내장 문자와 사전 구조 지식을 이용한 HTMLtoVXML 변환 에이전트 개발

장 영 건†

요 약

본 연구는 음성 인터넷 서비스를 위하여 기존의 HTML 컨텐츠를 자동으로 VXML 컨텐츠로 변환하여 사용하는 방법에 관한 것이다. 문서 변환은 HTML 문서의 해석과 내용 분리에 앞서 컨텐츠의 선택이 필수적으로 요구되지만, 이미 알고 있고, 관심이 있는 내용의 집단적 선택에 대하여는 아직까지 좋은 해법이 없어 보인다. 본 논문에서는 비구조적 데이터로 구성된 웹 문서로부터 필요한 정보 묶음을 선택하기 위하여 웹 문서에 포함된 문자열, 구조에 대한 사용자의 사전 지식을 상호 반응적으로 사용하는 방법을 제안하며, 이를 구현하여 그 타당성을 입증하였다. 제안하는 방식은 기존의 구조적 방식에 의한 선택보다 사용자의 의도를 정확히 반영하며, 문서 표현의 구현 기법의 변화에 강건한 장점이 있다. 또한 컨텐츠 분리 측면에서는 XML 또는 XHTML 문서변환을 중간 과정으로 사용하는 방식에 비하여 시간적으로 빠르며, 변환 과정의 부담이 작다.

Development of HTMLtoVXML Conversion Agent using Embedded Text and Priori Structural Knowledge

Young Gun Jang†

ABSTRACT

This paper presents a new agent which convert HTML contents to VXML contents automatically for voice services via web. In this paper, I propose an interactive hybrid sequential contents selection method to select desired contents fast and robustly from known web pages. It uses real time structural features as well as embedded text and/or priori structural knowledge such as link symbol position. To verify its effectiveness, a full agent system is implemented and tested. The method reflects user intention more accurately than conventional selections using structural features and is more robust to variations of HTML programming techniques. The agent is fast and has less computational burden than methods use XML or XHTML conversion as intermediate stage.

키워드 : HTML, VXML, 변환 에이전트(Conversion Agent), 컨텐츠 선택(Web Contents Selection), 컨텐츠 분리

1. 서 론

인터넷과 인트라넷에서 수행되는 작업은 그들이 포함한 컨텐츠나 그 정보를 사용하는 사람만큼 다양하다. 이 다양성, 정보와 서비스의 복잡성 때문에 일상적으로 반복되는 인터넷 작업을 자동화하려는 시도가 증가하고 있다. 이러한 시도는 특히 웹 컨텐츠의 가공자 입장에서는 급속하게 증가하는 웹 컨텐츠에 대하여 컨텐츠 제작, 관리의 비용과 시간을 절감하고, 인간 요소에 의한 실수를 최소화하는 입장에서 필수적인 것이 되고 있다. 웹 자동화는 폼을 채우는 것을 포함하여 임의적 웹 활동을 자동화해야 하며, 계획에 따라 또는 요구에 따라 동작해야 한다. 그러나 인터넷에서 운용되는 정보는 정적이지 않고, 지속적으로 내용이 변화하며, 그 구조조차 바뀌는 일이 비일비재하다. 웹 페이지의 완전한 자동 해석은 HTML이 갖는 컨텐츠의 내용이나 구조적 성질보다는 시각적 표현을 위주로 기술되었다는 성질 때문에 현재까지의 기술로는 불가능한 것으로 알려져 있다[1].

VXML의 등장 전에 그 원형 중에 하나인 VoxML이 모토로라에 의하여 발표되었으며, HTML을 VoxML로 변환하려는 연구가 Goose 등에 의하여 최초로 이루어졌다. Goose 등은 WWW의 전형적인 3층 구조에 HTML을 VoxML로 변환하는 기능을 갖는 VoxML-Agent를 추가하여 클라이언트, 에이전트, 웹 서버, 데이터베이스의 4층 구조를 갖는, 전화기를 사용하여 웹 접근이 가능한 Vox 포탈에 대한 연구를 발표하였다. VoxML-에이전트에 대해서는 Vox 포탈과의 상호작용에 대하여 중점적으로 언급되어 있고, 핵심 부분인 변환 기능의 설계에 대한 내용은 없다[2].

