국제입찰정보 통합시스템의 설계 및 구현
(The Integration System for International Procurement Information Processing)

윤 종 완 † 이 종 우 †† 박 찬 영 †††
(Jongwan Yoon) (Jongwoo Lee) (ChanYoung Park)

요 약 현존하는 상업용 웹 정보검색 시스템들이 전문성을 갖추지 못하고 있는 이유는 검색된 분야별 정보를 통합하고 가공하는 능력이 부족하기 때문이다. 따라서, 단순 검색이 아닌 실제 사용자가 원하는 웹 상의 의미정보를 추출하고 가공/통합하는 정보통합시스템의 필요성이 대두되었다. 본 논문에서는 분산된 이질의 웹사이트들에서 제공되는 특정분야의 정보를 추출 및 통합하는 정보통합시스템(TIC: Target Information Collector)을 구현하고, 구현된 시스템의 평가결과를 제시한다. 본 논문에서 대상으로 설정한 정보 영역은 국제입찰정보이다. 국제입찰정보는 전 세계 국가의 정부에서 필요로 하는 조달물품 및 서비스에 대한 공개 입찰자료이다. 본 논문에서는 전 세계의 국제입찰정보 제공 원천 사이트에서 공통 특성 정보를 자동 추출하기 위해 HTML 태그간 패턴을 사용한 정보위치지정 방법을 사용하였으며, 정보추출 및 통합을 위한 프레임워크 설계를 통해 큰 부담 없이 모든 원천사이트 별 정보추출 및 통합 코드를 작성할 수 있었다. 또한, 구현된 TIC을 약 8 개월 동안 운영한 결과 매우 단순한 기법을 사용하고도 거의 대부분의 중복정보가 제거된 고품질의 국제입찰정보를 수집할 수 있음을 확인하였다. 본 논문이 기여하는 바는 특정 범주에 속하는 공통정보를 추출 및 통합/가공하는 데에 필요한 시스템 프레임워크를 제시했다는 점이다.

키워드 : 정보통합, 국제입찰정보, 인터넷 검색, 설계, 구현

Abstract The lack of specialties of the existing commercial web search systems stems from the fact that they have no capabilities to extract and gather the meaningful information from each information domain they cover. We are sure, however, that the necessity for the information integration system, not just search system, will be likely to become larger in the future. In this paper, we propose a design and implementation of an information integration system called TIC(target information collector). TIC is able to extract meaningful information from a specific information area in the internet and integrate them for the commercial service. We also show the evaluation results of our implementation. For the experiments we applied our TIC to the international procurement information area. The international procurement information is publicly and freely announced by each government to the world. To automatically extract common properties from the related source sites, we adopt information pointing technique using inter-HTML tag pattern parsing. And through the information integration framework design, we can easily implement a site-specific information integration engine. By running our TIC for about 8 months, we find out it can remove considerable amount of the duplicated information, and as a result, we can obtain high quality international procurement information. The main contribution of this paper is to present a framework design and it's implementation for extracting the information of a specific area and then integrating them into a meaningful one.

Key words : Information Integration, international procurement information, internet search, design, implementation

· 이 논문은 2000년도 한림대학교 학술연구 조성비에 의하여 연구되었습니다.
† 비 회 원 : 코리아와이즈넛 연구원
webjoy@hananet.net
†† 종신회원 : 한림대학교 정보통신공학부 교수
jwlee44@hallym.ac.kr
††† 비 회 원 : 한림대학교 정보통신공학부 교수
cypark@hallym.ac.kr
논문접수 : 2001년 1월 30일
심사완료 : 2001년 11월 30일

1. 서 론

웹은 90년대 초에 시작되어 근 10년간에 역사상 유례를 찾아볼 수 없을 정도로 성장한 정보 네트워크가 되었다. 웹의 규모가 거대해지고 웹 서버 수가 기하급수적으로 증가함에 따라 웹 서버에 산재한 정보를 효율적으로 찾으려는 시도를 하게 되었고, 이러한 시도의 결과로 다수의 상업용 인덱스 기반 검색엔진이 출현하였다. 인덱스 기반 검색엔진은 가상공간 전체를 검색공간으로 설정하고, 물리적인 HTML 페이지를 검색대상으로 한다. 인덱스 기반 검색엔진은 다음과 같은 과정을 통해 정보 검색을 수행한다.

① 추출단계: 검색공간에서 검색대상을 추출.

② 인덱싱 단계: 검색대상 내에서 단어 인덱스를 생성하여 데이타저장소에 저장.

③ 검색 단계: 사용자가 입력한 검색어와 일치하는 인덱스를 데이타저장소에서 찾아 전달.

이와 같은 방식의 인덱스 기반 검색엔진은 검색공간의 급속한 팽창과 의미검색의 부재로 검색결과 품질저하 및 검색어 선정에 대한 사용자 부담 증가, 검색 시스템의 대규모화 등의 근원적인 문제를 내재하고 있다[1].

정보추출 방식의 근본적 발전이 지연되는 동안 보다 상세한 정보를 편리하게 취득하려는 웹 사용자들의 기대와 인덱스 기반 검색엔진 품질간의 괴리가 생겨났는데, 이 차이를 극복하기 위하여 검색엔진에 인공지능과 에이전트개념을 도입한 사용자 친화적 검색방식에 대한 연구가 최근에 활발히 진행되고 있다. 새로운 검색방식들 중에 특히 검색공간을 특정 정보공간으로 제한하고 그 정보공간의 대표적인 특징을 추출하여 통합하는 통합 정보 검색방식이 최근에 많이 연구되고 있다[2,3,4,5,6]. 정보통합시스템(information integration system)은 검색대상을 동질의 정보공간으로 한정하고 추출대상을 정보공간 대표특성(feature space)으로 규정함으로써 인덱스방식 검색엔진의 단점들을 해결한다. 가장 중요한 변화는 추출단계에서 이루어진다. 추출단계에서 제한된 정보원천 중 물리적 검색대상이 아닌 의미를 보유한 특성정보를 추출하고, 추출한 특성정보를 통합함으로써 정보품질이 크게 향상된다. 검색방식에 있어서도 키워드방식과 더불어 정보영역의 특성을 반영한 마법사 방식 질의나 메뉴판 방식 질의와 같은 다양한 검색방법이 채용되었다. 또한, 제한된 동일 정보공간에서 검색된 정보를 제공하므로 추출된 정보에 대한 상세 분류가 가능할 뿐만 아니라 검색공간을 제한하였기 때문에 웹 공간의 무한 팽창으로부터 받는 영향이 더욱 줄어들게 된다.

