韓國 컴퓨터情報學會 論文誌
第12卷 第1號, 2007. 3.

2007-12-1-1-6

내용기반 검색을 이용한 선박매매 정보추출 에이전트의 구현에 관한 연구

하 창 승*, 정 이 상**

A Study on the Implementation of Information Extraction Agency for Ship Sale and Purchase using Content Based Retrieval

Chang-Seung Ha *, Lee-Sang Jung **

요 약

정보 추출 작업에서의 처리지연은 인터넷 문서의 분류나 표현규칙이 아직 표준화되어 있지 않아 특정 요소에 대한 사용자의 정보 요구를 정확하게 인식하지 못하기 때문이다. 또한 정보추출에 wrapper 규칙을 사용하는 경우 같은 규칙을 서로 다른 문서에는 적용할 수 없는 확장성의 결여와 같은 문제점이 있다. 선박매매와 같이 선박의 거래를 위해 선박가격, 선박 제원, 인도 장소, 검사장소 등의 판매정보만으로도 거래가 가능한 경우에는 선박매매와 관련된 온톨로지(Ontology)를 이용하여 내용기반 검색 (content based retrieval)을 수행하면 선박 매매에 필요한 정보를 선택적으로 추출할 수 있다. 이 방법은 사이트마다 개별적으로 wrapper를 구성하거나 인터넷 문서에서 불필요한 정보를 단계적으로 제거해 나가는 방법을 개선하여 정보 추출 과정을 단순화 시키는 이점을 제공한다.

Abstract

Delay in the process of Information Extraction, IE, is largely due to inability to correctly recognize the user's information requirement of particular search factors. Especially if the wrapper rules are used in a search engine, the search generally fails to classify internet documents properly and efficiently since the application of the same wrapper rules lacks extensibility throughout various types of existing internet document. In case of buying or selling a ship, if the price range, type, place of delivery, inspection site and other information relevant to the sales would be available through the internet for proper retrieval, the sales could more readily succeed by using Ontology relating to sales or purchase information and by selectively searching for the desired information through the content based retrieval system. This system proposes to improve various wrapper systems existing throughout different internet sites and to eliminate unnecessary information tagged on the existing internet documents in order to create a more advanced information retrieval system.

▸ Keyword : information extraction, wrapper, ontology, content based retrieval

• 제1저자 : 하창승
• 접수일 : 2007.2.12, 심사일 : 2007.2.23, 심사완료일 : 2007. 3.19.
* 동명대학교 항만물류학부 조교수, ** 동명대학교 국제통상학과 조교수

Ⅰ. 서 론

최근 기업, 단체 및 개인에 이르기까지 많은 정보 제공자들이 인터넷을 통하여 정보를 제공함에 따라 인터넷을 통한 정보의 검색 또한 점점 일반화되고 있고 인터넷상의 정보양도 급속히 증가하고 있다. 현재 인터넷상에서 색인 가능한 문서의 수가 25억 개를 넘었고, 동적으로 생성되는 인터넷 페이지의 수는 5,500억 개 정도이며, 하루에 새롭게 생겨나는 인터넷 문서도 백만 개를 넘어서고 있다. 정보검색에서 정보량의 급속한 증대는 자신에게 유용한 정보를 찾는데 점점 더 많은 시간을 투자해야 하며 인터넷 문서에서 중심적 의미를 나타내는 특정 구성요소를 인식하여 처리하는 정보 추출 작업을 지연시키고 있다[1].

정보 추출 작업에서의 처리지연은 인터넷 문서의 분류나 표현규칙이 아직 표준화되어 있지 않아 특정 요소에 대한 사용자의 정보 요구를 정확하게 인식하지 못하기 때문이다. 또한 정보추출에 추출규칙을 사용하는 경우 같은 규칙을 서로 다른 문서에는 적용할 수 없는 확장성의 결여가 정보 추출 작업을 어렵게 한다[2].

이러한 문제를 해결하기 위해 기계학습(machine learning) 기법을 이용하는 지능적 에이전트가 필요하다. 지능적 에이전트는 추출을 원하는 텍스트 부분이 개발자에 의해서 표시된 몇 개의 예제 페이지를 학습하거나 정보 필터링을 통해 문서의 의미를 분석하여 카테고리로 분류하고 응용 영역별로 선택적인 정보를 제공하는 기능을 갖는다. 특히 가격이나 상표 등의 내용을 살펴보고 이를 비교하여 구입할 수 있도록 도와주는 비교쇼핑 에이전트에서는 다양한 사이트의 상품을 제품별, 가격별로 비교하여 사용자가 원하는 상품 정보를 빠르게 추출하여 통합하는 기능을 필요로 하기 때문에 더욱 지능적인 학습 기능을 필요로 한다[3].

