한국항해항만학회지 제26권 제3호(항해항만통합권 제3호), pp.337~344, 2002 (ISSN-1598-5725)
Journal of Korean Navigation and Port Research

선박매매정보 추출 에이전트 시스템 구조 설계에 관한 연구

박 남 규*

*동명정보대학교 유통경영학과 교수

Design for Information Retrieving Agent System for Ship Sale and Purchase

*Nam-Kyu Park**

**Division of Distribution Management, Tongmyung University of Information Technology, Busan 608-711, Korea*

요 약 : 선박매매사이트가 증가하고 있지만, 필요한 정보를 쉽게 찾아주는 에이전트에 관한 연구는 부족하다. 본 연구는 사이버공간상에 흩어져 있는 선박정보를 손쉽게 찾을 수 있는 지능형 에이전트의 설계 구조를 정의하는 연구로서 Wrapper 방식의 설계기법을 사용하였다. 본 연구가 추구하고자 하는 것은 선박매매 전자상거래 시스템의 매매대상 선박을 정보추출 에이전트를 이용하여 선박정보를 선택적으로 추출, 이를 소비자에게 제공하는 선박매매용 정보추출 에이전트의 기능을 시험적으로 설계하는 것이다. 특히 본 연구는 사이버 해운거래시스템이 실현되는 환경 하에서 이를 활성화하기 방안으로 B2B 선박매매 분야에 적용하였다는 점에서 의의가 있다. 본 연구의 결과 에이전트의 프로세스는 URL 읽기, 해당 URL 원시 데이터 가져오기, 태그처리 프로세스, 패턴분석 및 분석내용 저장하기로 구성되어 있음을 파악하였다. 또한 설계전략으로 "URL페이지 읽기" 프로세스와 "소스분석" 프로세스의 연계성 정도에 따라 분리시키거나 연계시킬 수 있음을 파악하였으며, 각각의 장단점이 비교 검토되었다.

핵심용어 : 지능형 에이전트, 패턴 분석, 선박매매 정보 추출 에이전트

Abstract : *Although the number of site for ship sale and purchase are increasing year by year, we can not find the agent system for retrieving the necessary data automatically and efficiently. The object of this paper is to find the design structure of the intelligent agent systems by using wrapper technology. This paper is composed of two contents : design of retrieving system for agent and its application to ship sale and purchase. This paper will be evaluated in terms that its target domain is ship sale and purchase. In the result of the study, agent process is composed of reading URL, taking the source data, processing tag, pattern analysis, and storing the contents analysed.*

Key words : *intelligent agent, pattern analysis, retrieving system, B2B, ship sale and purchase*

1. 서 론

최근 몇 년 사이에 인터넷(Internet) 공간상에서 거래되는 선박매매 사이트의 출현으로 인하여 사이버해운거래 시장이 형성되고 있다. 인터넷은 다양한 정보제공자에 의해 광범위한 분야의 정보를 멀티미디어 형태로 전달한다는 점에서 매우 각광을 받고 있지만 이용자의 증가로 무수한 정보들이 걸러지지 못한 채로 웹(Web)상에 존재하게 되었다. 따라서, 이러한 정보들을 체계적으로 정리하고, 이용자들의 구미에 맞게 가공된 정보제공의 필요성이 대두되게 되었다. 이와 같이 광범위하고 다양한 인터넷 정보 중에서 자신이 원하는 정보자원을 정확한 방법으로 많이 얻어 낼 수 있게 하는 것은 사용자가 인터넷 정보자원을 재활용하여 사용할 수 있다는 점에서 매우 중요한 문제이다. 인터넷에서는 다양하고 방대한 정보가 소비자에게 제시되므로 이에 대한 선별적인 정보제공이 필요하다. 즉, 전통적인 상거래에서 존재하는 점원의 도움처럼 사이버상의 상거래에도 여러 가지 도움을 받을 수 있는 지능을 가진 소프트웨어가 반드시 필요한 것이다.

이러한 문제를 해결하기 위해서 이용자가 원하는 것을 스스로 인지하고 판단하여 요구사항을 해결해주는 "지능적 대리인" 또는 "지능을 가진 도우미"라고 일컫는 지능형 에이전트(Intelligent Agent)의 필요성이 대두되었다. 지능형 에이전트는 그 다양한 기능으로 인하여 인터넷에서 매우 중요한 요소로 자리잡게 되었다.

우리 나라를 비롯해서 선진 해운국에서는 선박매매사이트를 개설, 사이버 공간상에서 거래를 수행하고 있지만 사이트별로 등록된 선박의 수가 많지 않을 뿐 아니라 거래대상 선박의 표현 방식도 사이트 및 국가별로 매우 달라 선박 매매 시 충분한 매매대상 선박정보를 획득하기 어려운 실정이다.

본 연구는 선박매매 시 고객에게 매매대상 선박정보를 제공해주는 선박매매용 에이전트 설계를 목표로 하고 있다. Bargainfinder[13], Pricewatch[25] 등의 쇼핑몰을 대상으로 하는 비교쇼핑에이전트는 상용화되어 운영되고 있지만 선박매매용 정보추출에이전트에 관한 연구는 시도되지 않고 있어 이에 관

* 종신회원, nkpark@tmic.tit.ac.kr, 051-610-8481

한 연구가 필요한 시점이라 판단된다.