시각적 인터페이스를 주 수단으로 하는 HTML 문서를 해석하여 음성 인터페이스를 부여하는 연구는 최근에 이루어지고 있으며, 보편적 접근을 위한 인터넷 내용의 적응 및 접근성 향상 도구를 위주로 한 문법적 속성에 근거한 코드 변환[3], HTML의 구조적 성질에 기반한 코드 변환[4-7], 의

† 정 회 원 : 청주대학교 컴퓨터정보공학과 교수
논문접수 : 2002년 9월 30일, 심사완료 : 2002년 11월 25일

미론적 코드 변환[8, 9], 수작업에 의한 주석코드 추가[1, 10] 등의 접근이 이루어지고 있으나, 좀 더 쉬운 항해와 이해를 위하여 웹 페이지의 재배치, 주석 달기 및 웹 객체를 변환하는 것을 목적으로 하고, 특정한 사용자 그룹에 대한 특별한 필요성의 해결을 목표로 하고 있어 총체적 일반화에는 이르지 못하고 있다. HTML 코드만을 대상으로 시각적으로 표현되는 레이아웃을 추정하고, 각각의 내용들을 분리해 내는 것은 컴퓨터는 물론이고, HTML에 대한 지식을 가진 인간도 완벽하게 처리할 수 없다. 따라서 변환 대상이 한정되며, 대상의 선택, 추출 및 분리에 지능적 기법이 요구된다. 변환기는 컨텐츠의 선택, 선택된 컨텐츠의 내용 분리와 추출, 추출된 내용을 미리 정의된 시나리오에 의하여 VXML 문서로 변환하여 생성하는 과정을 거친다. 변환기를 구성하는 핵심 기술은 사용자가 원하는 컨텐츠를 HTML의 표현방식에 관계없이 견고하게 컨텐츠를 선택하고 추출하는 방식과 HTML 문서의 통계적, 문법적, 구조적 특성을 활용하여 내용을 신뢰성이 있게 분리하는 것과 실용적 측면에서 가능한 한 한번의 접속으로 많은 컨텐츠를 변환하는 자동화율의 제고 기술이다. 내용 분리에 대하여 Embley[5], Buttler[6]와 최훈일[7]은 멀티 컨텐츠를 갖는 HTML 문서에서 게시판처럼 비슷한 형태의 내용이 나열되어 있는 멀티 컨텐츠를 내용 단위로 분리하는 방법을 제시하였으며, HTML 문서의 구조적 특징을 이용하여 내용 단위 분리의 기준이 되는 분리 태그를 추출하는 휴리스틱 알고리즘들을 제안하였다. 제안된 휴리스틱 알고리즘의 한계성 때문에 최훈일은 변환기의 적용 대상으로 멀티 컨텐츠의 형식을 게시판, 리스트 및 검색 결과로 한정하였다. 변환이 성공적이기 위해서는 한정된 적용대상을 하나의 웹 페이지에서 안정적으로 선택해야 한다. 그러나 구조적 특성을 이용한 컨텐츠 선택에 대한 접근 방식의 경우 하나의 웹 페이지의 다양한 표현 방식과 내용의 변화에 따라 컨텐츠의 선택이 달라지는 문제점이 있다. Embly와 Goose가 사용한 최소 자식 노드 방식은 사용자의 의도를 반영하지 못하는 근본적인 문제를 포함하고 있으며, 최훈일이 제안한 최소 자식 노드와 포함된 문자의 수의 합성에 의한 노드 선택 방식 역시 노드의 수와 포함된 문자수의 균형에 따라 선택이 달라질 수 있다. 따라서 구조적 특성을 이용하는 것 이외의 방법이 복합적으로 동원되어야 한다. 이 문제에 대한 연구로써 의미론적 접근 방식으로는 온토로지(ontology)를 사용하는 방식이 있다[8, 9] 이 방식은 의미를 기반으로 하기 때문에 자료가 XML 기반으로 서술되어야 하며, 온토로지를 해석하기 위한 RDF를 필요로 한다. 따라서 이미 알고 있는 웹 페이지의 컨텐츠 선택에 있어서는 온토로지를 기술하는 것이 복잡하며, 별도의 RDF가 필요하고, 기술언어와 문법에 대한 전문성이 요구되고, 정확하게 온토로지의 규격과 일치하지 않는 경우에는 내용상의 모호한 변화에 민감하게 반응하는 문제점이 있다[9]. 사용자의 의도를 반영하는 방법으로써 Bruce의 경우 사용자와 웹 페이지 사이의 상호 반응을 에이전트가 기록하여, 나중에 웹 페이지에 접속하였을 때 사용자가 선택한 동일한 작업을 에이전트가 수행하는 방법을 사용하였다[11]. 이 방식은 그 구현이 복잡한 단점이 있으며, 자동화 프로세스 중에는 HTML 위치정의 언어로 변환된 결과를 이용한다. MIT의 Lieberman은 예를 제시하여 훈련을 통하여 문자를 에이전트에 인식시키는 연구를 수행하였다[12] 이 방식은 비구조화된 웹 정보의 중간에 있는 의미 있는 문자의 패턴과 처리 방식을 에이전트에게 예를 제시하여 훈련시킴으로써 수동적 문법 작성에서 발생하는 전문성 요구와 오류 가능성을 제거하였다. 컨텐츠 선택에 대한 여러 연구들은 각각 장점을 갖고 있지만, 실용적 측면에서는 사용자에게 불필요한 정보 정의를 요구하거나 불필요한 인터페이스를 제공할 수 있으며, 구현의 복잡성 등으로 본 연구에서 필요한 견고하고, 편리한 컨텐츠 선택 방식은 아니다.

