해양레저정보 데이터베이스 구축을 위한 웹 탐색 컴포넌트 설계

최홍석*·정성훈*·안성환*·임재홍**

*한국해양대학교 대학원 · **한국해양대학교 전파·정보통신공학부 부교수

Design of Web Retrieval Component for Marine Leisure Information Database

Hong-Seok Choi*·Sung-Hun Jung*·Seong-Hwan Ahn*·Jae-Hong Yim**

*Department. of Electronics & Communication Engineering, Graduate School of National Korea Maritime University

**Division of Radio and Information Communication Engineering, National Korea Maritime University, Busan 606-791, Korea

E-mai l: guyver0815@hotmail.com

요 약

해양레저산업의 발달과 레저문화의 수요가 급증함에 따라 해양 안전 및 관련 정보를 제공하는 서비스에 대한 욕구가 증대하고 있다. 그러나 국내에서는 해양레저에 특화된 정보를 제공하는 서비스가 전무한 상황이다. WIPI 기반의 휴대 단말기 상에 디지털 화된 전자해도의 지리정보와 해양레저를 위한 각종 부가 정보를 제공하는 다운로드 형태의 콘텐츠를 개발하는 프로젝트의 일환으로 전자해도 및 부가 정보 DB를 구축하여 요구되는 콘텐츠를 제공하는 서버(CPS; Contents provider Server)가 필요하다. 본 논문에서는 수요자가 개인휴대단말기를 통해 해양레저정보를 요구했을 때 CPS가 정보를 제공할 수 있도록 예상되는 요구 정보를 데이터베이스화하는 웹 탐색 컴포넌트를 설계하여 각종 웹 상에서 시시각각으로 변화하는 정보들을 실시간으로 파싱하고 분류하는 웹 에이전트의 컴포넌트를 개발하고자 한다.

ABSTRACT

According as marine leisure industry has developed and the demand of leisure culture has increased rapidly, a desire about service which supply marine safety and connect marine information is enlarging. We wish to develop contents of download form that supply geographic information of Electronic Navigational Chart(ENC) in the marine that is digitalized to carrying along terminal of WIPI base and various informations for marine leisure. For this, DB that offer ENC and additional information should be constructed. Also, we need server (CPS; Contents provider Server) that offer required contents. In this paper, we design web retrieval component which store request information to database. When consumer required necessary information through personal mobile device, CPS can inform that. So, we wish to develop web retrieval agent component that parse informations in various World Wide Webs, and store to database.

키워드

해양레저, 웹 탐색 에이전트, 데이터베이스

I. 서 론

해양레저의 종류가 다양해지고 그 활동인구가 늘어나면서 해양안전 및 해양레저활동을 위한 관련정보에 대한 수요 또한 증가하고 있으며, 2001년 수산업법의 개정으로 인해 전국 연안어촌의 앞바다는 패류채취, 낚시, 스킨스쿠버, 체험활동, 주말어장 등의 다양한 프로그램으로 일반 국민들의 해양레저활동을 위한 장이 되고 있다[1].

해양수산청이나 기상청 등의 웹 사이트에서 제공하는 모바일 서비스를 이용하면 개인휴대단말기로 해양관련정보를 획득할 수 있다. 그러나 모바일 서비스에서 제공하는 것은 HTML 기반의 문서를 WML 기반의 문서로 변환한 것에 불과하

며 사용자가 필요로 하는 정보를 얻기 위해서는 여러 메뉴를 거쳐야하는 번거로움이 있다. 또한 제공되는 모바일 서비스가 각 이동통신사별 단말기의 플랫폼에 적합하지 않아서 사용자가 서비스 자체를 이용할 수 없는 경우도 있다. 이러한 문제는 웹 탐색 에이전트를 이용한 사용자의 요구 정보들을 수집하여 데이터베이스를 구축하고, 사용자 단말기의 플랫폼에 상관없이 정보의 검색이 가능한 WIPI 기반의 다운로드형 콘텐츠를 개발함으로써 해결할 수 있다.