일반적으로 정보통합은 크게 정보추출, 정보통합, 정보검색의 세 단계 절차로 이루어진다. 정보추출은 웹 데이타의 비정형적 구조로 인한 여러 어려움을 극복하는데 초점이 맞추어져 있어 가장 많은 연구가 이루어지는 분야이다. 대부분의 정보통합시스템은 비정형적인 웹 데이타에서 정형화된 정보를 추출하기 위하여 정보모델(information model)을 정의하고, 그에 부합되도록 정보원천별 변환모듈(wrapper)을 구현하고 있다. 정보추출 단계에서는 변환모듈을 관리하는 데에 드는 부담을 줄이기 위한 각종 방법들이 제시되고 있다. 정보통합 영역과 정보검색 영역은 대부분 기존 분산 데이타베이스 시스템의 구조를 차용하고 있으며, 추출된 정보간의 일관성유지와 정보검색을 위한 질의방법 등이 주로 연구의 대상이 되고 있는 실정이다[2,4,5].

본 논문에서는 실험대상 정보영역으로 국제입찰정보를 채택하였다. 국제입찰정보는 국가간 물리적인 시장장벽을 허물고 자유무역을 활성화시킬 목적으로 WTO 가입국들 간의 협약에 의해서 제공되는 공개정보로, 작게는 연필에서부터 크게는 비행기에 이르기까지 다양한 국가 조달품목을 대상으로 한다. 국제입찰정보영역의 특성은 (1) 입찰내용을 설명하는 제목, (2) 입찰정보 공고일, (3) 입찰 마감일, (4) 입찰 상세 내역, (5) 입찰정보 분류로 대표될 수 있으며, 입찰정보의 특성상 제목과 상세 내역은 대부분 국제통용어인 영어로 표현된 분명한 어조의 단문으로 구성되어 있다. 공고일과 입찰 마감일은 각국의 날짜 표현양식으로 제공된다. 그 특성상 입찰정보 대부분은 사이트 자체 데이타베이스 검색결과를 보여주는 동적 페이지 형태로 제공되고 있으며, 정보출력형태도 원천 정보소스마다 1 레벨 1 페이지 또는 1 레벨 다중페이지, 2 레벨 페이지 등으로 매우 다양하다[7]. 전 세계에서 발굴된 국제입찰정보 사이트에서 제공하는 입찰정보의 출력형태는 다음과 같이 4가지 출력형태로 분류된다.

- *단순목록형(1 레벨 1 페이지형식)*: 한 개의 HTML 페이지 안에 입찰제목과 입찰내역을 모두 나열 식으로 표현한 형태
- *파일목록형(2 레벨 1 페이지형식)*: 한 개의 HTML 페이지 안에 입찰제목이 나열되어 있고, 제목에 입찰내역 이진파일이 첨부 파일 형태로 링크된 형태
- *단순게시판형(2 레벨 2 페이지형식)*: 한 개의 HTML 페이지 안에 입찰제목이 나열되어 있고, 제목에 입찰내역 HTML 파일이 링크된 형태(단, 첨부파일은 없음)
- *복합게시판형(3 레벨 2 페이지형식)*: 단순 게시판

형과 동일하나, 한 개 이상의 이진 첨부파일이 입찰내역 HTML 페이지 안에 링크된 형태

본 논문에서 제시하는 정보통합시스템 TIC은 웹 상에 산재해 있는 특정 정보영역에 대한 상업용 정보통합 시스템 구축을 궁극적인 목표로 하고 있다. 따라서, TIC을 설계하고 구현하는 데에 있어서 우리는 상업화의 핵심 요구 조건인 추출품질을 최우선의 판단기준으로 설정하였다. 추출단계의 변환모듈은 HTML태그간 문자열 패턴을 통해 정보추출위치를 지정하는 정보위치지정(information pointing) 기법을 사용하여 구현하였다. 또한, 국제입찰정보원천의 특징인 복합적인 정보출력형태를 수용하기 위하여 다단계 페이지에 걸쳐 분리되어있는 정보를 추출하도록 이동제어 기능을 부여한 변환모듈을 설계하였다. 통합단계에서는 일관성 유지를 위하여 신규 추출된 정보와 기존정보와 중복여부를 판단하여 중복된 정보를 제거시키는 데이타 일관성 유지/관리기법과 사용자 인터페이스 일관성 유지기법을 적용하여 추출된 정보의 품질을 향상시켰다. 마지막으로 질의방법은 관계형 질의어인 SQL을 채택함으로써 상업용 웹 게이트웨이 솔루션(예. 마이크로소프트사의 액티브 서버 페이지, 선마이크로시스템즈사의 자바 서블릿)에서 쉽게 호출하여 검색할 수 있도록 구현하였다.

2. 기존 연구

웹의 급속한 발전으로 웹 페이지를 매개체로 표출되는 정보량이 급증함에 따라 정보통합 연구에서 웹 상의 정보가 중요한 정보원천으로 부각되었다. 웹 상의 정보를 통합하려는 시도는 이질 정보원천으로부터 추출된 데이타를 통합하는 중개자(mediator) 방식의 정보통합 시스템과 여러 웹 사이트 안에 산재한 웹 데이타 통합만을 위한 정보추출 시스템으로 가시화되었다. 이 시스템들은 공통적으로 자신들만의 정보모델을 정의하고 이를 변환의 대상으로 또는 통합질의의 대상으로 삼고 있기 때문에, 웹 페이지에서 추출한 정보를 자신들의 정보모델로 변환하는 변환기(wrapper)를 구비하고 있다. 정보모델은 각 시스템의 특성에 맞게 설계되었으며, 주로 질의 언어와 밀접한 관계를 가진다. 변환기는 일반적으로 규칙기반, 휴리스틱, 패턴 매칭, 추론학습 등의 기법을 사용하여 데이타 변환을 시도하고 있으며, 최근에는 특정 정보영역에서 범용 정보영역으로 확장하려는 방향으로 연구가 진행 중에 있는 상황이다. 또한, 변환기 관리부담을 줄이기 위해 변환기 자체를 자동 생성하는 방법들이 많이 연구되고 있다.

기존의 정보통합 시스템들은 관계형 데이타베이스, 텍스트, 파일 시스템 등이 정보원천 이었으나, 최근 들어 웹 페이지 상의 데이타 통합 모듈에 대한 연구가 활발히 진행되고 있다. 대표적인 시스템으로는 TSIMMIS[2], ARANEUS[4], ARIADNE[5]가 있다. TSIMMIS와 ARANEUS는 전문가의 수 작업에 의존한 수작업 생성 방식 변환기를 사용하였다. 일반적으로 수작업 생성기법은 소수의 원천소스가 있는 경우와 포맷 변화가 적은 경우에 적합한 기법이다. 따라서, 정보영역이 추가되고 각 원천소스의 포맷 변화가 빈번할수록 수 작업 방식은 변환기를 갱신하는데 필요한 노력이 증대되어 바람직하지 않게 된다. 이와 같은 관리 문제점에 대한 보완으로 ARIADNE는 휴리스틱을 사용하여 변환기를 반자동으로 생성한다[8].