본 연구는 선박매매정보를 추출하여 선박구매자에게 제공하는 선박매매정보 추출 에이전트를 설계하는데 연구의 목적을 두고 있다. 본 연구의 목표를 달성하기 위해 제 2장에서는 정보에이전트에 관한 선행연구를 검토하여 에이전트에 관한 연구 성과를 개관하였으며, 제3장에서는 선박매매 거래 방식을 분석하였다. 제4장에서는 선박매매 정보를 제공하는 사이트를 조사하여, 연구대상 사이트를 선정하였다. 제5장은 정보를 지능적으로 추출하는 에이전트시스템을 설계하였으며, 제6장에서는 결론을 도출하였다.

2. 정보 에이전트 선행 연구

본 장에서는 에이전트의 여러 응용 분야 중에서 인터넷상의 정보를 처리해주는 정보 에이전트에 관하여 검토하고자 한다. 인터넷상에서의 정보를 처리하는 지능형 에이전트는 크게 정보검색 에이전트, 정보필터링 에이전트, 정보통합 에이전트, 정보추출 에이전트의 네 가지로 분류할 수 있다.

정보검색 에이전트(Information Retrieval Agent)는 사용자가 원하는 정보를 찾아주는 역할을 수행하며 검색엔진이 대표적인 예가 된다. 검색엔진은 검색로봇(Search Robot), 인덱스(Index), 질의서버(Query Server)의 3가지 요소로 구성된다. 검색로봇은 주기적으로 웹 공간에 존재하는 문서를 수집하여 인덱싱할 수 있도록 도와주며 인덱스는 검색로봇이 모아준 문서를 데이터베이스에 저장하는 작업을 한다. 그리고 질의서버는 사용자의 질의 검색어를 입력받아서 인덱스를 참조하여 검색결과를 출력해준다. 이 부분에 관한 연구는 많은 연구가들에 의해 시도되었다. 검색언어[14],라우팅 프로토콜[15], 하이퍼 텍스트 항해시스템[16], 의미론적 네트워크[17], 네트워크관리 시스템[18], 소프트웨어 에이전트[19] 및 메타데이터 관리[20,21]에 관한 연구가 있었다.

정보필터링 에이전트(Information Filtering Agent)는 사용자의 구미에 맞도록 정보를 가공하고 걸러주는 지능형 에이전트로서 기본적으로 끊임없이 유입되는 정보 중 필요한 것이 무엇이고 필요 없는 것이 무엇인지를 판단하여 필요하지 않은 것은 무시함으로서 사용자에게 알맞은 정보를 제공하여 준다. 정보필터링에서는 사용자가 관심을 가지는 사항에 대한 사용자 정보 프로파일이 중요한 역할을 한다. 정보필터링 과정은 전자메일이나 뉴스그룹의 정보와 같은 정보스트림을 사용자의 정보 프로파일과 비교하여 관심이 있는 정보만을 걸러서 저장한 후 사용자가 볼 수 있게 한다. 사용자는 정보필터링 과정을 거친 결과를 본 후 그것이 실제로 자신이 원하는 것이었는지를 알려주게 되는데 이를 관련성 피드백이라 하며 이 과정을 거치면서 사용자 정보 프로파일을 재구성한다. 다수의 정보필터링 에이전트 시스템이 연구용 또는 상용으로 제시되었다. 정보필터링 에이전트는 어떤 정보를 대상으로 필터링 작업을 하는가에 따라 분류될 수 있는데 크게 웹문서 필터링 에이전트, 상용뉴스 필터링 에이전트, 유즈넷뉴스 필터링 에이전트로 나눌 수 있다. 웹문서 필터링 에이전트의 예로는 WebFilter[1], Webcatcher[2], Point Subscription[3], Smart Marks[4] 등이 있고, 상용뉴스 필터링 에이전트에는 NewsHound[5], Farcast[6], PointCast Network(PCN)[7] 등이 있으며, 유즈넷뉴스 필터링 에이전트에는 NewsClip[8]과 SIFT[9] 등이 있다.

