추출용 웹 로봇뿐 아니라 웹 포탈 구축용 일반 웹 로봇, 다양한 검색 사이트를 통합적으로 연동하여 하나의 검새 창에서 제공하는 메타 검색엔진 기능도 별도로 제공한다.<br><br>-. 인터넷의 발달과 함께 끝없이 생성되는 정보의 홍수 속에서 사람이 직접 수 많은 웹 사이트를 방문할 필요 없이 외부에서 필요한 정보를 자동으로 수집해주므로 시간과 비용을 절약하여 기업의 경쟁력을 높여준다.<br><br>-. 멀티쓰레드, HTTP, Socket, 자동 로그인 처리, Javascript 처리 등 고도의 에이전트 기술을 접목하여 강력한 기능에 고성능으로 온라인 정보를 수집해준다.<br><br>-. 수집한 데이터는 규칙에 따라 처리된 후 데이터베이스, 검색엔진 혹은 그 밖의 외부 응용 프로그램과 자동 혹은 프로그램으로 연계하여 이용된다. | | |

0 구성도




0 기능적 특징

- 웹 문서, 게시판, CGI 등 다양한 온라인 문서 수집 지원
- 첨부 파일 및 이미지 등의 선별적 다운로드 기능
- 자동 로그인 지원(Cookie, Cookie2, SSL, Session, Web Server 보안)
- Form 자동 실행 및 결과 페이지 수집
- Javascript 처리 지원
- HTML Header/Body 정보 선별적 수집
- 수집한 문서에서 규칙에 따라 메타 정보만을 추출
- 스크립트 형식으로 규칙을 정의
- 데이터베이스 및 검색엔진과 연동하여 수집 정보 색인 및 활용
- 웹 포탈용 웹 로봇, 통합 메타 검색 엔진 별도 제공


0 데이타크롤러 세부기능

| | |
|---|---|
| 파일 다운로드 | HTML 파일에 첨부된 동영상, 이미지, 기타 바이너리 파일을 HTTP와 FTP를 통하여 다운로드할 수 있다. |
| 로그인 후 수집 | 일반적인(Cookie, Session, 웹 서버 등) 로그인 과정을 거쳐야 볼 수 있는 문서를 자동으로 로그인 하여 수집할 수 있다. |
| Form 자동 실행 | Form을 자동으로 실행하여 나오는 결과를 수집할 수 있다. |
| Javascript 처리 | Javascript 함수를 호출하는 경우 환경파일에 등록된 내용에 따라 자동으로 수행한 결과를 수집할 수 있다. |
| 선별적 수집 | 사이트의 모든 URL을 수집할 수도 있고, 특정한 메뉴 밑에 있는 URL만을 수집할 수도 있습니다. 또한 수집할 필요가 없는 URL의 형태를 지정하여 수집 대상에서 제외시킬 수도 있다. |
| 선별적 데이터 추출 | 문서 전체를 데이터로 만들 수도 있고, 필요한 메타 데이터 - 제목, 본문, 날짜 등 - 만을 추출하여 데이터로 만들수도 있다. |
| XML 자동 변환 | 수집하여 추출한 정보를 XML 형태로 변환하여 제공할 수 있어서 XML 응용 시스템과 쉽게 연동이 가능하다. |
| 날짜 포맷 지정 | 추출한 데이터가 날짜인 경우, 원하는 포맷으로 쉽게 변환할 수 있어서 별도의 작업 없이 날짜 포맷을 한가지로 통일할 |

| | | |
|---|---|---|
| | | 수 있다. |
| | 링크 정보 자동 관리 | 문서에 나타나는 링크 정보를 보존할 수 있으며, 문서 내에 URL www - 혹은 http://로 시작하는 스트링 - 이 있으면 그것을 하이퍼링크로 자동으로 생성할 수 있다. |
| | 이미지 보존 | 문서에 나타나는 이미지를 보존할 수 있습니다. 단 이미지 자체를 다운로드 하는 방식이 아니라, 이미지의 링크를 보존하는 방식이다. |
| | 문자 세트 자동 변환 | 수집하는 문서의 문자 세트를 설정하면 손쉽게 Unicode등 다른 문자 세트로 자동으로 변환할 수 있다. |
| | 갱신된 문서만 수집 | 매번 모든 문서를 다시 수집하는 방식이 아니라 이미 수집된 목록을 보존하고 갱신된 문서나 새로 추가된 문서만을 수집할 수 있다. |
| | BR 태그 삽입 | BR 태그를 자동으로 삽입할 수 있으므로, 문서의 형태를 보존하여 수집할 수 있다. |
| | 단순 분류 및 자동 분류 가능 | 수집하는 정보를 정의된 환경 파일에 따라 카테고리 매핑으로 단순 분류하거나 자동 분류 모듈을 연동하여 정해진 지식맵에 수집한 정보들을 자동으로 분류할 수 있다. |
| | 사이트 관리도구 | 웹 기반의 사이트 관리도구를 제공하여 수집대상 사이트와 수집 일정을 웹 상에서 손쉽게 관리할 수 있다. |
| | 웹 포털 및 메타 검색 도구 | 웹 포털을 구축하기 위한 사이트 전체 웹 페이지 수집 기능과 다양한 검색 사이트를 하나로 통합하여 검색해 주는 통합 메타 검색 기능을 별도로 제공한다. |
| | 게시판 형태 수집 | 게시판이나 방명록 형태의 웹 페이지를 연동하여 수집할 수 있습니다. 즉 리스트 상에 있는 정보와, 리스트를 누르면 나오는 실제 페이지의 정보를 통합하여 하나의 데이터로 만들 수 있습니다. 다운로드한 파일은 파일시스템에 저장 된다. |
| **시스템운영환경** | **O/S (지원가능)** | Linux, Unix, Windows 기반 사용 가능 |
| | **Database** | Oracle, MS-SQL, Mysql등 가능 |

| | | |
|---|---|---|
| | H/W 최소사양 | HDD 1G이상, CPU P3 500Mhz 이상, 최소128MB RAM |
| | 기타 App | Java, JSP, ASP, PHP, C |
| 기타환경 | | |

## 첨부 #1 :

- 제품소개서 & 참조사이트

http://www.ntechs.com/