비교쇼핑 에이전트가 활용되는 국제간 선박매매 거래는 국내외의 여러 선박 포탈사이트를 통해 사이버 공간상에서 수행되고 있다. 하지만 등록된 정보량이 적어 사이트별로 제공하는 정보는 매우 제한적이며 등록된 선박정보의 표현방법도 서로 달라 선박 매매 시 충분한 사전 정보를 제공하기 어려운 실정이다. 또한 사용자가 인터넷으로 여러 선박정보를 비교하면서 거래하고자 할 때 여러 사이트의 정보들이 단일화된 포맷으로 통합되지 않아 각 사이트를 개별적으로 접근해야하는 어려움이 있다[4]. 따라서 선박매매를 사이버 상에서 구현하기 위해서는 선박 포털 사이트로부터 선박 가격이나 선박 제원 정보를 보다 쉽게 추출하고 이를 비교 검토할 수 있는 정보추출 에이전트에 대한 연구가 필요하다.

본 연구는 유연성과 확장성을 갖는 선박매매 정보추출 에이전트를 구현하는데 연구의 목적을 두고 있다. 본 연구의 목적을 달성하기 위해 2장에서는 에이전트 시스템의 선행연구를 검토하였으며, 3장에서는 정보추출 에이전트를 설계 및 구현한다. 4장에서는 구현된 에이전트를 실험 및 평가하고, 5장에서는 결론을 도출한다.

Ⅱ. 에이전트 시스템의 선행연구

2.1 정보추출 에이전트

정보 추출 에이전트는 인터넷 문서에서 원하는 특정부분의 정보를 선택적으로 추출하는 작업을 지원하는 지능형 에이전트 시스템이다. 정보추출 에이전트의 성능은 확장성과 범용성의 정도에 달려 있는데 이러한 확정성과 범용성은 서로 다른 인터넷 문서에서 얼마나 유연하게 추출규칙이 적용될 수 있는 가에 달려 있다. 정보추출 에이전트에 대한 관련 연구는 인터넷 문서에서 원하는 부분 정보를 추출하는 wrapper의 생성과 정보추출의 자동화를 중심으로 이루어졌다.

인터넷 문서의 정보추출을 위한 전통적인 연구에는 ARIADNE[5], STALKER, MORPHEUS 등이 있다. 이러한 시스템은 휴리스틱(heuristic) wrapper를 채용하고 있어 정보 추출방법이 단순하고 특정한 부분 텍스트를 추출하는 데는 정확하고 유용한 특성을 제공한다. wrapper의 반자동 생성에 대한 대표적인 연구로는 ShopBot[6], HLRT[7], XWRAP, WHIRL[8] 등이 있다. 이 가운데 ShopBot은 wrapper induction을 비교쇼핑 도메인에 적용하여 자동으로 쇼핑몰 사이트의 상품정보 추출을 위한 wrapper를 갖고 있다. XWRAP는 wrapper를 학습하기 위해 최소한의 HTML 문서를 계층구조의 트리로 구성하고 의미 부분에 대한 사용자 입력을 받도록 하고 있다[9]. Kushmerick은 여러 wrapper class를 제안하였고 WHLRT는 이 class 중의 하나이다. 이 기법은 인터넷 문서에 국한하지 않고 여러 도메인의 문서로부터 정보추출이 가능하다.

wrapper의 자동 생성을 위해 자연어 처리를 이용하는 연구에서는 자연어로 구성된 문서에서 정보추출 패턴을 학습하는 AutoSlog[10], LIEP[11], CRYSTAL[12, 13, 14] 등이 있다. 이 시스템은 자연어 처리에 초점을 맞추었

기 때문에 인터넷의 문서를 학습하기에는 제약조건이 많으며 이러한 제약조건을 줄이는 것이 인터넷 문서를 학습할 수 있는 관건이 된다. 비자연어 문장의 경우에는 자연어 처리 알고리즘을 적용할 수 없으며 각 문서의 형태에 대한 분석 연구가 필요하다. 이러한 연구에는 WIEN[15], WHISK 등이 있다. WIEN은 인터넷상의 많은 정보들이 상관관계가 있는 데이터로 존재하고 있다고 보고 라벨이 포함된 데이터로부터 wrapper를 자동으로 생성하기 위한 귀납법을 제안하였다. WHISK는 인터넷의 구조화 문서에서부터 비구조화 문서까지 다양한 형태의 문서에서 wrapper의 추출 패턴을 생성하는 학습 기능을 가지고 있다.