정보통합 에이전트(Information Integration Agent)는 이형질의 여러 정보소스로부터 정보를 검색하여 단일화된 형태로 통합하여 보여주는 작업을 수행한다. 메타 검색엔진이나 비교쇼핑 에이전트 시스템들이 대표적인 예가 된다. 정보통합 에이전트의 필요성은 여러 가지로 기술할 수 있지만 그 중에서도 다수의 정보소스를 사용자가 하나 하나 접근하여 검사하는 노력을 줄여주고 각 정보 사이트에서 사용자에게 불필요하다고 판단되는 것을 걸러주는 점을 들 수 있다. 따라서 정보검색 에이전트와 정보필터링 에이전트의 통합적인 개념을 가지고 있으며 메타검색 엔진과 같은 개념도 정보통합 에이전트와 맥락을 같이 한다고 볼 수 있다. 정보 통합 에이전트의 또 다른 예로는 비교쇼핑 에이전트 시스템을 들 수 있는데 최초의 비교쇼핑 에이전트 시스템인 BargainFinder[10]는 음악CD를 판매하는 온라인상의 전자상거래 사이트를 연결하여 사용자의 구매편의를 도모하였다. 즉, 사용자가 앨범의 제목이나 가수 이름과 같은 검색어를 BargainFinder의 검색창에 입력하면 BargainFinder는 이를 개별 전자상거래 사이트의 검색에 필요한 특정 입력형태로 자동 변환한 후 각 전자상거래 사이트로 보내 검색을 수행하며 출력된 검색결과를 사용자에게 단일화된 형태로 통합하여 보여주는 것이다. BargainFinder와 같은 비교쇼핑 에이전트 시스템은 수많은 전자상거래 사이트에서 판매 중인 제품의 특징 및 가격 등을 사용자에게 한 번에 제공함으로서 사용자의 효율적인 제품구매 의사결정을 도울 수 있다.

정보추출 에이전트(Information Extraction Agent)는 인터넷 문서에서 원하는 특정부분의 정보를 선택적으로 추출해내는 작업을 수행하며 Wrapper라 불리는 추출 규칙을 각 정보소스에 대해 생성하여야 한다[12]. 정보추출 에이전트의 성능은 확장성과 범용성의 정도에 달려있는데 이러한 확장성과 범용성은 서로 다른 인터넷 문서에 대해 추출규칙이 얼마나 유연하게 적용되어지는가에 달려있다. 만일 특정 문서마다 일일이 새로운 추출규칙을 만들어야 하는 경우 같은 추출규칙을 다른 문서에는 적용할 수 없기 때문에 확장성이 떨어진다. 수동적으로 규칙을 구성하는 대부분의 시스템이 이 부류에 속한다. 확장성을 가지기 위해서는 일반적인 프로시저 또는 프로그램이 존재해서 처음 접하는 문서에 대해서도 이 프로그램을 통해 자동적으로 추출규칙을 얻어낼 수 있어야 한다. Wrapper 생성과 인터넷 문서의 정보추출에 대표적인 연구로는 ShopBot[22], HLRT[12], ARIADANCE[23], WHIRL[24] 등이 있다. 이중 ShopBot은 wrapper induction을 비교쇼핑 도메인에 적용하여 자동으로 쇼핑몰 사이트의 상품정보 추출을 위한 wrapper를 학습하는 것으로서의 특징을 갖고 있지만, 규칙성이나 바이어스를 강하게 사용하기 때문에 쇼핑몰이 제한되는 한계를 가지고 있다.

3. 선박매매 거래 방식 분석

3.1 선박매매 거래 절차 분석

선박매매는 전문화된 상거래이며 선박매매중개인(Sales & Purchase Broker)에 의해서 행해진다. 선박매매중개인은 선박의 판매자와 구매자를 위하여 행동하며, 경우에 따라서는 다른 선박매매중개인을 위해 행동하기도 한다. 선박매매시장은 국제시장이며, 선박은 고철용 또는 운항용으로 매매된다. 선박판매시 제시해야하는 정보는 Table 1과 같다.

선박매매의 프로세스는 (1)중개인에게 선박 판매/구매 의뢰 (2)중개인의 중고선 시장 데이터베이스 현황검색 (3)매도인과 매입인의 협상: 선박명세 및 기술사항의 정보교환 (4)매입인의 선박검사 (5)계약체결 (6)지불 (7)협상 후 본선인도의 7단계 주요 프로세스와 이후 본선인수에 따른 추가 프로세스로 이루어진다. 다음 Fig. 1 은 선박매매 거래 처리도이다.

Table 1 Information for ship sale

유 형	내 용
선박의 안정성	선급협회(classification society)
선박 제원	호출부호(call sign), 등록번호, 선박 중량톤수(ship's deadweight), 제원(dimensions) 및 흘수(draught), 건조년월(year of build), 건조장소(place), 조선소(shipbuilder), 용적(cubic capacities), 갑판형상(deck arrangement), 해수저장용량(water ballast capacities), 선창개수(number of holds, hatches), 기기명세(machinery details and builders), 엔진마력(horse power), 속도 및 연료 소비량(speed and consumption), 연료저장용량(bunker capacity), 검사장소(special and classification survey position), 경하배수톤수(light displacement including propeller details)
판매 정보	선박가격 및 검사 장소 및 인도장소

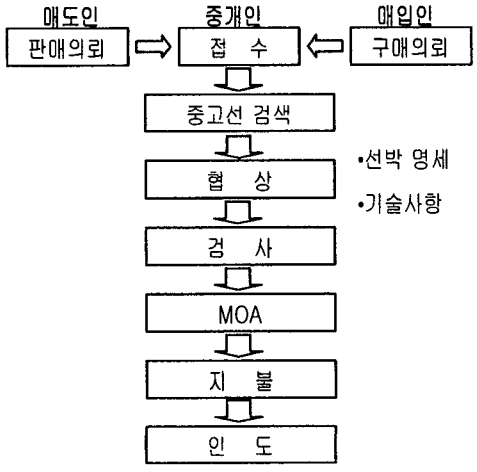

Fig. 1 Procedure of ship sale and purchase

3.2 선박매매정보 홍보사례

선박매매 정보는 해사신문 및 선박매매 사이트를 통해 게재되고 있다. 오프라인상에서의 사례와 전자상거래상에서의 홍보방식에는 근본적으로 차이가 없지만, 전자상거래 사이트는 보다 자세한 데이터를 제공할 수 있다는 장점이 있다(Table 2 참조). 예를 들면 판매선박의 주요명세는 주로 선박의 주요제원인 선박길이, 넓이, 건조년도, 용량, 선박타입, 가격, 국적 및 선급 등이지만 보다 자세한 정보를 원할 경우 링크로 연결된 상세 페이지를 제공하고 있다.