본 논문에서는 수요자가 지리정보 및 기상정보, 낚시정보, 뉴스 등의 각종 정보를 실시간으로 제공받을 수 있도록 해양관련정보를 수집하여 데이터베이스를 구축하는 웹 탐색 에이전트를 설계한다. 논문의 구성은 2장에서 관련연구로 해양레저, 웹 에이전트, robots.txt 파일에 대해서 기술하며, 3장에서는 제안하는 웹 에이전트의 설계에 대해 기술하고, 마지막 4장에서는 결론 및 향후 연구 과제에 대해 기술하였다.

II. 관련연구

2.1 해양레저

해양레저란 통상 물에 접하여 행해지는 모든 레저 스포츠 활동을 일컫는다. 넓은 의미로는 해양레저활동을 영위하기 위한 관련 교육, 시설 및 장비의 생산까지 포함하는 경우도 있다. 해양레저는 크게 행동적인 동적 해양레저와 비행동적인 정적 해양레저로 나누어지며 그 이용 형태에 따라 스포츠형, 친수형, 크루즈형, 낚시로 나누어 정리할 수 있다[2].

2.2 웹 에이전트

웹 에이전트는 로봇 에이전트, 웹 크롤러, 웹 스파이더 등으로 불리기도하며, 웹 서버를 순회하면서 수많은 정보를 수집하는 프로그램이다. 수집한 정보를 사용자에게 보여주는 방법만 다를 뿐, 웹 브라우저와 유사하며 일반적으로 통계 분석, 유지 보수, 미러링, 리소스 발견, 복합적인 사용의 용도로 쓰인다[3][4].

(1) 통계 분석(Statistical Analysis)

웹 에이전트는 웹 서버를 발견하거나 서버의 수를 세는데 사용될 수 있다. 서버 당 문서의 평균수를 포함해서 파일 타입의 분포, 웹 페이지들의 평균 사이즈, 상호 연결성의 깊이 등의 통계를 분석할 수 있다.

(2) 유지 보수(Maintenance)

웹 에이전트는 웹 서버의 유지 보수에 사용될 수 있다. 하이퍼텍스트 구조를 유지하는데 주된 어려움중의 하나가 다른 페이지에 대한 링크가 끊기는 것(dead links)이다. 웹 에이전트는 HTTP 헤더의 Last-Modified 필드를 참조하여 링크의 변경을 알 수 있으며, 이 사실을 서버 관리자에게 알려준다.

(3) 미러링(Mirroring)

웹 에이전트는 웹 페이지의 내용을 복사하는데 사용될 수 있다. 웹 페이지의 서브 트리를 검색하고 로컬에 저장한다. 이러한 미러링의 용도에 특성화된 웹 에이전트를 웹 미러링 툴이라고 한다.

(4) 리소스 발견(Resource discovery)

대표적인 웹 에이전트의 용도로써 데이터베이스와 연동시켜 검색엔진을 동작시키는데 사용된다. 그림 1은 기본적인 웹 에이전트의 동작을 나타낸다. 웹 탐색 에이전트는 초기화된 URL 정보를 바탕으로 HTTP 상에 존재하는 웹 서버의 문서 위치를 파악하고 수집, 분석한다. 그리고 문서에 연결된 문서들을 추출하는 방식으로 동작한다.

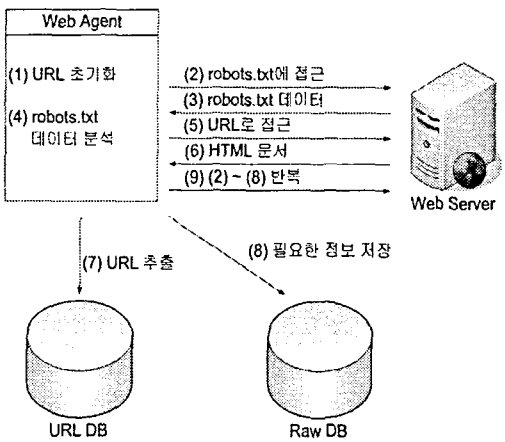

그림 1. 웹 에이전트의 동작 방식

(5) 복합적인 사용(Combined Uses)

웹 에이전트는 위에 언급한 4가지의 용도 중 한 가지 이상에서 복합적으로 사용될 수 있다. 본 논문에서 설계한 웹 에이전트 역시 리소스 발견, 미러링, 통계 분석의 기능을 번갈아 가며 사용한다.