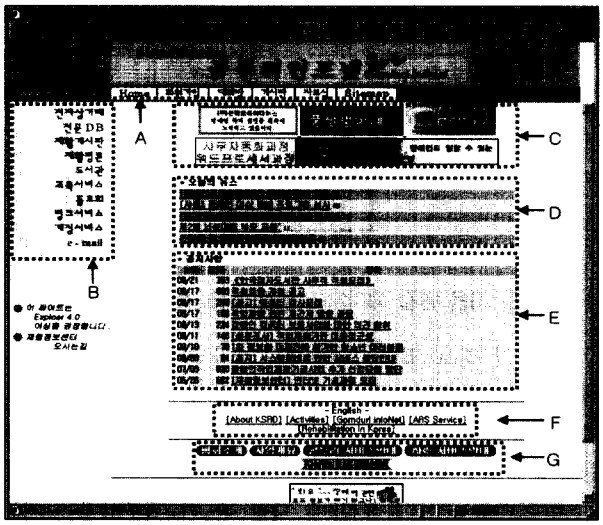
본 연구에서는 ① 전술한 컨텐츠 선택의 신뢰성 문제를 해결하기 위하여 구조적 특성과 함께 사용자의 사전 지식이나 에이전트와의 상호반응을 이용하며, 문자 정보를 이용하여 내용 기반 접근 방식의 포괄성과 명확성을 확보하고, 예를 통한 문자 인식 방식의 간편성과 전문적 지식을 필요로 하지 않는 것과 같은 장점을 유지하면서 각 방식의 단점을 보완하는 새로운 순차적 혼성 컨텐츠 군 선택 방식을 제안한다. ② 변환기의 자동화율을 높이기 위한 방식으로써 해석하고 있는 문서에서 다음 문서와의 연결관계에 대한 사용자의 사전지식을 단계에 따라 상호반응적으로 제공하는 스키마기반의 연속 문서 인식 방식을 제안한다. ③ 앞서 제안한 ①과 ②의 방식이 적용되는 HTMLtoVXML 변환기의 전체 구조, 기본적 체계인 내용 분리의 방식과 분리된 내용을 정해진 시나리오에 따라 VXML로 변환하는 구체적인 구현 방식을 제시한다. 2장에서는 변환 대상 문서유형을 제시하고, 3장과 4장에서 본 논문에서 제안한 컨텐츠 선택 방식과 연속 웹 문서 인식 방식을 서술하며, 5장에서는 제안한 방식과 결합한 HTMLtoVXML 변환 에이전트의 구조와 구현 방식을 서술하고, 내용 분리 방식을 서술할 것이다.

2. 변환 가능한 HTML 문서 유형과 구조 특성에 의한 선택

웹 저작자는 다양한 시각적 효과를 사용하여 가능한 많은 정보를 한 페이지에 표현하고자 한다. 이 정보는 시각적으로 단체화하여 조각나 있다. 예를 들면 (그림 1)과 같은 웹 페이지는 A, B, C, D, E, F, G의 조각이 하나의 페이지에 표현된다. 이 조각 중에서 어느 조각이 이 페이지에서 제공하고자 하는 주요 정보인지를 판단하고, 해당 정보를 내용 단위로 분리하여, 그 내용 단위를 기준으로 VXML[13] 문서로 변환하여야 한다.