2.2 수동형 Wrapper방식의 문제점 분석

정보추출은 수집된 문서로부터 특정한 패턴을 분석하여 필요한 정보를 획득하는 작업이다. 일반적으로 정보추출의 과정은 wrapper이라고 불리는 추출규칙에 의존하여 왔다[16]. wrapper는 주어진 문서에서 문서의 중심적 의미를 나타내는 특정 구성요소를 인식하여 추출하는 규칙이다. 기존의 wrapper는 정보 추출을 위해 각 사이트마다 개별적인 wrapper를 수동으로 생성하였다[17][18].

wrapper를 수동형으로 생성해야하는 경우에는 사이트마다 결과를 출력하는 형태가 다르므로 n개의 사이트에 대해서 n번의 wrapper 생성 작업을 해야 한다. 수동형 wrapper는 정보 소스의 특정 패턴에 맞추어 개별적으로 구성해야 하기 때문에 추출의 정확성은 높지만 유연성, 효율성, 확장성의 측면에서 문제점이 있다.

wrapper을 이용한 정보추출 작업은 본질적으로 확장성을 어렵게 만드는 것 외에 다음과 같은 문제점을 지니고 있다. 첫째, 정보소스의 문서들은 원칙적으로 사람들이 읽기 편하도록 작성되었기 때문에 프로그램이 쉽게 처리할 수 있도록 문서 구성시의 포맷 관행 등에 대한 정보를 제공해주는 사이트는 거의 없다. 둘째, 한 사이트에서 사용된 독특한 포맷 관행이 다른 사이트에도 적용될 가능성이 거의 없기 때문에 사이트가 추가되는 경우 새로운 wrapper가 구성되어야 한다. 셋째, 사이트들이 자주 포맷을 바꾸기 때문에 이전에 만들었던 wrapper가 동작하지 않는 경우가 많다.

wrapper를 자동으로 생성하는 일부 연구에서도 도메인 지식의 획득과 표현의 어려움, 그리고 여러 정보소스 사이에 나타나는 문서형태의 구조적 이질성 때문에 정확한 정보추출이 이루어지지 못하고 있다[9].

또한 기존의 정보 추출 방법은 제거 프로세스를 프로그램 소스에서 구동하여 불필요한 태그들을 단계적으로 제거해 나가는 방법을 사용하고 있기 때문에 처리 절차가 복잡하고 반복된 작업이 많아져서 제거 단계마다 적정성을 평가해야 하는 문제점이 있었다.

이와 같은 문제점 때문에 선박매매와 관련하여 현재 Marine net, Shippingnet.com, Partnership, Shipping Portal 등과 같은 국내외 선박정보 포털 사이트에서 선박매매 정보를 제공하고 있지만 거래 정보의 표현방법이 사이트별로 서로 달라 일관된 정보를 획득하고 분석하기가 매우 어려운 실정이다.

따라서 선박매매와 같이 선박의 거래를 위해 선박가격, 선박 제원, 인도 장소, 검사장소 등의 판매정보만으로도 거래가 가능한 경우에는 선박매매와 관련된 온톨로지(Ontology)를 이용하여 내용기반 검색(content based retrieval)을 수행하면 선박 매매에 필요한 정보를 선택적으로 추출할 수 있다. 이 방법은 사이트마다 개별적으로 wrapper를 구성하거나 인터넷 문서에서 불필요한 정보를 단계적으로 제거해 나가는 방법을 개선하여 정보 추출 과정을 단순화 시키는 이점을 제공한다.

III. 정보추출 에이전트의 설계 및 구현

3.1 정보추출 에이전트의 구성

본 연구에서 개발하는 선박매매 정보에이전트 시스템은 그림 1과 같이 로봇모듈, 추출모듈, 데드링크 테이블(dead link table), URL 테이블, 색인 DB로 구성된다. 로봇모듈은 선박 포털 사이트로부터 선박매매 정보를 수집하는데 로봇모듈의 핵심기능은 검색로봇이 담당한다. 검색로봇은 정해진 시간에 따라 URL 테이블로부터 URL을 읽어 해당 선박매매 사이트의 웹 서버로 접근하여 인터넷 페이지를 읽어온다. 읽혀진 인터넷 문서에서 검색로봇은 하이퍼링크를 확인하고 하이퍼링크의 구조를 따라 다니며 하부 문서를 추출한다. 검색로봇은 다시 그 문서에서 참조되는 다른 사이트의 인터넷 문서들을 순환적으로 탐색하고 관련 정보를 추출하는 작업을 반복한다.