2.3 robots.txt

robots.txt는 로봇 배제의 표준(A Standard for Robot Exclusion)을 수행하는 하나의 방법이다. 웹 서버에서 웹 에이전트에 대한 접근 정책을 명시하는 것으로서, 웹 서버가 원하지 않는 웹 에이전트에 대하여 선택적으로 접근을 허용할 수 있다.

robots.txt의 포맷은 '<field>:<value>'로 기술되며 주석은 '#'을 사용하여 처리한다. <field>는 웹 에이전트의 이름을 명시하는 'user-agent'와 방문하지 말아야할 URL을 명시하는 'disallow'로 나

뉘고 대·소문자를 구분하지 않는다. 그림 2는 microsoft의 robots.txt 파일을 캡처한 것이다.

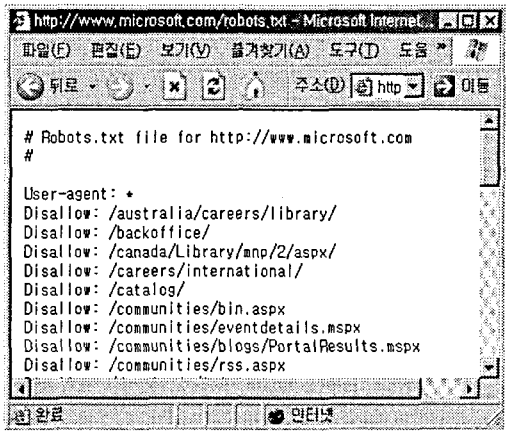

그림 2. Robots.txt file of http://www.microsoft.com

III. 웹 탐색 컴포넌트의 설계

웹 에이전트는 Search 모듈, Parser 모듈, DB 모듈로 구성된다. 그림 3은 설계한 웹 에이전트의 동작 알고리즘이다. 각 모듈은 수집한 정보를 가공, 처리하기 위해 상호 연동이 가능하고, 각 모듈에서 처리된 데이터를 공유할 수 있다.

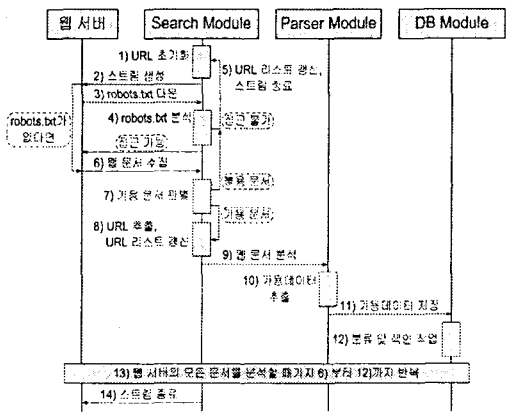

그림 3 Algorithm of web agent

3.1 Search Module

Search 모듈의 가장 큰 목적은 웹 서버에 접근해서 HTML 문서를 가져오는 것, 접근한 웹 문서가 가용문서인지 불용문서인지를 판별하는 것, 마지막으로 URL 리스트를 갱신하는 것이다.