4. 선박매매 사이트 조사

사이버 공간에서 선박매매 전자상거래를 하는 국내외 사이트들이 매년 증가하고 있다. marine-net.com, shippingnet.com, shipbroker.net, marinedigital.com, e-jan.com. ebizmarine.com 등이 대표적인 웹사이트들이다. 이들 웹사이트의 개요 및 주요 서비스를 소개하면 다음의 Table 3과 같다.

선박매매 사이트에서 매매할 선박을 제공하는 형식과 내용은 사이트별로 차이가 있지만 내용은 대동소이하다. Shipbroker.net의 경우 구매할 선박은 참조번호, 선박유형, 간단한 설명, 용량, 건조년도, 가격, 게재일자로 구성되며, 판매대상 선박일 경우는 선박제원 부분에 국적과 선박길이 등 보다 구체적인 정보가 부가되어 구성된다. Shipbroker.net은 회원제로 운영되고 있으며, 가격 및 매매 조건은 당사자끼리 협상하도록 하고 있다. 이에 비해 marine-net.com은 선주 또는 중개인이 판매 선박명세서 및 검사장소를 사이트에 등록하며, 선박판매는 경매방식을 채택하고 있다.

본 연구에서 목표로 하는 시스템은 선박매매정보를 추출하는 지능에이전트를 개발하는 것이므로, 선박매매정보 검색이용이한 shipbroker.net의 사이트를 대상으로 분석하고자 한다.

shipbroker.net은 아래의 운영절차에 의해 선박매매 거래가 이루어진다.

Table 2 Advertisement case for ship sale and purchase

해사신문사례	온라인 사이트 사례[11]
EWL Suriname container 8020 MTDW on 6.581 m Built 1982 Rcikmers Classed GL ice strengthened 127.67mLOA 117.23mLBP 20.1mbeam 2 decks 2 holds 2 hatches 10060 grain 10010 bale 582 TEU 50 reefer CR: 2 x 35T 1 X Deutz Koeln RSBV12M540 6000 BHP 1 thruster 15.5K on 23.5 DO(80.5 CAP) HV(599.5 CAP) DM 15 million	Posted Date :09-17-01 Year Built : 1995 Capacity : 5,778 DWT Vessel Type : Reefer Price : 5,600,000(FOB) Length :97.49m Nationality:Japan Service speed: Sell as Scrap : no Class society :Others

Table 3 Function of web site relating to ship sale and purchase

사이트명	사이트 운영방식	주요 서비스
marine-net.com	일본 이토츄상사, 해사프레스, MOL,K-Line, Hitachi, Fujitsue 등 6개사가 공동출자해 설립한 해운, 조선 포털 사이트이다.	선주 또는 중개인이 판매선박 명세서 및 검사장소를 사이트에 등록한다. 판매는 경매방식을 채택하고 있다.
shipping net.com	현대상선에서 운영하고 있으며, 현대종합상사와의 제휴로 활성화가 되고 있는 해운 조선사이트이다. 선박매매, 용선업무, 화물중개, 운임입찰, 구인구직, 정기선 스케줄 조회, 해운시장정보 서비스 등 다양한 서비스를 제공하고 있다.	매도선박 및 매수선박 조건을 등록하고 검색할 수 있다.
shipbroker.net	Young Sun Trading 사에서 운영하고 있으며, 선박매매, 챠터링, 조선업, 수리업을 취급하는 사이트이다.	선박매매시 판매 및 구매 선박에 관한 주요정보 및 상세정보를 등록하며 이를 검색할 수 있다.
ebizmarine.com	노르웨이 해운사인 파터너쉽사에서 만든 전문포탈사이트로서 챠터링, 선식, 구인구직 등의 서비스를 제공한다.	챠터링 서비스는 선박매매 및 용선서비스로 구분되어 있으며 목적에 따라 선박을 등록하고 화물을 등록하게 되어 있다.

1. 사용자는 Web Browser를 사용하여 Shipbroker.net에 접근한다.
2. Fig. 2 가 나타나면 선박판매희망자와 선박구매희망자는 희망사항을 ADD-SHIP FOR SALE, ADD SHIP FOR PURCHASE를 선택하여 사이트에 등록한다.
3. 선박을 구매하려고 하는 구매인은 판매대상선박정보인 VIEW-SHIP FOR SALE의 정보를 검색한다.
4. 검색결과는 Fig. 3에서 보여지고 있으며, 상세정보를 원하는 사람은 상세화면으로 연결하여 볼 수 있다.