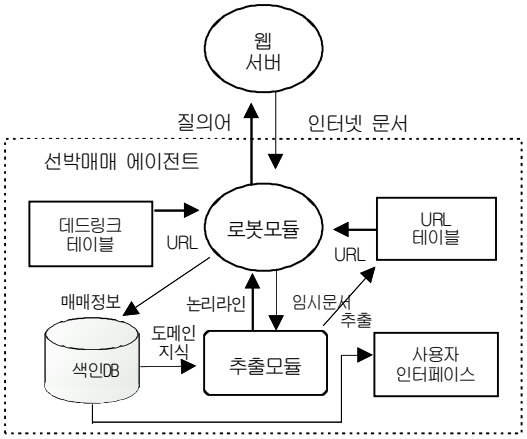

그림 1. 선박매매 정보에이전트 시스템의 구조
Fig 1. Structure of information agent system for ship sale and purchase

추출모듈은 수집된 문서에서 도메인지식, 속성정보, 제어정보를 이용하여 필요한 정보의 위치와 구조, 포맷 등을 파악하여 사용자에게 필요한 정보를 추출하는 기능을 담당한다. 추출모듈에서는 HTML형식의 인터넷 문서에서 관련성이 없는 태그와 속성 및 일반 텍스트 문자들을 제거하여 정보추출의 대상이 되는 논리라인을 생성하고 다시 토큰화(tokenization) 시키는 과정을 거쳐 핵심 구성요소 만을 추출한다.

데드링크 테이블은 문제가 있는 사이트의 방문을 제한하여 불필요한 방문으로 인한 병목현상과 데드 링크된 사이트를 배제시켜 검색 결과의 신뢰성을 높일 목적으로 사용되는 파일이다. URL 테이블은 검색 로봇이 방문할 사이트의 초기 URL을 제공하며 로봇이 접근한 선박매매 사이트와 관련이 있는 사이트의 URL을 추출하여 지속적으로 방문 사이트의 정보를 갱신하고 확장할 목적으로 사용되는 파일이다.

3.2 정보추출 에이전트의 설계

3.2.1 로봇모듈

로봇모듈의 핵심인 검색로봇의 기본 동작은 그림 2와 같이 초기 URL로 주어진 특정한 웹사이트로부터 시작하여 하이퍼링크를 따라 연결된 다른 모든 웹 페이지들을 방문하는 항해전략을 따른다. 일단 초기 페이지가 선정되면 페이지내의 모든 링크주소들을 큐(queue) 속에 저장하고 다음 URL을 큐에서 선택하여 웹 페이지를 가져오는 작업을 반복하면서 선박매매 관련 정보를 수집한다. 검색로봇의 동작 과정은 다음과 같다.

- 단계 1 : 방문할 URL 주소를 URL 테이블로부터 받아 로봇을 실행한다.
- 단계 2 : 호스트 이름에 따라 데드링크 테이블에 접근한다.
- 단계 3 : 데드링크 테이블의 내용을 읽어 들여 주기억장치 내에 벡터 테이블을 구성한다.
- 단계 4 : 벡터 테이블의 내용을 분석해 자신이 들어갈 수 있는 사이트인지 아닌지 비교한다.
- 단계 5 : 자신을 배제하는 사이트가 아니면 해당 사이트에 접근한다.
- 단계 6 : 접근한 사이트의 문서를 전달받아 임시 파일로 저장한다.
- 단계 7 : 추출모듈을 호출하고 임시파일을 전달한다.
- 단계 8 : 추출모듈로부터 전달받은 핵심 구성요소를 저장정보로 색인화 하여 색인 DB에 저장한다.
- 단계 9 : URL 테이블에서 다음에 방문할 URL을 전달받고 단계 4에서 단계 8까지의 과정을 반복한다.

검색로봇의 동작에 필요한 주요 클래스로 URL객체, InputSttreamReader객체, BufferedReader객체 등이 사용된다. 먼저 URL 테이블에 등록된 도메인 이름을 이용하기 위해 URL 클래스 객체가 사용 되었으며 URL 클래스 객체를 네트워크상의 논리적인 입력 스트림(stream)으로 이용하기 위해 InputStreamReader객체와 입력 스트림을 버퍼링 하기 위해 BufferedReader객체를 사용하였다.