Search 모듈은 URL 리스트에 의해 초기화된 URL을 따라 웹 서버에 접근하여 robots.txt 파일을 검사하고 접근여부를 판단한다. 접근이 가능하다면 웹 서버의 홈페이지에 접근한다. 접근이 불가능 하다면 URL 리스트를 삭제하고 다음 URL 리스트를 읽어온다. URL 리스트를 삭제하는 이유는 접근할 수 없는 URL을 중복해서 읽어오는 일을 방지하고 웹 서버의 부하를 줄이기 위해서이다.

robots.txt 파일을 검사하여 웹 서버의 홈페이지에 접근하면 그 페이지의 HTML 문서를 텍스트 파일로 다운받는다. 텍스트 파일로 다운받는 이유는 가용 문서 판별, URL 리스트 추출, 가용 데이터 추출 등 반복적으로 수행될 파싱 과정에서 사용되는 인덱스 값으로 char 혹은 string 값을 사용할 것이기 때문이다.

HTML 문서를 다운받고 나면 가용 문서 판별을 시작한다. 가용 문서로 판별이 되면 그 문서에서 링크된 URL을 추출한다. 추출한 URL을 URL 리스트와 비교해서 중복된 내용은 삭제하고 새로운 것만 덧붙여 URL 리스트를 갱신한다. 불용문서로 판별이 되면 문서를 삭제하고 그 문서의 URL도 URL 리스트에서 삭제한다.

3.2 Parser Module

Parser 모듈의 목적은 Search 모듈에서 가용 문서로 판별된 문서를 분석하여 가용 데이터를 추출하는 것이다. 해양레저에 관련된 정보에는 낚시, 스쿠버, 섬 여행 등 레저활동에 관한 정보도 있지만 그 외에 일기, 조석, 조류, 파고 등 환경에 관한 정보도 포함된다. Parser 모듈에서는 다운받은 HTML 문서에서 이러한 레저정보나 환경정보를 파싱한다.

그림 4는 (사)한국해양레저보트협회의 웹 사이트에서 제공하는 섬 여행정보에 관한 웹 페이지이다.

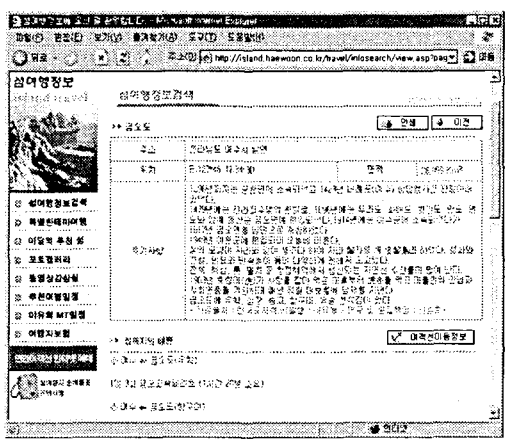

그림 4 Island travel information

HTML 문서를 작성할 때 웹 브라우저에서 보여줄 이미지를 깔끔하게 처리하기 위해 일반적으로 <table> 태그를 많이 사용한다. <table> 태그

의 특징은 테이블을 구성하기 위해 <tr> 태그와 <td> 태그가 반복적으로 사용된다는 것이다. Parser 모듈이 가용 데이터를 파싱하기 위해서는 <td> 태그와 </td> 태그 사이에 < 와 > 에 묶여 있지 않은 문자열의 유·무를 검사한다. < 와 > 에 묶여있는 문자열이 없는 <td> 태그를 시작점으로 하여 파싱을 시작하며, 공백을 나타내는   등의 기타 기호를 나타내는 문자열에 대한 예외 처리는 따로 정의해 둔다.

테이블의 특성상 하나의 <tr> 태그에 여러 개의 <td> 태그가 나열되어 있기 때문에 진행 중인 파싱의 종료점을 </tr> 태그로 정하고 종료점 이후부터 다시 파싱을 시작하도록 한다. 여러 번의 파싱을 반복적으로 수행하여 그 내용을 ini 파일로 출력한다.

3.3 DB Module

DB 모듈의 목적은 *Parser* 모듈에서 출력한 ini 파일을 DB 테이블로 변환하는 것이다. 그림 5에서와 같이 ini 파일은 텍스트 형태로 구성되어 있다. [와] 를 사용하여 그룹제목을 정의하고 그 밑에 = 를 사용하여 변수명과 변수값을 정의한다.