대표적인 정보추출 시스템으로는 ShopBot[3], MORPHES[5] 등이 있다. 정보통합 시스템과 달리, ShopBot과 MORPHES는 이질 정보원천을 수용할 수 없기 때문에 정보원천이 웹 페이지만으로 제한된다. 또, 정교한 질의처리나 복잡한 정보모델보다는 단순한 형태

표 1 정보 추출 및 통합 시스템의 특징

시스템이름 (연도)	개발기관	목적	대상정보		변환기		질의 언어
			영역	정보모델	변환방식	생성방식	
TSIMMIS (1994)	Standford Univ.	정보 통합	제한 없음	Object Exchange Model	규칙기반	수동	OEM QL
ShopBot (1996)	Washington Univ.	정보 추출	상품 정보	Domain Property	휴리스틱 검색 패턴매칭 추론학습	자동	없음
ARANEUS (1997)	Univ di Roma Tre	정보 통합	제한 없음	ARANEUS Data Model	규칙기반	수동	ULIXES PENELOPE
ARIADNE (1998)	Univ. of South California	정보 통합	제한 없음	Nested Hierarchy	휴리스틱 검색	반자동	KQML
MORPHEUS (2000)	한양대학교	정보 추출	상품 정보	Ontology	추론학습 노이즈제거	자동	없음

의 정보모델만을 제공한다는 단점도 있다. ShoptBot과 MORPHEUS는 모두 변환기에 중점을 둔 시스템으로, ShopBot은 휴리스틱, 패턴매칭, 추론학습 기법을 사용하여 상품정보영역에 대한 변환기를 자동 생성한다. MORPHEUS는 ShopBot의 웹 페이지 구조에 대한 전제를 완화하고, 강화된 추론학습 기법을 적용하여 보다 유연한 방식으로 변환기를 자동 생성한다. 하지만 이 두 시스템은 추출한 정보에 대한 통합처리 부족과 빈약한 정보모델로 인해 상업용 정보통합 시스템의 핵심 요구사항인 통합정보 품질 향상에는 크게 도움을 주지 못한다고 할 수 있다. 표 1은 기존 정보 추출 시스템 및 통합 시스템들의 특징을 요약하고 있다.

3. 설 계

본 논문에서 제시한 국제입찰정보 통합시스템인 TIC은 전세계 국가에서 제공하는 국제입찰 원천 데이타를 수집한 후, 이를 검색 및 분류가능한 양질의 국제입찰정보로 가공하고 통합함으로써 단일 국제입찰정보 취득창구를 제공하기 위한 목적으로 설계되었다. TIC 설계에 있어서 우리는 상업적 가치를 기본 설계 기본 요건으로 정하였기 때문에 통합정보 품질향상을 위해 TSIMMIS, ARANUS, ARIADNE 시스템과 같이 정보모델과 질의언어를 채택하였고, 웹 상의 정보 추출에 대한 부족한 기능을 보완하기 위해 ShopBot과 MORPHEUS 시스템과 같이 웹 상 정보의 추출방식을 적용하였다. 이와 같은 혼성형 시스템으로 TIC을 설계함으로써 ShopBot, MORPHEUS에서 부족한 질의처리능력을 강화하였고, TSIMMIS, ARANUS, ARIADNE 시스템에서 부족한 웹 정보 추출 능력을 보완하여 상업용 품질을 제공하는 웹 정보통합 시스템을 설계하고 구현하였다. 또한, 통합정보 품질향상을 위해 TSIMMIS, ARANUS 시스템과 같이 추출정보에 오류가능성이 높고 복합적인 정보추출 형태에 대처하기 어려운 변환기 자동 생성방식보다는 수작업 생성방식을 채택하였다. 이밖에 기존 정보통합 시스템이 정보를 담고 있는 전후 페이지들만을 대상으로 적용하는 것과 달리 TIC은 정보 효용성 판단을 위해 접근성 테스트를 수행하여 의미 없는 정보추출을 방지하였고, 추출 후 통합단계에서는 데이타 및 사용자 인터페이스 일관성 유지기법을 사용하였다.

TIC의 정보모델은 관계형 모델에 기반을 두고 있는데 관계형 모델은 현재 상업용 검색시스템에서 가장 많이 사용하는 모델로써 관계형 데이타베이스 시스템을 정보저장소로 사용한다. 그림 1에서 볼 수 있듯이 TIC

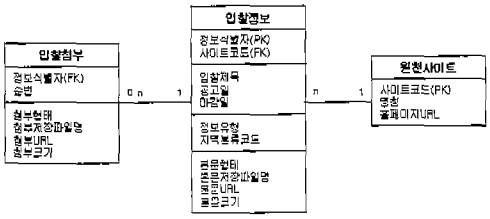

그림 1 TIC의 관계형 정보 모델(큰 사각형 하나가 관계형 정보모델 구성요소인 엔터티 하나에 해당함)

의 정보 모델은 원천사이트와 입찰정보, 입찰첨부 등 총 3개의 작은 엔터티(entity)로 구성되며 다음과 같은 세 가지 부류의 정보를 담고 있다.

- 국제입찰 정보 원천사이트 정보 : 원천사이트 엔터티
- 국제입찰 정보 분류 정보 : 입찰정보 엔터티의 지역분류 속성
- 국제입찰 정보 : 입찰정보/입찰첨부 엔터티

국제입찰 정보영역이 갖는 일반적인 특성은 입찰정보 엔터티의 여러 속성 항목에 의해 표현되는데, 입찰정보 엔터티에는 입찰내용을 규정하는 입찰제목, 공고일, 마감일, 지역분류코드 속성 등이 포함되어 있다. 하지만 입찰정보 엔터티는 입찰에 대한 메타데이타의 역할만을 하기 때문에 입찰정보에 대한 상세 내역은 본문에 첨부된 파일을 참조해야 한다. 따라서 입찰정보 엔터티에는 첨부 문서의 형태(예를 들면, HTML, DOC, PDF 등)에 대한 속성도 포함된다. 본문에 첨부된 자료는 입찰첨부 엔터티로 표현된다. 원천사이트 엔터티와 입찰정보 엔터티는 1:(1...N)의 관계를 가지며, 입찰정보 엔터티와 입찰첨부 엔터티는 1:(0...N) 관계를 가진다.

입찰정보는 시간제약 속성을 가지고 주기적으로 갱신된다. 국제입찰정보는 원천사이트 내의 데이타베이스가 갱신됨으로 인해 동적으로 갱신되기도 하고, 정적 페이지로 구성된 입찰정보 페이지의 URL이 변경됨으로 인해 정적으로 갱신되기도 한다. 동적 갱신방식은 URL이 대부분 고정되어 변동이 많지 않으나, 정적 갱신방식은 정보가 갱신되면서 신규 URL이 생성되었는지의 여부를 추적하기 어렵다. 따라서 입찰정보의 갱신여부를 추적하기 위해서는 접근성(accessibility) 검증이 필요하다. TIC은 원천사이트의 루트 URL에서 필요한 입찰정보를 보유한 목적 페이지(target page)에 도달하기 위한 이동경로(navigation path)를 각 입찰정보 원천마다 지정하여 원천소스 변경을 초기에 검출할 수 있도록 함으로써 추출정보의 효용성을 높이고자 하였다.