HTML을 VXML로 변환하기 위해서는 먼저, HTML 페이지를 컨텐츠의 내용을 중심으로 조각 단위로 나누고 이 조각 단위로 VXML 문서를 생성해야 한다. 그러나 HTML 태그는 정보의 시각적 표시 방법만을 나타낼 뿐 XML 태그

처럼 정보에 대한 의미를 포함하고 있지 않기 때문에 조각 단위로 분리하기가 어렵다.



(그림 1) 웹 페이지 형태

따라서 본 논문에서는 컨텐츠가 시각적으로 보여지는 내용별로 단체화되어 구성되었다고 가정하였다. 이 가정은 대부분의 웹 페이지의 코드 검증을 통하여 개연성이 입증되었다. 조각 단위로 분리된 컨텐츠에 대한 상세한 내용분리는 태그의 구조적 특징을 통계적으로 분석한 발견적 기법을 사용하였다. 따라서 이 기법을 사용하여 분리할 수 있는 유형은 제한되며, 정성적으로는 비슷한 형태의 컨텐츠가 나열되어 있는 형태이다. 본 논문에서 사용한 변환 방식을 적용할 수 있는 문서 형태는 게시판 유형, 리스트 유형, 검색결과 유형으로 나눌 수 있다. 이런 유형의 HTML 문서는 트리 구조로 표현할 때 동일한 형태의 자식 노드를 많이 갖고 있고, 그 내용 속에 많은 문자를 포함하고 있는 점을 이용하여 컨텐츠의 위치를 구할 수 있다. (그림 1)에서 VXML로 변환 가능한 내용은 공지사항(E)과 오늘의 뉴스(D)이다.

HTML 문서를 구조적으로 분석하기 위한 첫 번째 단계로써 트리 구조로 재구성한다. HTML 문서를 트리 구조로 구성하는 이유는 트리 구조가 HTML 문서 구조를 분석하기가 좀더 용이하기 때문이다.

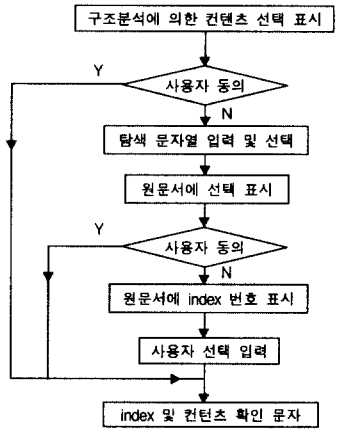
웹 페이지에서 구조적 특징만을 사용하여 주요 컨텐츠 군을 선택하는 방법으로써 변형된 최소 서브 트리 방식을 고안하였다. 컨텐츠를 포함하는 최소 서브 트리란 전체 트리 구조에서 컨텐츠를 포함하며, 크기가 가장 작은 트리를 말한다. 컨텐츠를 추출하기 전에 컨텐츠가 포함되어 있는 서브 트리를 먼저 추출하는데 이 서브 트리는 (그림 1)에서의 각각의 조각에 대응된다. 각 노드의 자식 노드 수를 구하여 가장 많은 자식 노드를 갖는 노드를 루트로 하는 트리를 최소 서브 트리로 한다. 보통 컨텐츠가 나열되어 있는 형태의 웹 페이지에서는 자식 노드의 수가 가장 많은 노드를 루트로 하는 서브 트리에 주요 컨텐츠가 포함되어 있을 확률이 가장 많다. 그러나 최소 서브 트리를 추출할 때 자식 노드의 수만을 고려할 경우 메뉴가 많은 웹 페이지에서는 최소 서브 트리가 메뉴를 포함하는 서브 트리가 된다.

본 논문에서는 많은 자식 노드를 갖는 노드의 내부 텍스트 크기를 이용하여 최소 서브 트리를 추출하는 방법을 제안한다. 즉, 자식 노드를 많이 가지고 있는 노드들을 구하여 이 노드 내에 있는 텍스트의 크기를 구하여 텍스트의 크기가 가장 큰 노드를 루트로 하는 서브 트리를 최소 서브 트리로 한다. 이 방법을 이용하면 메뉴와 같이 나타내는 텍스트의 크기가 작고, 자식 노드가 많은 서브 트리를 최소 서브 트리로 지정하는 오류를 피할 수 있다.