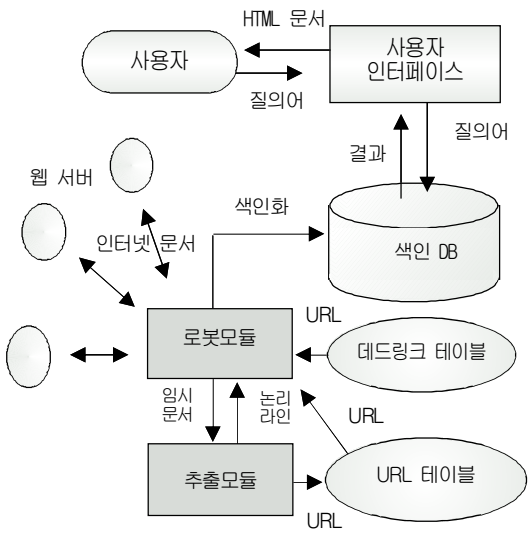

그림 2. 검색로봇의 정보 수집을 위한 동작
Fig 2. Operation for information retrieve of search robot

사용자 인터페이스는 사용자와 색인 DB 사이에서 상호 동작을 통해 정보의 입출력을 지원하는 모듈이다. 인터페이스 모듈은 색인 DB의 IP 어드레스와 포트번호를 객체화하고 사용자가 질의한 검색어를 SQL언어로 캡슐화하여 원격에 분산된 데이터베이스 서버와 연동된다. DB 드라이버와 내부 연결에 이용되는 JDBC 드라이버로 Thin Driver가 이용된다. 그러나 색인 데이터베이스와 사용자간의 실제적인 연결은 "java.sql" 패키지와 "java.net" 패키지에 포함된 클래스의 API 메소드를 통해서 이루어진다.

3.2.2 추출모듈

인터넷에서 사용되는 구조화 문서는 질의를 만족시키는 데이터베이스의 일부 내용을 일정한 출력형식으로 표현한 형태를 지닌다. 즉 실제 데이터와 그 데이터가 나타내는 의미를 메타 데이터를 통해 명시적으로 표현하는 문서 형식이다. 여기서 데이터의 의미를 기술하는 메타 데이터를 라벨(label)이라고 하며 라벨과 데이터의 값이 쌍으로 구성되어 있는 문서를 라벨 문서라고 한다. 이제까지 대부분의 정보추출 에이전트는 이러한 라벨 문서에서 사용자에게 필요한 정보를 라벨을 이용하여 구조화된 부분의 정보를 추출하도록 하였다. 라벨을 이용할 경우에는 추출하고자 하는 라벨과 대응되는 데이터의 위치와 구조, 포맷 등을 나타내는 wrapper를 필요로 하는데 wrapper는 해당 문서에 종속성이 강하여 추출 규칙의 형식을 별도로 기술해야 했다.

본 논문에서는 기존의 연구와는 달리 라벨을 이용하지 않고 내용에 기반한 검색을 통해 필요한 정보를 확인하고 관련 정보를 그림 3과 같이 추출 한다.

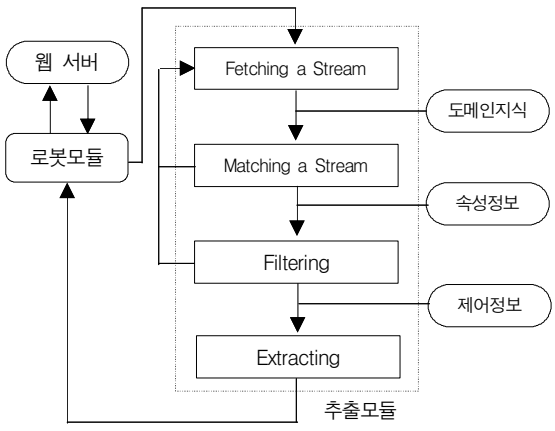

그림 3. 추출모듈의 정보 추출을 위한 동작
Fig 3. Operation for information retrieve of extraction module

정보 추출과정은 선박매매 분야의 도메인 전문가로부터 선박매매와 관련된 온톨로지(ontology)를 입력받아 그림 4의 도메인지식을 구성한다. 다음으로 로봇모듈에서 제공된 인터넷 문서를 한 라인씩 가져와 도메인지식에 입력된 도메인지식과 비교한다. 만약 일치하는 문자열을 지니고 있으면 불필요한 HTML 태그들을 해체하는 제거(empty)과정을 거쳐 HTML 태그가 제거된 문자열 스트림(string stream)인 논리라인을 생성한다. 논리라인에서 정보의 신뢰성을 높이기 위해 그림 5의 속성정보들을 이용한다. 속성정보는 도메인지식을 보조하기 위한 정보로써 도메인지식과 문자패턴이 일치하더라도 속성정보를 충족하지 않으면 그 논리라인은 버려진다. 속성정보까지 충족되면 그림 6의 제어정보를 이용하여 온톨로지와 연결된 추가적인 데이터들을 추출한다. 제어정보는 문서에서 관련 데이터들이 존재하는 위치에 대한 추가 정보를 포함하고 있다.