정보추출을 위해서 TIC은 다양한 필터를 사용하는데, 필터란 추출할 문자열 패턴을 보유한 소프트웨어 모듈을 의미한다. TIC에서 사용하는 필터 종류에는 정보추출용 필터와 이동제어용 필터가 있는데, 정보추출용 필터는 입찰정보 추출 필터와 제거필터로 구성된다. 추출 필터는 목적 페이지에서 입찰정보를 추출하기 위한 문자열 추출패턴을 보유하게 되고, 제거 필터는 추출된 데이타에서 불필요한 데이타를 제거하기 위한 제거패턴을 보유한다. 추출패턴을 사용해 추출하려는 정보가 담겨져 있는 페이지 내의 특정 위치를 찾고 나면, 추출필터는 추출패턴이 지정한 해당 영역을 추출한다. 추출 필터가 추출패턴을 찾을 때 사용하는 알고리즘에는 정규 표현(regular expression)을 이용하는 방법과 HTML 태그간 문자열 일치를 이용하는 방법이 있는데 본 논문에서는 후자를 사용하였다. 정규 표현은 표현력은 풍부하지만 사용하기 힘들고 시스템에 부하는 많이 주는 등의 단점이 있는 반면, HTML 태그간 문자열 일치 방식은 사용하기 쉽고 간단하면서도 큰 오버헤드 없이 원하는 문자열을 추출할 수 있다는 장점이 있다. HTML 태그간 문자열 일치 방식에서는 두개 이상의 HTML 태그를 분리자(delimiter)로 사용하여 분리자 사이의 문자열 블록을 추출하는데, 이를 위해 소요되는 HTML 태그 파싱(parsing) 오버헤드는 정규 표현 처리에 드는 오버헤드보다 매우 적다고 할 수 있다. 추출필터는 경우에 따라 다중-패스 필터로 구성되기도 하지만 원천사이트의 정보 출력 형태는 정형화되어 있기 때문에 대부분의 경우 단일-패스 필터만으로도 충분한 정보를 추출할 수 있다.

원천사이트의 페이지가 복합적인 출력형태로 구성된 경우에는 페이지 간 이동을 제어해야 하는데, 이는 이동제어 필터가 수행한다. 이동제어 필터는 루프 검출필터와 종료조건 검출필터로 구성된다. 루프 검출필터는 반복적으로 무한히 출력되는 정보목록 페이지에서 다음 페이지로 이동하기 위한 이동제어 정보를 추출한다. 종료조건 검출필터는 루프종료 여부 판단을 위한 이동제어 정보를 추출한다. 정보변환기는 이 두 필터를 사용하여 정보추출과 동시에 페이지 간 이동을 제어한다. 그림 2에 캐나다 뉴브룬스위크주 입찰사이트(http://www.gnb.ca/)에서 입찰정보를 추출하기 위한 정보변환기의 이동제어필터 적용 예를 보이고 있다. 이동제어필터의 입력데이타에는 정적 데이타와 동적 데이타 두 가지가 있다. 정적 데이타는 다단계 접근성 테스트용 가교페이지를 지정하기 위한 이동경로(URL 목록)와 입찰정보 검색페이지인 폼 페이지에서 입찰정보를 담고 있는 페이지를 검색하기 위한 폼 데이타가 있다. 정적 데이타는 입찰 사이트에 변경이 없는 한 계속 유지된다. 반면에 동적 데이타는 수집기간에 게시된 입찰정보의 유무에 따라서 변동되는 데이타로 정보 변환기의 이동제어용 입력데이타로 사용된다. 이 같은 동적 데이타는 루프 검출필터, 종료조건 검출필터에서 추출된다.

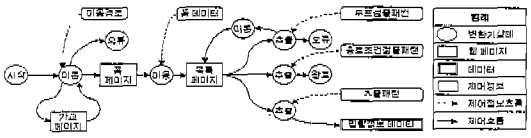

그림 2 정보변환기 상태천이도 및 추출 흐름도(캐나다 뉴브룬스위크주 입찰사이트)

정보통합 단계에서는 일관성 유지기법을 적용하여 추출정보의 품질을 향상시킨다. 국제입찰 입찰공고는 공고대상지역이 중첩됨으로 인해 상이한 원천사이트들의 입찰정보가 서로 중복되기도 하고, 수집 주기와 갱신 주기와의 차이로 인해 단일 사이트 정보가 중복 추출되기도 한다. 중복 추출의 원인을 정리하면 다음과 같다.

- TIC 외적인 요인: 같은 입찰정보를 두 곳 이상의 원천사이트에서 동시 혹은 시간 차이를 두고 공고할 경우 사이트간 중복 추출이 발생할 수 있다.
- TIC 내적 요인: TIC에서 정보를 수집하는 모듈의 실행 주기가 사이트 내의 한 입찰공고 지속 기간보다 짧음으로 인해 발생하는 현상으로써 단일 사이트 내에서 같은 정보를 여러 번 추출하는 부작용을 유발한다. 단일 사이트 중복 추출 현상이 사이트간 중복 추출에 비해 좀 더 빈번히 발생한다.

중복추출 현상은 통합정보의 효용성을 크게 저하시킬 뿐만 아니라 저장 공간의 효율성을 저하시키고, 수집된 정보의 일관성을 저해하는 큰 요인이다. 게다가 중복 수집으로 인해 시스템에 불필요한 부하를 주게 된다. 따라서, TIC에서는 중복추출을 최대한 방지하기 위해, 대부분의 입찰정보가 보유하고 있는 공통 특성을 활용하여 중복을 제거하는 도메인 의존 방식을 도입하였다. 이 방식은 서로 다른 입찰공고의 제목이나 공고일, 마감일, 원문, 분류 등의 속성이 동일한 경우는 거의 없다는 가정에 기반한다. TIC에서는 사이트 코드, 제목 길이, 원문 크기(바이트 수), 마감일을 동일성 판단기준으로 설정하였으며, 이들 속성이 동일한 입찰공고는 같은 공고로 간주하여 중복 추출되지 못하도록 하였다.

4. 구 현

본 논문에서 제시한 정보통합시스템 TIC은 Windows

NT 4.0 환경에서 구현되었다. 실험 환경은 수집서버와 DB서버 두 대의 Windows NT서버를 사용했으며, 수집서버에는 TIC 엔진을 탑재하였고 DB서버에는 국제 입찰정보 저장소로 SQL Server 6.5를 사용하였다. 수집서버에서 DB서버에 접근하는 API로는 ADO(Active Data Object)[10]를 사용하였으며 모든 모듈은 Visual C++ 6.0을 사용하여 C++로 구현하였다.