3. 내용 기반의 상호 반응적 컨텐츠 군의 선택

HTML로 작성된 문서의 구조적 해석을 통한 콘텐츠 군의 선택은 어떠한 규칙 기반을 사용하여도, 일반적으로 사용자의 의도를 적용할 수는 없고, 하나의 웹 페이지에서 다수의 컨텐츠를 선택할 수 없다. 따라서 구조적 특징을 사용하여 자동으로 선택된 컨텐츠가 사용자의 의도와 맞지 않을 때, 그 의도를 반영하여 다시 컨텐츠를 선택할 수 있는 메커니즘이 필요하다. 제안하는 방식은 사용자의 상호 반응적 선택을 원칙으로 하여 자동 선택과 다중의 사용자 선택을 적용 빈도에 따라 순차적으로 제공하는 혼성 선택 방식이다. 첫 번째 구조적 방식을 적용하여 그 적합성을 선택하고, 맞지 않는 경우에는 두 번째 문자정보에 의한 예 제시를 통한 선택을 행하며, 그 결과를 웹 페이지의 합성 결과로써 확인하는 방식을 취한다. 이 때 예로 제시된 문자열은 변환결과로써 생성되는 문서들의 제목이 된다. 또한 결과를 확인하여 컨텐츠 군의 선택이 잘 못 되었을 경우, 선택은 번호 또는 변호와 논리적 연산으로 지정하는 방식을 취한다. 본 논문에서는 웹 페이지에 포함된 문자열을 예로써 사용하여 에이전트를 훈련시키는 방법을 사용하였다. 웹 문서의 저작자는 일반적으로 컨텐츠의 내용적 특성을 단순하고, 효과적으로 전달할 수 있는 키워드를 문자가 그래픽을 사용하여 사용자에게 전달하려고 한다. 따라서 이 방식은 저작자의 의도된 표현 능력과 사용자의 보여지는 웹 페이지에 대한 통찰력을 이용하며, 해당되는 컨텐츠의 키워드를 선택하고, 복사하여 에이전트에게 전달하여, 이미 구조 해석을 위하여 작성된 서브 트리 중에서 해당 키워드를 포함하는 서브 트리를 변환 대상으로 지정한다. 그러나 그래픽 표현의 경우에는 키워드의 문자 복사가 불가능하며, 다중의 컨텐츠 선택이 필요하거나, 컨텐츠의 논리적 합성이 필요한 경우에는 이 방식을 채용할 수 없다. 컨텐츠의 합성은 잘못된 선택 또는 시각적으로는 같은 컨텐츠 군으로 보이지만 구조적으로 별개의 구조로 구성된 웹 문서에 특히 유용하게 사용될 수 있다. 문자열에 의한 선택이 문제가 되는 경우, 구조 해석의 결과에 따라 형성된 컨텐츠의 인덱스 번호를 원 문서에 포함시켜 합성된 웹 문서를 사용자에게 시각적으로 제시하고, 사용자의 판단에 따라 선택한 인

텍스 번호를 에이전트에게 전달하여 컨텐츠를 선택하는 방식을 취한다. 이 방식은 복수의 선택이 가능하며, 논리적 합성을 지원한다. 논리적 합성 연산자는 병합을 의미하는 + 연산자만을 사용하였으나, 필요에 따라 추가가 가능하다.

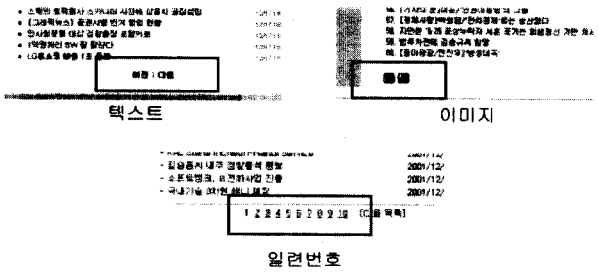


(그림 2) 순차적 혼성 컨텐츠 군 선택 방식의 흐름도

해당 구현 알고리즘은 (그림 2)에 표시하였다. 본 방식은 기본적으로 트리 구조와 같은 구조 해석 방식을 노드 수와 문자수와 같은 정보를 이용하는 규칙 기반과 결합하였고, 해당 규칙기반이 근본적으로는 사용자 의도를 반영하지 않으며, 동일한 시각적 효과를 내는 웹 문서 저작 기법이 다수 존재하며, HTML의 해석에서 가능한 모든 기법에 대한 적응이 사실상 불가능하다는 점에서 자동 선택이 실패할 경우에는 예를 통한 의미 기반적 접근과 시스템의 해석 결과를 사용자가 상호 반응적으로 선택하는 순차적 혼성 기법을 채용한 것이 특징이다.