Category	Ontology1	Ontology2	Ontology3	Ontology4	··
Ships for purchase	Tanker	Bulk	Container	Cargo	··
Ships for sale	Bulk	Container	Tug	RoRo	··

그림 4. 도메인지식 테이블
Fig 4. Domain knowledge table

Category	Attribute1	Attribute2	···
Ships for purchase	font	href	···
Ships for sale	font	href	···

그림 5. 속성정보 테이블
Fig 5. Attribute information table

Category	# of Skip	# of Read	···
Ships for purchase	2	6	···
Ships for sale	3	6	···

그림 6. 제어정보 테이블
Fig 6. Control information table

추출모듈 도메인지식 및 관련정보를 사용하여 선형(vessel type), 선박의 크기(capacity), 건조년도(year built), 선가(vessel price), 게시일(post date) 등의 선박매매정보를 추출하는 세부과정은 다음과 같다.

- 단계 1 : 로봇모듈에서 제공된 인터넷 문서를 한 라인씩 잘라서 읽기 위해 BufferedReader 클래스 객체를 생성한다.
- 단계 2 : BufferedReader 객체에서 제공하는 readLine() 함수를 이용하여 입력 스트림이 Null이 될 때까지 반복해서 문서를 읽어 들여 문자열 인스턴스를 생성한다.
- 단계 3 : 데이터베이스의 테이블에서 도메인지식, 속성정보, 제어정보를 읽어 들여 벡터테이블을 생성한다.
- 단계 4 : String 클래스의 indexOf()함수를 이용하여 문자열 인스턴스와 도메인지식 벡터테이블의 문자열과 패턴 매칭을 수행한다.
- 단계 5 : 단계 3에서 일치된 문자열 인스턴스의 HTML 태그들을 해체하기 위해 empty()함수를 호출하여 논리라인을 생성한다.
- 단계 6 : 선택된 논리라인의 신뢰도를 높이기 위해 속성정보 벡터테이블을 이용하여 관계없는 논리라인은 제거하는 필터링 작업을 수행한다.
- 단계 7 : 제어정보 벡터테이블을 이용하여 스킵(skip)할 논리라인과 반복해서 읽어 들일 논리라인을 결정한다. 최종적으로 논리라인이 확정되면 토큰화 과정을 거쳐 결과를 로봇모듈에 전달한다.

empty()함수는 HTML 태그들을 제거하기 위해 구성한 사용자 정의함수이다. 이 함수는 의미 있는 문자열만을 추출하기 위해 문자열 인스턴스의 문두와 문미에서 태그의 시작 기호인 '<'을 인식하고 걸러내는 작업을 반복적으로 수행한다.

IV. 실험 및 평가

본 연구에서 구현한 선박매매 정보추출 에이전트는 JSP 언어를 사용하였고 색인데이터베이스 및 관련 데이터베이스는 오라클 8.1을 사용하여 JDBC를 통해 연동하였다. 선박매매 사이트에서 제공하는 형식과 내용은 사이트별로 차이가 있지만 내용은 대동소이하기 때문에 시스템 성능의 평가는 'Young Sun Trading'에서 운영하고 있는 shipbroker.net을 대상으로 실험하였다.

Shipbroker.net의 경우 판매할 선박과 구매할 선박을 구분하여 정보를 제공하는데 각 페이지의 내용형식은 참조번호, 선박유형, 간단한 선박정보, 용량, 건조년도, 가격, 게재일로 구성된다. 그림 7은 Shipbroker.net 사이트의 선박구매 정보를 보여주는 화면이다.

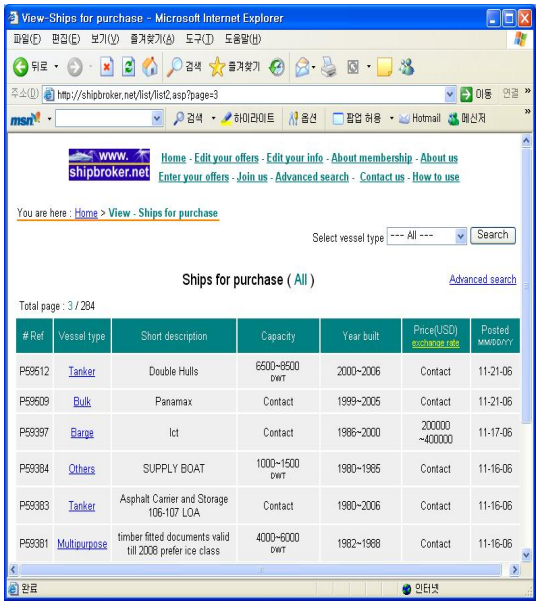

그림 7. Shipbroker.net 사이트의 선박구매 정보
Fig 7. information for ship for purchase of shipbroker.net