TIC은 클라이언트/서버 웹 구조에서 클라이언트 측 소프트웨어에 해당하므로 정보추출을 위해 웹 브라우저의 기능을 구현해야 했지만, TIC에서는 이동 및 추출결과를 화면상에 출력할 필요가 없고 구현의 복잡성 때문에 웹 브라우저의 기능 중 반드시 필요한 기능만을 선택하여 차용 혹은 개발하였다. TIC에서 필요한 웹 브라우저의 기능 중 HTML 페이지 요청 및 추출을 위해 사용되는 TCP/IP 데이타 통신, HTTP 프로토콜 해석기는 Microsoft에서 제공하는 Win32 Internet Function[9]을 차용하였고, 필터구현에 필수적인 HTML 문법 해석기는 그 용도에 맞도록 간략하게 구현하였다.

TIC의 정보변환기는 통합정보의 품질향상을 위해 수작업 생성방식을 통해 정확한 입찰정보만을 웹 페이지 내에서 추출한다. 이에 따른 추출기 생성부담을 줄이기 위해서 TIC은 모든 웹 사이트에서 공통적으로 필요한 기능을 C++의 상속개념을 이용하여 클래스 계층도를 만들어 코드재사용성을 극대화하였다. 즉, 정보변환기는 클래스계층도의 최하단에 위치한 CSite클래스를 상속하고 CSite에서 정의된 가상멤버함수를 구현함으로써 정보추출에 필요한 최소한의 코드로 생성 가능하도록 구현하였다.

정보변환기에서 가장 핵심부문은 필터구현이었으며, 추출하려는 정보추출패턴, 제거패턴, 루프패턴, 종료조건 패턴들은 이 필터에 의해서 신속하고 정확하게 추출되어야 하므로 성능을 특히 고려했다. TIC은 다수의 원천 사이트에서 많은 필터를 사용하므로 필터의 성능향상이 전체 시스템의 성능에 막대한 영향을 주기 때문에 많은 연산을 요하는 정규식-기반 HTML 해석기 대신 약식 HTML 해석기를 구현하였고, 아울러 추출하려는 패턴을 태그단위로 선정하여 추출패턴의 위치를 지정하는 HTML 태그 간 일치방식으로 구현하였다. 이 해석기도 CSite 클래스 안에 멤버함수로 구현하여 모든 변환기에서 원하는 패턴을 지정하여 필요한 정보를 추출할 수 있도록 구현하였다.

TIC은 크게 정보통합 프레임워크와 정보변환기, 그리고 작업제어기로 구성된다(그림 3). 정보통합 프레임워크는 모든 정보변환기가 공통적으로 사용할 서비스를 제공하기 때문에 C++클래스를 표출하는 MFC 확장

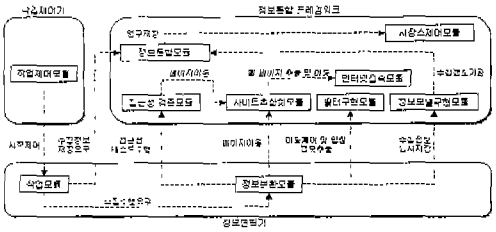

그림 3 TIC을 구성하는 모듈 구성도

DLL(Dynamic Link Library)형태로 구현되었다. 정보변환기에 제공하는 공통 서비스는 인터넷 접속모듈, 저장소 제어 모듈, 접근성 검증 모듈, 사이트 추상화 모듈, 필터 구현 모듈, 정보모델 구현 모듈, 정보 통합 모듈들로 구성된다. 정보변환기는 원천사이트 별로 수작업에 의해 하나씩 실행파일 형태로 구현되었으며, 정보변환기들의 실행 시점을 제어하는 작업제어기도 실행파일 형태로 구현하였다.

4.1 TIC 동작 예

TIC은 모두 10개의 하부 모듈들을 가지며 각 모듈들은 유기적인 관계를 가지며 동작한다. 시스템 구현내역을 설명하기 위해 구현된 TIC 모듈간의 동작을 정보변환기를 중심으로 캐나다 뉴브룬스위크 주정부 사이트(http://www.gnb.ca) 실례를 적용하여 기술한다. 구현된 정보추출 및 정보통합 과정을 단계별로 구체적인 데이타와 함께 기술하면 다음과 같다.

캐나다 뉴브룬스위크 주정부 사이트는 국제입찰정보 출력형태 중 복합게시판형에 해당하고, 입찰정보를 검색할 수 있는 입찰정보 검색페이지까지 8개의 가교 페이

표 2 정보변환기의 이동 경로 예(캐나다 뉴브룬스위크 주)

순서	정보 변환기 이동 경로	페이지 설명
1	http://www.gnb.ca/	홈페이지
2	http://www.gnb.ca/index-e.asp	영문 홈페이지
3	http://www.gnb.ca/dept.htm	부서 홈페이지 링크 모음
4	http://www gnb.ca/dss-mas/index.htm	구매부 홈페이지
5	http://www.gnb.ca/0337/Index.htm	구매부 구매 페이지
6	http://www.gnb.ca/0337/Purche.asp	구매부 입찰정보 페이지
7	http://www gnb.ca/pur-achat/nbop.asp	구매부 입찰 공개 네트워크
8	http://www.gnb.ca/pur-achat/SearchTenders.asp	입찰정보 검색 페이지

지를 가진다. 입찰정보를 찾기 위해 정보변환기는 사이트추상화모듈과 인터넷접속모듈을 이용하여 주 정부 홈페이지부터 표 2에 나타난 이동 경로를 따라 입찰정보 검색 페이지(폼 페이지)에 접근여부를 타진하여 입찰정보의 접근성을 검사한다.

정보변환기는 HTML FORM 구문으로 구성된 입찰정보 검색 페이지에서 주 정부의 모든 기관에서 공고된 입찰정보를 검색하기 위해, 사전에 고안된 폼 데이타인 "Category=1&Organization=1& Product=N0&Submit=Search+Tenders&LastDayes=All"를 사이트추상화모듈에 전달하여 검색결과를 담고 있는 목록페이지를 얻어낸다. 그림 4는 웹 브라우저를 사용하여 얻은 입찰정보 검색목록에 포함된 입찰정보 요약본 한 개의 모습이다. 이 요약본 HTML 소스에서 정보변환기는 추출필터를 사용하여 TIC의 정보모델 중 입찰정보 엔터티 속성들인 입찰제목, 공고일, 마감일, 본문형태, 본문저장파일명, 본문 URL들을 추출한다. 그림 5는 그림 4의 입찰정보에 대한 HTML 소스와 HTML 태그 간 일치방법에 의해서 국제입찰 엔터티 속성들을 추출할 때 적용하는 속성패턴들을 함께 보이고 있다. 각 속성 추출패턴은 목록 페이지 안에서 고유한 문자열이 되어야 하므로 추출 대상 속성을 감싼 태그보다 상위의 태그로 확장하면서 추출 패턴이 선정된다. 선정결과는 그림 5에서 역상으로 표기되어 있다.

각 입찰정보 속성들은 입찰공고 요약본 HTML 소스에서 다음과 같은 방식으로 추출된다. 추출결과는 메모리 상에 임시 저장되고, 추출과정을 모두 마친 후 통합과정을 수행하면서 표준화된 값 혹은 형태로 다시 한번 변환된다.