Sale & Purchase-Ship
Add-Ship for sale(Free) View-Ship for sale
Add-Ship for Purchase(Free) View-Ship for purchase

Fig. 2 Menu for registering and retrieving ship sale and purchase

# Ref	Vessel type	Short description	Nationality	Capacity	Length (M)	Built (Year)	Price (USD)	Postec mm/dd/yy
S5651 Pic	Passenger	1994 GREECE REBUILT PULLMAN SEATS 120P SALOON300 P,2BARS	Sweden blt	400 Passengers	161.60	1977	Inquire	09/22 /01

Fig. 3 Display of ship sale and purchase

5. 선박매매용 정보추출 에이전트 설계

5.1 정보추출 에이전트 프로세스 정의

정보추출은 한 문서에서 그 문서의 중심적 의미를 나타내는 특정 구성요소를 인식하여 추출하는 작업을 가리킨다[12]. 정보추출을 위한 프로세스는 다음과 같이 정의된다.

(1) 정보추출 에이전트는 최초의 기동을 위하여 URL 테이블을 참조한다. 따라서 에이전트가 방문해야하는 특정 선박매매 사이트의 URL 테이블의 구축이 선행되어야 한다.

(2) 로봇에이전트는 저장된 URL 테이블을 참조하여 관련 선박매매 사이트를 방문, 매매 정보가 있는 해당 페이지를 인덱싱하여 추출규칙을 생성한다. 특정부분의 필터링을 위해 생성된 추출규칙은 웹페이지 내의 불필요한 부분을 제거한 후 웹페이지의 일정한 구조패턴을 분석하여 필요한 정보를 추출, 그 결과를 테이블에 저장한다. 이러한 구조분석 기반의 선택적 웹페이지 정보추출 방식은 기존의 정보추출 에이전트에서 사용되던 텍스트 기반의 언어적, 시각적 분석기술의 단점을 극복할 수 있도록 설계되어야 한다.

(3) 정보추출 에이전트에 의해 원시데이터의 형태로 추출되어진 선박매매정보는 색인작업을 거쳐 데이터베이스에 저장되며, 검색 및 다양한 사용자요구에 유연하게 대응할 수 있도록 정보검색시스템과 같은 사용자 인터페이스로 구현되어진다.

지금까지 간략하게 설명한 정보추출 에이전트의 기능을 컨텍스트 다이어그램으로 표현하면 다음과 같다.

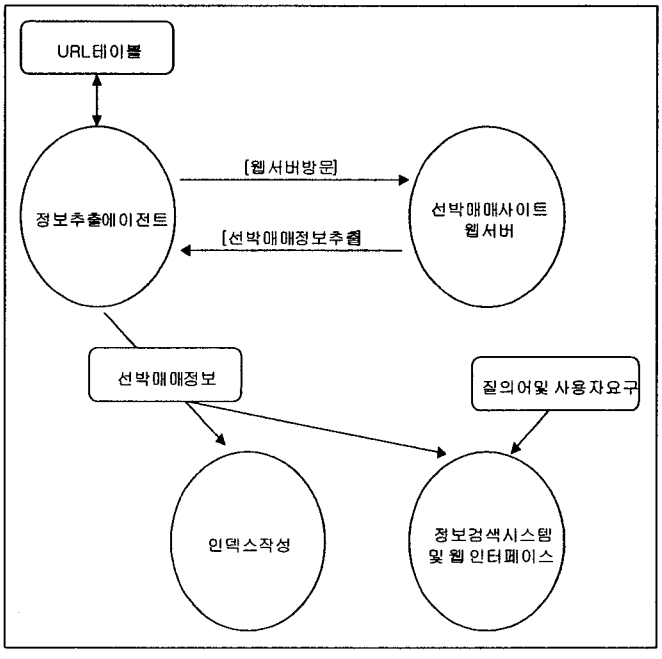

Fig. 4 Context of retrieving agent for ship sale and purchase

Fig. 4 는 정보추출 에이전트 과정을 다이어그램으로 나타낸 것이다. 정보추출 에이전트는 데이터베이스에 등록된 URL을 통해 선박매매 사이트의 웹서버로 이동하여 필요한 문서를 인덱싱하며 구조분석기법을 이용하여 웹페이지 내부의 특정 정보를 선택적으로 추출, 최종사용자에게 제공하는 기술적 방식을 개념적으로 표현하고 있다.

5.2 데이터수집프로세스 설계 전략

Fig. 4 의 원시데이터를 저장하는 선박매매용 에이전트의 기능을 하위 수준의 프로세스로 분할하게 되면, 데이터 수집, 데이터 구조분석, 데이터 저장의 3개 프로세스로 분해된다.

데이터 수집 프로세스는 정해진 시간에 따라 URL이 저장되어 있는 데이터베이스 테이블로부터 URL을 읽어 해당 선박매매 사이트의 웹서버로 정보추출 에이전트가 이동하여 웹페이지를 읽어오는 기능을 수행하는 프로세스이다.