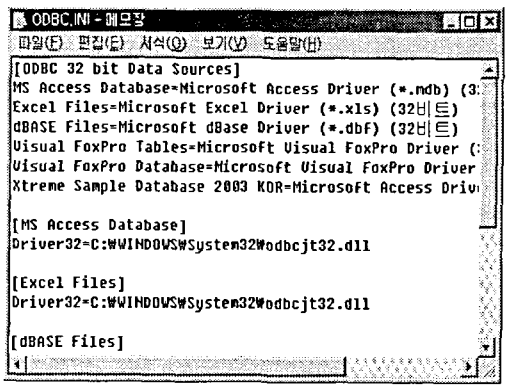

그림 5 ODBC.INI File

DB 모듈에서는 그룹명에 따라 테이블을 분류하고 변수명에 따라 각 칼럼을 구분함으로써 DB 테이블을 완성한다. CPS와 연동되는 메인 데이터베이스가 있음에도 불구하고 DB 모듈을 두어 테이블을 미리 만드는 이유는 1차적으로 인덱싱 작업을 함으로써 사용자에게 좀 더 정확한 데이터를 제공하기 위함이다.

웹 에이전트는 완성된 테이블을 메인 데이터베이스에게 넘겨준다. 또한 수집된 데이터의 정보와 웹 에이전트의 수집활동을 관리자가 시각적으로 확인할 수 있도록 View 기능과 Log 활동을 수행한다.

웹 에이전트는 최신의 데이터를 지속적으로 수집하기 위하여 데몬 형태의 서비스를 수행하며 수시로 웹을 탐색하고 데이터베이스 내의 정보를 갱신하는 역할을 수행한다. 구축된 데이터베이스는 CPS에서 사용자의 모바일 단말기로부터 정보요청이 있을 때 가용정보를 제공할 수 있도록 한다.

IV. 결 론

해양레저산업의 발달과 함께 무선 네트워크를 이용하는 첨단 기기간의 응용기술의 필요성이 높아지고 있다. 본 논문에서는 이러한 응용기술을 접목하고자 WIPI 기반의 모바일 콘텐츠와 CPS를 지원하기 위해서 가용정보를 웹으로부터 추출하고 분류·색인 작업을 거쳐 데이터베이스를 구축하는 웹 탐색 에이전트를 설계하였다.

웹 서버로의 불필요한 접근을 방지하고, 접근하는 웹 서버의 부하를 줄이고자 URL 리스트를 여러 번에 걸쳐 갱신하도록 하였다. 또한 DB 모듈에서 1차 분류·색인 작업을 거쳐 메인 데이터베이스에서 콘텐츠 이용자에게 양질의 정보를 제공하고자 하였다.

향후 본 논문에서 다루지 않은 이미지, 오디오, 동영상 등의 멀티미디어 정보와 HTML 문서로 표현되는 정보에 대한 파싱 작업을 보완하여 탐색의 정확도를 높이고, DB 모듈에서의 분류·색인 작업에 있어서 사용자의 요구 빈도수를 적용하여 *사용자에게 더욱 최적화된 서비스를 제공할* 수 있도록 하고자 한다.

후 기

본 연구는 2005년도 산학연컨소시엄사업단의 지원을 받아 수행되는 과제의 일부임을 밝힙니다.

참고문헌

[1] 해양수산부, http://www.momaf.go.kr
[2] 반석호, 국내 해양레저와 레저선박 산업의 현황 및 전망, 대한조선학회지 제39권 제1호, pp.36-44, 2003
[3] 김동범, 웹 로봇 에이전트의 효율적인 인터넷 정보검색, 한국정보과학회 학술발표논문집 29(2), pp.574- 576, 2002
[4] 신성수, 학술지목차DB(QTOC)를 활용한 해외 학술정보 수집에이전트 시스템, 제20회 한국정보처리학회 추계학술발표대회 눈문집 제10권 제2호, pp.813- 816, 2003