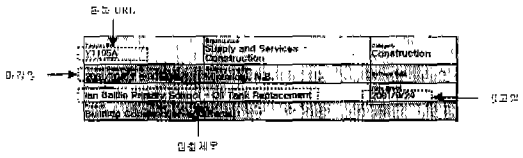

그림 4 입찰공고 요약본 예(캐나다 뉴브룬스위크 주)

■ 입찰제목: 입찰제목 추출패턴을 사용하여 직접 추출한다("Ian Baillie Primary School - Oil Tank Replacement").

■ 공고일: 공고일 추출패턴을 사용하여 추출한다("2001/9/24").

■ 마감일: 마감일 추출패턴을 사용하여 추출한다("2001/10/5 / 3:00 P.M.").

■ 본문형태: 뉴브룬스위크 주 입찰사이트에서는 입찰정보 본문을 HTML로 제공한다("HTML").

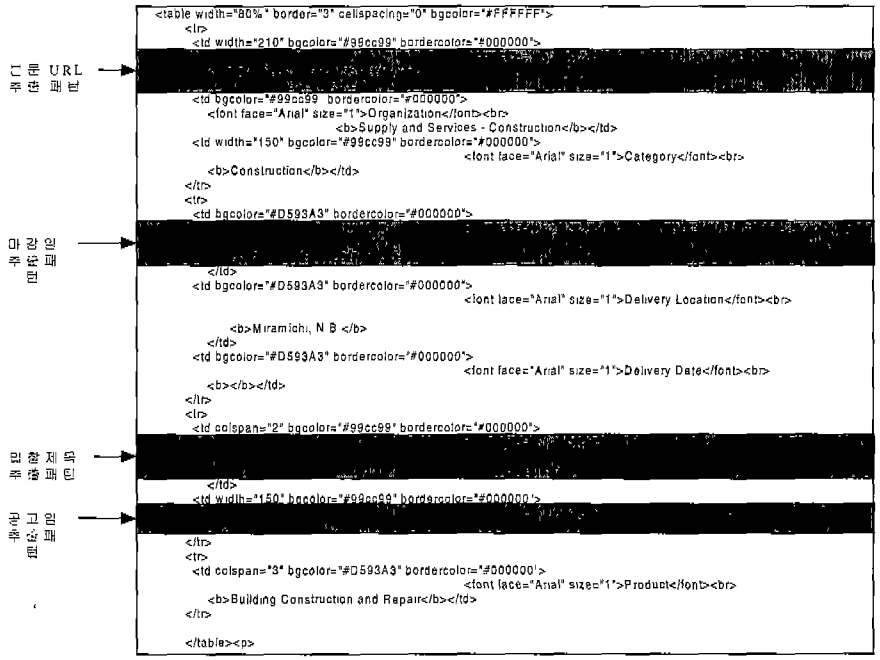

그림 5 입찰공고 요약본 HTML 소스와 입찰정보 엔터티 속성 추출 패턴 예(캐나다 뉴브룬스위크 주)

■ 본문저장파일명: 중복방지를 위해 밀리초 단위로 측정한 추출 시간으로 파일명을 생성하고, 추출된 본문 URL을 인자로 사이트추상화모듈을 통해 본문 HTML을 다운로드하여 저장한다 ("NEWBW20010927011027121.HTM").

■ 본문URL: 본문URL 추출패턴을 사용하여 추출한다 ("SearchTenders3.asp?Id2=2474").

■ 본문크기: 본문저장파일의 크기를 바이트단위로 기록한다.

뉴브룬스위크 주 사이트가 복합게시판 형태이기 때문에 입찰정보 요약본 목록이 여러 페이지에 나뉘어 출력될 수 있다. 이런 경우를 위해 정보변환모듈에서는 필터 구현모듈의 도움을 받아 다음 페이지로 이동을 위한 폼 데이타(다음버튼)가 있는지 검출하면서 페이지 이동제어를 수행한다. 정보변환기는 폼 데이타를 목록 페이지에서 추출하여 사이트추상화모듈을 통해 다음 목록페이지로 이동하면서 입찰정보들을 계속해서 추출하게 된다. 다음 페이지로 이동하기 위한 루프검출패턴 이외에 페이지 이동을 중지할 종료조건검출패턴도 매번 목록페이지로부터 추출되어 페이지 이동제어에 사용된다. 종료조건이 검출되면 정보추출과정은 모두 끝나게 된다. 이렇게 추출된 정보들은 정보모델 구현모듈이 관리하는 메모리상의 임시 저장소로 저장된다.

정보통합과정은 정보모델 구현모듈이 보유하고 있는 미가공 입찰정보들을 영구 저장소로 기록하면서 배치작업으로 수행된다. 통합작업은 추출된 정보의 데이타 일관성 유지를 위해 수행되는 입찰정보 중복 검출 및 제거, 그리고 TIC의 관계형 정보모델에서 정의된 표준형태로의 속성필드 변환, 시각적 일관성 유지를 위한 HTML 소스 조정 및 입찰제목 변환작업 등으로 구성된다. 각 속성별 상세 통합작업 및 그 결과는 다음과 같다.

■ 입찰제목: 시각적 동일성 향상을 위해 전부 대문자로 변환한다("IAN BAILLIE PRIMARY SCHOOL - OIL TANK REPLACEMENT").

■ 공고일: TIC의 표준날짜형태인 연월일 포맷으로 변환한다("20010924").

■ 마감일: TIC의 표준날짜형태인 연월일 포맷으로 변환한다("20011005").

■ 본문형태: 변환하지 않는다("HTML").

■ 본문저장파일명: 변환하지 않는다("NEWBW20010927011027121.HTM"). 단, 본문 HTML은 시각적 동일성 향상을 위해 스타일쉬트를 제거하고 이미지 태그를 제거한다.

■ 본문URL: 이동 경로를 바탕으로 전체 URL을 생성한다("http://www.gnb.ca/pur-achat/SearchTenders3.asp?Id2=2474").

■ 본문크기: 변환하지 않는다.

위와 같은 통합과정을 거친 입찰정보는 저장소제어모듈을 통해 영구히 관계형 저장소에 저장되고, 정보검색 단계에서 관계형 질의 언어를 사용하여 검색된다.

5. 평 가

본 논문에서 설계 및 구현한 TIC의 효용성을 평가하기 위해 국제입찰정보를 제공하는 전 세계 57개 원천사이트를 발굴하여, 제공하는 입찰정보 출력형태에 따라 이들을 4가지 유형으로 분류하였다. 그리고 각 유형 별로 수집한 입찰정보의 개수와 수집을 위해 작성한 정보변환기의 소스코드 평균 라인 수, 접근성 테스트용 평균 경로이동 횟수, 적용된 정보모델, 그리고 중복 제거율 등을 측정하였다. 표 3과 표 4에서 보이고 있는 측정 결과는 각 원천사이트마다 정보변환기를 구현하여 8 개월간의 운영결과를 토대로 작성한 것이다. 상당한 수준의 중복 제거율을 달성했다는 점에서도 유추할 수 있듯이 우리는 본 논문에서 개발한 정보 통합 시스템 TIC이 상업용 서비스로 만족할 만한 품질을 보인다는 것을 확인할 수 있었다. 또한 최소한의 관리 인력만으로도 유지보수를 충분히 수행할 수 있었다.