데이터 수집을 위한 프로세스의 설계는 2가지 방식이 가능한데 첫 번째 방식은 URL 정보를 수집하는 모듈과 이를 분석하는 모듈을 완전히 별도로 분리해서 설계하는 것이며, 두 번째 방식은 한 건의 URL을 읽어서 이를 분석한 후 그 결과를 저장한 다음 다른 URL정보를 읽는 것이다.

첫 번째 방식을 사용하게 되면 두개의 프로세스가 별도로 작동하게 되어 프로세스 상호간 영향을 주지 않기 때문에 프로세스의 독립성이 보장되어 전문적 역할을 수행할 수 있는 장점이 있다. 이는 프로세스의 응집력을 높이며, 나아가 프로세스 사이의 결합도를 제로 상태로 만들어 완전한 독립성을 유지할 수 있지만, 수집프로세스가 읽어온 데이터를 일단 수집테이블에 저장해 두어야 하기 때문에 데이터베이스의 용량을 많이 요구한다는 단점이 있다.

두 번째 방식을 사용하여 프로세스를 설계하면, 수집프로세스와 분석프로세스가 모듈커플링을 하게 되어, 수집하여 온 페이지 정보가 그대로 분석프로세스에 이전되게 된다. 이 경우 프로세스의 결합도가 존재하기 때문에 선행 프로세스의 결과가 후행 프로세스에 영향을 주게 된다. 이는 랜덤번호 발생 등을 처리하기 위한 메모리의 사용량을 기하급수적으로 증가시켜 처리 속도를 급격하게 저하시키는 결과를 초래하는 단점은 있지만, 읽어온 원시페이지를 데이터베이스 내에 저장할 필요가 없다는 장점이 있다. 이 경우 수집프로세스가 분석 프로세스의 호출 시 넘겨주어야 하는 인자로는 원시페이지의 내용, 페이지의 URL, 랜덤번호 등이 있으며 랜덤번호의 생성을 위한 배열처리용 메모리가 급격하게 증가된다.

5.3 데이터 구조분석 프로세스 설계

데이터 구조분석 프로세스는 다시 "제거 프로세스", "분석" 및 "데이터 추출"의 3가지 요소로 구성된다. 제거 프로세스는 필요정보를 포함하고 있지 않은 영역의 태그를 제거하는 프로세스로서 다음의 세부 모듈로 구성된다.

가) 제거 프로세스

(1) 선박매매 내용을 연결하는 프로그램 소스에 대해 분석을 시도해 보자. 문서의 시작은 <html>로 시작하며 </html>로 끝난다. 다음은 문서의 머리말로서 <head>..</head>가 나타나며, 이 태그 안에 <title>Shipbroker Net </title>, 환경을 설정하는 <meta name="description" content="Classified listings boats for sale or charter from shipbroker,..."> 및 사용할 스크립트언어인 <script language="JavaScript">..</script>가 정의된다. 또한 사용자 인증 함수인 function check() 등도 여기서 사용된다. 이들이 우선 제거 대상 태그가 된다.

(2) 이들이 제거되면 본문인 <body> 시작되며 이 태그의 속성 bgcolor="white" text="black" link="blue" vlink ="purple" alink="red" 등이 제거되어야 한다.

제거 모듈(face를 제거할 경우의 예)

```
for ($i =0 : $i < $temp_count ; $i++)
$analpage =~ /face='/i:
'face 단어를 공백으로 치환
```

Fig. 5 Deletion Module

나) 분석 프로세스

(1) 이 단계는 패턴매칭 및 필터링 단계로서 웹페이지내 특정정보패턴을 분석하는 모듈과 이를 기반으로 한 테이블내의 속성제거 모듈로 구성된다. 제거 과정을 마친 후 분문에 남게 되는 것은 <table> 태그와 <tr> <td> 태그이다. 이들 태그들은 여러 개가 존재하게 되는데 그중 어느 한 부분이 필요할 것인지 결정해야 한다. 이를 위한 <table> 태그의 속성을 제거하는 작업을 시행한다.

분석 모듈

1. 동일한 패턴 발견
   : 한 화면에 여러개의 테이블형태가 나올 경우, 이 중 동일한 형태의 모듈을 찾고 난 다음,
2. 테이블의 양이 많은 것을 선택
   : 동일한 패턴의 테이블이 여러 개 등장하여도 행의 수가 많은 테이블에 필요정보가 들어 있는 경우가 많이 있기 때문이다

Fig. 6 Analysis Module

(2) 다음 단계로 <table>을 제거하며 최종<tr> <td>구조를 기준으로 필요한 테이블을 판단한다. 아래 테이블의 경우 제거대상 태그는 <p>, <b>, <img>이며 속성은 , font, size, face, color 등이다.

다) 추출단계

(1) 이 단계는 "제거"와 "분석" 프로세스를 통해 도출된 원시코드의 내용을 단순히 화면에 표시하는 단계로서 더 이상의 설명이 필요없다. 이렇게 추출된 정보는 Fig. 8

의 속성제거후의 결과소스와 이의 브라우징 된 화면 Fig. 7로 예시할 수 있다.