성능 평가는 검색된 결과의 정확율(precision ratio)과 재현률(recall ratio)을 측정하는 방법을 사용하였다. 일반적으로 정확율과 재현률은 정보검색 성능 평가의 척도로서 사용되며 식 (4.1), (4.2)와 같이 정의된다.

$$\text{정확율} = \frac{\text{추출된 문서 개수}}{\text{검색된 전체 문서들 개수}} \quad (4.1)$$

$$\text{재현률} = \frac{\text{추출된 문서 개수}}{\text{대상에 존재하는 문서들 개수}} \quad (4.2)$$

shipbroker.net사이트의 ship for purchase(All)을 대상으로 본 연구에서 구현한 선박매매 정보추출 에이전트를 각 페이지의 내용에 접근시키는 실험은 그림 8과 같으며 실험에서 표 1과 같은 결과를 얻었다.

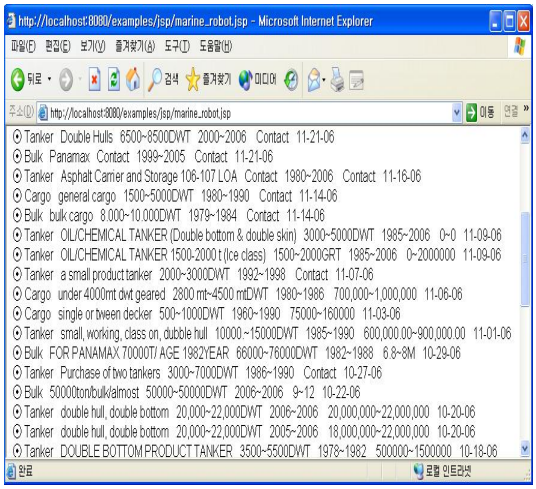

그림 8. Shipbroker.net 사이트의 선박구매 정보 추출 결과
Fig 8. screen shot of extracting the ship information for purchase in shipbroker.net

표 1. 실험 결과
Table 1. An experiment result

Sample Site	# of Ontology		# of target items(T)	# of extracted items(E)	# of correctly extracted items(C)	Recall R=C/T	Precision P=C/E
	D	A					
ship for purchase (All)	5	0	2880	2375	1560	54%	66%
	5	1	2880	1584	1584	55%	100%
	16	1	2880	2822	2822	98%	100%
	16	2	2880	2822	2822	98%	100%

D: 도메인지식, A: 속성지식

실험은 Shipbroker.net 사이트의 선박구매 관련 데이터 2880개를 대상으로 하였고 도메인지식과 속성지식의 수를 변화 시키면서 실험결과가 측정되었다.

표 1에서와 같이 도메인지식은 정보의 재현률과 관련이 있고 속성지식은 정보의 정확율과 관련이 있었다. 속성지식의 수와 정확율은 서로 비례하지 않아서 속성지식의 수를 증가시켜도 정확율은 더 이상 개선되지 않았다. 즉 정확율을 향상시키기 위해서 결정력이 높은 한 두개의 속성지식이 필요하며 이 속성지식이 정보 추출의 정확도를 결정하였다. 반면 도메인지식의 수와 재현률은 서로 비례하며 도메인지식의 수가 증가할수록 재현률이 높아졌다. 즉 온톨로지의 수가 16개인 경우 선박유형이 Landing craft와 Warship 같은 특수한 경우를 제외하고는 대부분 정보 추출이 가능하였다.

실험결과 본 연구에서 시도한 내용기반 검색방법이 기존의 wrapper 시스템이나 기계학습형 정보추출 에이전트에 비해 유연성과 확장성이 높으며 정확율과 재현률에서도 높은 성능을 보였다.

V. 결론

규칙성에만 의존하는 wrapper의 여러 문제점을 보완하기 위해 본 연구에는 선박매매와 같이 필수정보를 미리 파악할 수 있는 경우에는 대상의 내용을 인식하는 내용기반 검색을 이용하여 선박매매관련 정보를 추출한다. 내용검색을 위해 온톨로지를 지닌 도메인지식과 신뢰성을 높이기 위해 속성지식을 이용하였다. 이 방법은 수동적으로 규칙을 구성해야 하는 대부분의 시스템에서도 도메인지식을 구성하면 확장성과 유연성을 가질 수 있고 추출프로세스가 처음 접하는 문서에서도 선택적으로 정보를 추출 할 수 있었다. 따라서 본 연구에서는 기존의 시스템에서 발생했던 경직성을 일부 개선하고 wrapper를 작성해야하는 문제점을 보완할 수 있는 새로운 가능성을 보여주었다.