표 3 정보출력 형태별 실험결과

원천사이트 분류	사이트	수집 정보	정보변환기 소스코드 평균 라인 수	평균 경로이동 횟수	정보모델
단순목록형	5(8.77%)	5,155	291	3.8	입찰정보
파일목록형	8(14.04%)	3,477	296	2.63	입찰정보
단순 게시판형	34(59.65%)	40,744	219	2.59	입찰정보+ 입찰첨부
복합 게시판형	10(17.54%)	7,139	261	2.7	입찰정보+ 입찰첨부
전체	57	56,515	244	2.72	

표 4 중복제거적용 결과

원천사이트분류	중복제거 전 입찰정보 수	중복제거 후 입찰정보 수	중복 제거율
단순목록형	20,093	5,155	74.3%
파일목록형	11,307	3,477	69.2%
단순게시판형	162,498	40,744	74.9%
복합게시판형	14,708	7,139	51.5%
전체	208,606	56,515	72.9%

원천사이트를 출력형태에 따라 분류한 결과 단순 게시판형이 59.65%로 가장 많았으며, 따라서 단순 게시판형 원천사이트로부터의 수집 개수와 사이트 당 수집 개수도 각각 40,744개와 1,198개로 가장 많았다. 정보변환기 관리부담을 가장 잘 표현해주는 정보변환기의 소스코드 평균 라인 수와 접근성 테스트를 위한 평균 경로이동 횟수는 각각 244라인과 2.72회로 나타났다. 원천사이트 당 평균 244라인만으로 정보변환기를 구현할 수 있었던 것은 정보변환기에서 사용하는 공통 기능들을 정보통합 프레임워크로 모은 후 정보변환기에서는 프레임워크의 공통 서비스를 사용하도록 설계했기 때문이다. 이렇게 하면 특정 사이트에만 적용되는 이동경로 및 추출필터 만을 원천사이트마다 구현하면 되므로 평균 244라인만으로도 충분히 구현할 수 있었다. 또한 정보변환기는 정보출력형태 별로 유사한 코드 패턴을 보이기 때문에, 유사한 출력형태를 보유한 정보변환기 코드내의 코드패턴을 클래스간 상속을 활용하여 재사용 함으로써 개발시간을 많이 줄일 수 있었다. 적은 코딩부담과 코드패턴 재사용 결과 소수의 관리 인력만으로도 코드 관리 및 TIC 운영이 가능했다. 평균 2.72 회의 경로이동만으로 원천사이트 루트에서 원하는 정보를 담고 있는 페이지에 도달할 수 있었기 때문에 접근성 검증 또한 정보변환기 구현에 부담을 주지 않았다.

설계 시 가장 우선적으로 고려했던 정보품질은 프레임워크의 중복제거 모듈에서 전체 대상 원천사이트에 대해 평균 72.9%의 중복데이타를 제거함으로써 고수준의 정보품질을 확보할 수 있었다. 원천사이트 유형별 중복 제거율을 살펴보면 표 4에서도 보이고 있듯이 단순 출력형일수록 중복 제거율이 높다는 것을 알 수 있는데, 이는 출력형태가 단순한 원천사이트일수록 정보중복이 많이 발생하기 때문인 것으로 판단된다. 예를 들어, 목록페이지를 취득하는 방법에 있어서 단순형의 경우는 날짜기준검색을 통한 목록페이지 생성기능이 제공되지 않지만, 반면에 복합게시판형에서는 날짜기준검색을 바탕으로 원하는 목록페이지만을 동적으로 생성할 수 있다. 따라서, 단순형일수록 이전 주기에서 추출되었던 내용이 다음 주기에서도 반복적으로 추출될 가능성이 높아지게 되므로 중복 제거율도 높아졌다고 할 수 있다. 한편 정보변환기에서 중복제거 기능을 사이트 별로 구현하지 않아도 되도록 설계하였기 때문에(즉, 정보통합 프레임워크에서 일괄적으로 중복을 제거함) 정보변환기 관리에 드는 부담을 상당히 절감할 수 있었다.

8개월간의 시험 운영을 통해 확인한 결과 TIC 시스템은 기존 웹 정보추출 시스템과는 달리 복잡한 정보출력형태를 갖는 원천사이트에서도 원활하게 정보를 수집할 수 있다는 것을 확인하였다. 기존 웹 정보추출 시스템에서는 주로 단순목록형과 파일목록형 만을 대상으로 정보를 수집하는 것이 일반적이다. 표 3에서도 알 수 있듯이 TIC이 정보 통합의 대상으로 설정한 정보원천사이트의 77.19%가 복잡한 형태의 게시판형 사이트이기 때문에,

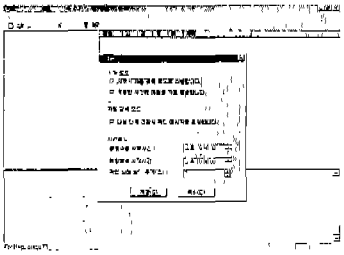

그림 6 TIC 설정 화면

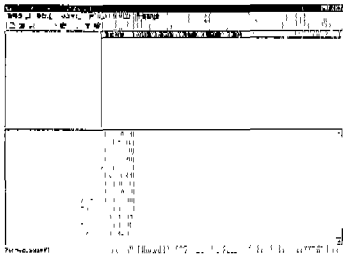

그림 7 TIC 정보수집전 화면

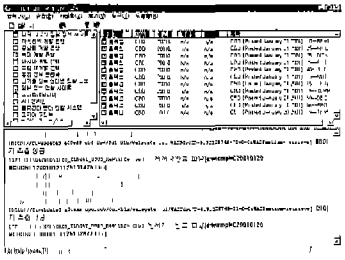

그림 8 TIC 정보수집 화면

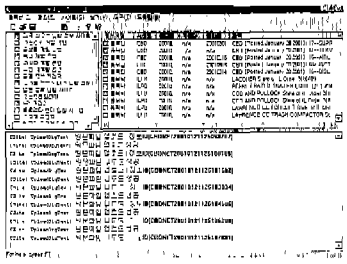

그림 9 TIC 정보수집 저장 화면

기존 정보추출 시스템을 사용해 이들 사이트에서 정보를 추출하려 했다면 정보취득에 한계가 있었을 것이다. 반면에, TIC은 관리부담을 적절히 유지하면서도 복잡한 형태의 이동경로와 출력형태를 보유한 원천사이트에서도 정보를 무리 없이 수집할 수 있음을 확인하였다.