P6340	Cargo	Looking for cargo ship	2000~3000 DWT	1989~2001	Inquire	11/90/01

Fig. 7 Actual screen after test of retrieving system

원시 소스코드

```
<tr>
<td width="40" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2">P6340</font><font face="Arial"
size="2"></b></font></b></font></p></td>

<td width="74" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2"><ahref='../list/detail2.asp?ship_no=6340'>Carg
o</a></font></p></td>

<td width="175" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2">looking for cargo ship</font></p></td>

<td width="112" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2">2000~3000</font><font face="Arial"
size="1"><br>DWT</font></p></td>

<td width="110" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2">1989~2001 </font></p></td>
<td width="91" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2">Inquire</font></p></td>

<td width="62" bgcolor="#EFEFEF">
<p align="center"><font face="Arial"
size="2">11-09-01</font><font face="Arial"
size="2"></p></td></tr>
```

속성제거 후 결과소스

```
<tr>
<td>P6340</td>
<td><a href='../list/detail2.asp?ship_no=6340'>
Cargo</td>
<td>looking for cargo ship</td>
<td>2000~3000 DWT</td>
<td>1989~2001</td>
<td>inquire</td>
<td>11-09-01</td>
</tr>
```

Fig. 8 Source code and code after deleting attributes

태그를 제거 한 후, 남는 것을 요약한 것이 Fig. 8 의 속성제거 후 결과소스에 나와있다. 이러한 일련의 구조분석기법을 기반으로 추출된 원시 선박매매정보는 색인화 작업을 거쳐 데이터베이스에 재가공 하여 저장된다.

5.4 시스템구조도

본 시스템의 구조를 그림으로 표현하면 Fig. 9와 같다.

시스템 구조도는 모듈간의 관계를 계층적 구조로 표현하는 소프트웨어 공학의 도구로서 시스템의 전체구조를 표현하는데 유용하다. 시스템구조도를 보면 메인 모듈은 전체 시스템을 관장한다. 메인 모듈의 하위 시스템에 "에이전트 작동시간 설정 모듈", "URL 주소읽기 모듈"," 웹사이트 방문 모듈", "웹페이지 읽기 모듈"로 구성되어 있다.

웹페이지 읽기 모듈은 다시 "구조분석기법정보추출하위 모듈"을 거느리고 있으며, 이 모듈은 "제거", "분석", "추출" 및 "저장 모듈"을 하위 모듈로 거느리고 있다.

5.5 시스템 테이블 구조 제안

(1) URL 저장 파일 구조: 검색엔진용 로봇 에이전트는 첫 번째 사이트에 도달하면 링크된 URL을 따라 다른 사이트로 이동하면서 필요한 정보를 수집한다. 그러나 정보추출 에이전트는 선박매매 사이트만 방문해야 하며 방문한 후에도 필요한 정보만을 선택적으로 추출하여 수집해야하기 때문에 특정 URL이 저장될 테이블이 존재해야한다. URL 테이블 내의 어트리뷰트는 URL 뿐만 아니라 다양한 속성필드를 유지하고 있을 필요가 있는데 그 이유는 향후 범용적 시스템을 구축할 때 유연한 확장성을 확보할 수 있기 때문이다. 따라서 URL 테이블 내부 속성 필드로 문서분류코드를 설정할 필요가 있다. 문

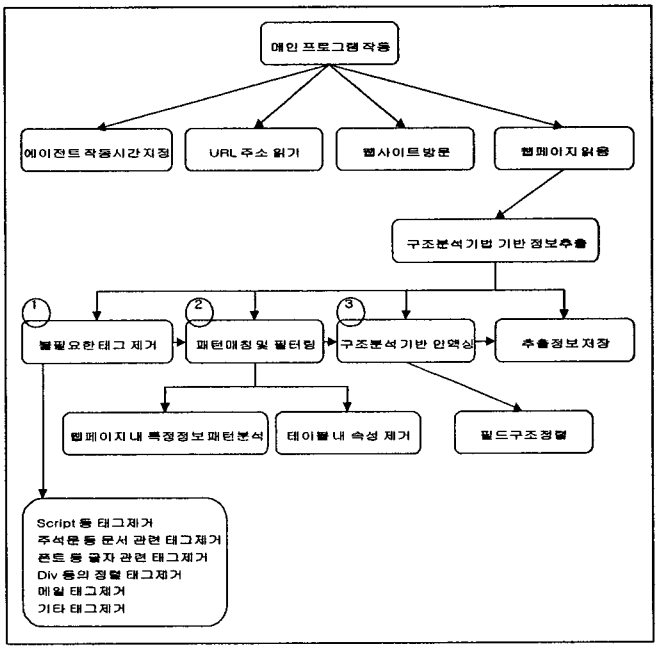

Fig. 9 Architecture of information retrieving agent system for ship sale and purchase

서분류코드는 선박매매용 에이전트가 범용적 에이전트로서 작동하기 위한 이형질의 다양한 수집정보를 체계적으로 분류할 수 있다. 문서분류코드의 테이블은 일련번호 + 문서대분류코드 + 문서의 중분류코드 + URL + 대분류명 + 중분류명 + 사이트명 으로 구성되며, 이 속성중 일련번호가 주요키로서의 역할을 하게 된다.