본 연구는 이러한 비교쇼핑 에이전트에서 선박매매와 같이 판매정보를 사전에 확인할 수 있는 경우 각 사이트마다 공통적으로 적용할 수 있는 특정 온톨로지를 사용한 결과 주어진 정보소스로부터 정보추출이 가능하였다. 향후 연구과제로는 각 사이트마다 다른 구조를 갖는 제어규칙을 자동적으로 설정할 수 있는 개인화된 작동학습 에이전트의 연구를 계속하고자 한다.

참고문헌

〔1〕 하창승, 류길수, "사례기반 추론을 이용한 지능형 웹 검색 에이전트의 설계 및 구현", 한국컴퓨터정보학회논문지, 제8권 제1호 pp20-29 2003

〔2〕 최중민, 인터넷 정보추출에이전트, 정보과학회지 18권 5호 pp. 48-53 2000

〔3〕 N. Kushmerick, Gleaning the web, IEEE Intelligent systems, vol.14, no2, pp. 20-22, 1999

〔4〕 박남규, 선박매매정보 추출 에이전트 시스템 구조 설계에 관한 연구, 한국항해항만학회지 제26권 제3호 pp. 337-344 2002

〔5〕 J. Ambite, N. Ashish, G. Barish, C. Knoblock, S. Minton, P. Modi, I. Muslea, A. Philpot, S. Tejada, "ARIADNE: A System for Constructing Mediators for Internet Sources," ACM SIGMOD International Conference on Management of Data, pp. 561-563, 1998.

〔6〕 R. Doorenbos, O. Etzioni, D. Weld, D, "A Scalable Comparison-Shopping Agent for the World Wide Web," First International Conference on Autonomous Agents, pp. 39-48, 1997.

〔7〕 N. Kushmerick, D. Weld, R. Doorenbos, "Wrapper Induction for Information Extraction," International Joint Conference on Artificial Intelligent, pp. 729-735, 1997.

〔8〕 W. Cohen, "A Web-based Information System that Reasons with Structured Collections of Text," Second International Conference on Autonomous Agents, pp. 400-407, 1998

〔9〕 서희경, 양재영, 최종민 "준구조화된 정보소스에 대한 지식기반의 Wrapper 학습 에이전트" 정보과학회논문지 소프트웨어 및 응용 제29권 제1호 2002.2

〔10〕 E.Riloff, Automatically Constructing a Dictionary for Information Extraction Tasks, *Proceedings of the Eleventh Annual Conference on Artificial Intelligence (AAAI-93)* pp. 811-816, 1993

〔11〕 S. Huffman, Learning Information Extraction Patterns from Examples, *Workshop on New Approaches to Learning for Natural Language Processing, IJCAI-95* pp. 127-142, 1995

〔12〕 S. Solderland, CRYSTAL Inducing a Conceptual Dictionary, *Preceedings of 15th International Conference on Artificial Intelligence(IJCAI-95)* pp. 1314-1319, 1995

〔13〕 S. Soderland, D. Fisher, and W. Lehnert., Automatically Learned vs. Hand-crafted Text Analysis Rules, *Technical Report TE-44 at Center for Intelligent Information Retrieval,* University of Massachusetts, 1997

〔14〕 S. Soderland, Learning Text Analysis Rules For Domain-Specific Natural Language Processing, *University Massachusetts Amherst. Department of Computer Science PhD thesis*1997

〔15〕 Nicholas Kushmerick, wrapper Induction for Information Extraction, *Proceedings of 15th International Conference on Artificial Intelligence (IJCAI-95)* pp. 729-735, 1995

〔16〕 N.Kushmerick(1999), "Gleaning the Web", IEEE Intelligent Systems, vol. 14 no.2, pp. 20-22

〔17〕 P. Atzeni, G. Mecca. and P. Merialdo, Semi-structured and Structured Data in the Web: Going Back and Forth, *ACM SIGMOD Workshop on Management of Semi-structured Data* pp. 1-9, 1997

〔18〕 J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. Breunig, V. Vassalos, Template-based Wrappers in the TSIMMIS System, *ACM SIGMOD International Conference on Management of Data* pp. 532-535, 1997

저 자 소 개

하 창 승

1984년 2월 한국해양대학교 항해학과 졸업 (공학사)
1992년 2월 한국해양대학교 전자통신공학과 (공학석사)
2004년 2월 한국해양대학교 전자통신공학과 (공학박사)
1996년 9월 - 2006년 2월 동명대학 부교수
2006년 3월 - 현재 동명대학교 항만물류학부 조교수

정 이 상

1988년 2월 인제대학교 경영학과 졸업 (경영학사)
1991년 2월 부산대학교 경영학과 (경영학석사)
1998년 2월 부산대학교 경영학과 (경영학박사)
1996년 3월 - 2006년 2월 동명대학 부교수
2006년 3월 - 현재 동명대학교 국제통상학과 조교수