그림 6, 7, 8, 9는 TIC이 실제 수행되는 화면을 보이고 있다. 그림 6은 수집 스케줄과 기타 실행에 필요한 정보를 설정하는 화면으로 설정정보는 TIC 설치 후 한 번만 설정하면 시스템에 저장되어 TIC 수행 중에 이용된다. 그림 7은 정보수집 시작 직전 화면으로 작업 스케줄러만이 수행되는 모습을 담고 있다. 작업 스케줄러는 그림 6에서 설정한 수집시간이 경과하면 정보변환기들을 구동시킨다. 정보변환기들이 구동되어 수집이 시작된 화면이 그림 8이다. 구동된 정보변환기에서 국제입찰정보 수집이 완료되면 그림 9에서 보듯이 수집된 정보를 저장한다. 수집정보 중에 파일형태인 본문파일과 첨부파일은 FTP를 사용하여 지정한 서버에 저장한다. 기타 입찰정보는 후처리작업을 거친 후에 데이타베이스에 저장한다.

6. 결론 및 향후 연구과제

본 논문에서는 상업용 정보통합시스템 설계와 구현에 필요한 프레임워크를 제시하였고, 이를 국제입찰정보 영역에 적용하여 국제입찰 정보통합시스템을 구현하였다. 그리고 구현된 정보통합 시스템을 약 8 개월 동안 운영한 결과를 분석 평가하였다. 평가 결과 총 57 개의 대상 원천사이트에 걸쳐 평균 73%의 높은 중복정보 제거율을 얻을 수 있었으며, 기존 정보검색 시스템과는 달리 복잡 게시판형 출력 형태의 원천사이트로부터도 성공적으로 정보추출이 가능하였다. 본 논문이 기여하는 바는 특정 범주에 속하는 공통정보를 추출 및 통합하는 데에 필요한 시스템 프레임워크를 제시했다는 점과 구현 시 발생하는 문제점들을 실제 국제입찰정보 분석을 통해 제시함으로써 다른 범주의 웹 컨텐츠 추출 및 통합 연구에도 도움을 줄 수 있다는 점이다. 비록 본 논문에서 사용한 검색 알고리즘이 새로운 것은 아니라 할지라도 특수 목적의 정보 검색/통합 시스템에서는 검색 기법보다는 검색된 정보의 통합/가공이 정보의 품질에 더 큰 영향을 미친다는 점도 알 수 있었다.

향후 연구과제로는 본 논문에서 구현된 TIC 시스템을 바탕으로 궁극적으로 기업간 범용 정보통합시스템 개발을 들 수 있다. 즉, 현재 TIC에서 설계/구현한 정보통합 프레임워크는 주로 국제입찰정보 추출을 위한 것이지만 이를 확장하여 사용자가 원하는 범주를 지정하기만 하면 그 범주에 속한 정보들을 통합하는 범용 정보통합 프레임워크를 설계하는 일이다. 이를 위해 우선 현재 TIC의 정보모델이 사용하고 있는 정적 분류체계를 보다 유연한 인공지능형 동적 분류방식으로 개선해야 할 것이다. 그래야만 더욱 향상된 정보통합 품질을 얻을 수 있기 때문이다. 또한, GUI 사용자인터페이스 상에서 필터 템플릿과 이동경로 템플릿을 사용자가 쉽게 제작/사용할 수 있도록 함으로써 수 작업에 의한 코드패턴 재사용 빈도를 줄여야 할 것이다. 이렇게 하면 정보변환기 개발부담을 한층 더 줄일 수 있을 것이다.

참 고 문 헌

[1] P. M. G. Apers. Identifying internet-related database research. In *Proceedings of the Second International EastWest Database Workshop*, Klagenfurt, Workshops in Computing, pages 183-193. SpringerVerlag, 1994.

[2] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *Proc. of IPSJ Conference*, pages 7-18, 1994.

[3] B. Doorenbos, O. Etzioni, and D. Weld. A scalable comparison-shopping agent for the world-wide web. In *Proc. of the First Int'l Conf. on Autonomous Agents*, pages 39-48, February 1997.

[4] Atzeni, P., Mecca, G., and Merialdo, P. Semistructured and structured data in the web: going back and forth. In *Proceedings of ACM SIGMOD Workshop on Management of Semi-structured Data*, pages 1-9, 1997.

[5] C. Knoblock, S. Minton, J. Ambite, N. Ashish, P. Modi, I. Muslea, A. Philpot and S. Tejada, Modeling web sources for information integration. In *AAAI '98*, 1998.

[6] J. Yang, H. Seo, N. Koo, J. Choi, J. Kim, S. Kim, K. Lee, and H. Ham, A More Scalable Comparison Shopping Agent, *Engineering of Intelligent Systems(EIS 2000)*, pp. 766-772, Paisely, Scotland, 2000.

[7] Chidlovskii, B., Borghoff, U., and Chevalier, P. Towards sophisticated wrapping of web-based information repositories. In *Proceedings of 5th International RIAO Conf.*, pages 123-135. 1997.

[8] Naveen Ashish and Craig A. Knoblock. Semi-automatic wrapper generation for internet information sources. In *Proceedings of the Second IFCIS International Conference on Cooperative Information Systems (CoopIS)*, Charleston, SC,

1997.
[9] Microsoft, "Microsoft Win32 Internet Functions Reference", MSDN Online Web Workshop, 2000.
[10] Microsoft, "ADO Programmer's Guide", MSDN Library, 2000.

윤 종 완
1993년 한국과학기술원 전기 및 전자공학과(학사). 1996년 한국과학기술원 정보 및 통신공학과(공학석사). 1996년 ~ 1998년 현대전자 정보시스템사업본부 연구원. 1998년 ~ 2001년 현대정보기술 정보기술 연구소, 선임 연구원. 2001년 ~ 현재 코리아와이즈넷 연구원. 관심분야는 인터넷 정보 추출 및 통합 시스템, 검색 시스템, 무선 데이타 통신

이 종 우
1990년 서울대 컴퓨터공학과 졸업(학사). 1992년 서울대 컴퓨터공학과 석사과정 졸업(석사). 1996년 서울대 컴퓨터공학과 박사과정 졸업(박사). 1996년 ~ 1998년 현대전자산업(주) 과장. 1998년 ~ 1999년 현대정보기술(주) 책임연구원. 1999년 ~ 현재 한림대학교 정보통신공학부 조교수. 관심분야는 운영체제, 병렬/분산 운영체제, 클러스터 시스템, 전산금융

박 찬 영
1987년 서울대학교 전자공학과 학사. 1989년 한국과학기술원 전기 및 전자공학과 석사. 1995년 한국과학기술원 전기 및 전자공학과 박사. 1991년 ~ 1999년 삼성전자 정보통신총괄 네트웍사업부 선임연구원. 1999년 ~ 현재 한림대학교 정보통신공학부 조교수. 관심분야는 고속 통신망, ATM, 통신 ASIC, 통신 프로토콜, xDSL, 차세대 인터넷