(2) 선박매매데이터저장파일

선박매매 관련 데이터를 저장할 수 있는 데이터는 번호 + 사이트명 + 사이트 주소 + 소스주소 + 일자 데이터 구조로 정의된다. 선박매매매 데이터의 예를 들면 번호는 임의 번호를 지정하고, 사이트 명은 shipbroker.net으로 정의 된다. 사이트 주소는 http://www.shipbroker.net이 입력되며, 소스 주소는 http://shipbroker.net/list/list1.asp로 정의 된다. 일자는 데이터를 입력한 일시가 기록된다.

6. 결 론

본 연구는 선박매매 전자상거래 시스템의 매매대상 선박을 정보추출 에이전트를 이용하여 선박정보를 선택적으로 추출, 이를 소비자에게 제공하는 선박매매용 정보추출 에이전트의 기능을 시험적으로 설계하는 것을 주요 목표로 하였다. 본 연구의 결과 에이전트의 프로세스는 URL 읽기, 해당 URL 원시 데이터 가져오기, 태그처리 프로세스, 패턴분석 및 분석내용 저장하기로 구성되어 있음을 파악하였다. 또한 설계 전략으로 "URL페이지 읽기" 프로세스와 "소스분석" 프로세스의 연계성 정도에 따라 분리시키거나 연계시킬 수 있음을 파악하였으며 각각의 장단점을 비교 검토하였다. 본 연구는 특정 도메인에 한정했기 때문에 새로운 도메인에 적용될 수 있을 지에 관한 의문은 남아있다. 따라서 향후 도메인의 폭을 넓혀 범용적인 연구를 실시하여 강력한 확장성을 가질 수 있는 정보추출 에이전트의 연구개발이 필요할 것으로 예상된다.

후 기

본 연구는 동명정보대학교 2001학년도 교내학술연구비 지원으로 이루어 진 것입니다.

참 고 문 헌

[1] WebFilter, http://ils.unc.edu/webfilter.

[2] Webcatcher, http://plum.tuc.noao.edu/webcatcher/webcatcher.html.

[3] Point Subscription, http://www.pointcom.com.

[4] Smark Marks, http://www.netscape.com/comprod/smartmarks.html.

[5] NewsHound, http://www.sjmercury.com/hound.htm.

[6] Farcast, http://www.farcast.com.

[7] PointCast Network, http://www.pointcast.com.

[8] NewsClip, http://www.clarinet.com/newsclip.html.

[9] SIFT(Stanford Information Filtering Tool), http://sift.stanford.edu.

[10] BargainFinder, http://bf.cstar.ac.com/bf.

[11] http://www.shipbroker.net.

[12] N.Kushmerick,(1999),"Gleaning the Web", IEEE Intelligent Systems, vol. 14 no.2, pp.20-22

[13] http://www.bargainfinder.co.kr.

[14] P. Ein-dor, I Spiegler,(1995),"Natural language access to multiple database: a model and a prototype", Journal of Management Information Systems vol.12, pp.171-197

[15] D. Flater, Y.Yesha,(1993), "An Information retrieval system for network resources", in: Proceedings of the Workshop on Next Generation Information Technologies and Systems(NGITS'93)

[16] P. Francis,(1996), Ingrid: A self-configuring information grid, in: Proceedings of SIGIR'96 workshop on Networked Information Retrieval(NIR'96)

[17] P. shoval, P. Ein-dor, R. Gilal, I Spiegler,(1996), A meta knowledge base and a search mechanism for distributed, heterogeneous databases, in:Proceedings of the Americas Conferences on Information System

[18] V. Kashayap. A.Sheta,(1998), Semantic heterogeneity, in global information systems: Current Trends and Direction, Academic Press, New York

[19] D Boles et al.,(1996), MeDoc information broker-harnessing the information in literaturd and full text databases in :Proceedings of SIGIR '96 Workshop on Networked Information Retrieval(NIR '96)

[20] C.Hsu, (1990), "The metadatabase project at Rebsselaer", ACM Sigmod Record vol. 20 no.4, pp.758-776

[21] C Hsu, M. Bouziane, L Rattner, L.Yee,(1991), "Information resources management in heterogeneous distributed environments": a metadatabase approach, IEEE Transactions on software Engineering vol. 6, pp.604-625

[22] R. Diirenbose, O. Etzioni, D., Weld,(1997),"A Scalable Comparison-Shopping Agent for the World Wide Web", First International conference on Autonomous Agents, pp. 39-48

[23] J. Ambite, N. Ashish, G. Barish, C. Knoblock, S. Minon, P. Modi, I. Muslea, A. Philpot, S. Tejada, (1998), "ARIADNE: A System for Constructing Mediators for Internet sources", ACM SIGMOD International Conference on Management of DAta, pp. 561-563

[24] W. Cohen, (1998), "A Web-based Information System

that Reasons with Structured Collections of Text", Second International Conference on Autonomous Agents, pp. 400-407

[25] http://www.oilpricewatch.com.

원고접수일 : 2002년 04월 15일
원고채택일 : 2002년 07월 15일