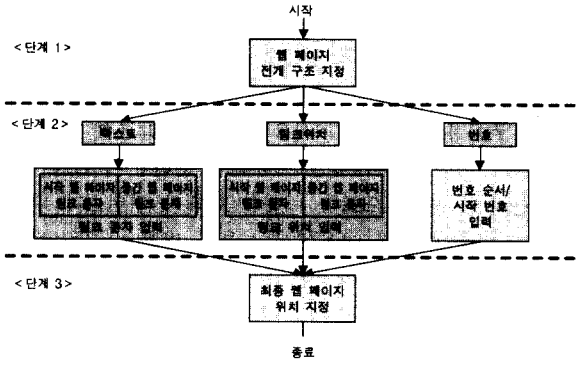
4. 사전지식을 이용한 상호반응적 스키마 기반의 연속적 웹 문서 인식

일반적으로 동일한 성격과 구조를 갖는 컨텐츠의 수가 많을 경우 하나의 웹 문서에 표시할 수 있는 컨텐츠의 수는 한정되어 있기 때문에 해당 컨텐츠는 여러 개의 웹 문서로 구성되며, 이 웹 문서들은 링크를 통해 서로 연결되게 된다. 그러나 HTML 코드만으로는 다음 문서에 대한 링크가 어느 것인지를 알 수가 없기 때문에, 한번에 자동으로 추출하기란 불가능하다. 이 문제를 해결하기 위해서는 다음 문서와의 연결관계에 대한 사용자의 사전지식을 단계에 따라 상호반응적으로 제공함으로써 전체 컨텐츠를 좀 더 쉽게 추출할 수가 있는데, 이런 사전지식 정보는 HTML 코드를 분석하여 해당 링크 태그의 경로를 지정하는 것이 아니라 웹 문서에 표시되어 있는 마지막 컨텐츠를 기준으로 몇 번째의 링크인지와 그 링크의 형태에 대한 정보만을 제공함으로 쉽게 다음 문서의 URL을 추출할 수 있다. 이 때 웹 문서들간의 링크는 일반적으로 마지막 컨텐츠 다음에 위치하게 되며, 다음 문서에 대한 연결을 나타내는 링크의 형태는 (그림 3)과 같이 텍스트, 이미지 및 일련번호의 세 가지로 구분할 수 있다.



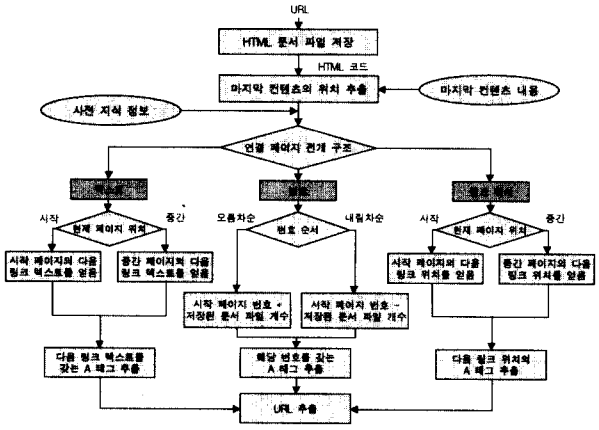
(그림 3) 다음 연결 페이지의 연결 링크 형태 및 링크 위치

사전지식 정보 입력 스키마는 (그림 4)와 같이 3 단계로 나눌 수 있다. 첫 번째 단계에서는 웹 문서들의 다음 문서에 대한 연결 링크 형태가 무엇인지를 지정하고, 두 번째 단계에서는 연결 링크의 형태에 따라 링크 문자, 링크 위치, 일련번호 순서 등 링크에 대한 좀 더 세부사항을 지정하고 마지막 단계에서는 몇 번째 문서까지를 추출할 것인지 즉, 마지막 문서의 위치를 지정한다.



(그림 4) 사전지식 정보 입력 스키마

사전지식이 주어지면 이를 이용하여 (그림 5)와 같은 과정을 통해 다음 연결 문서의 URL을 추출할 수 있다.



(그림 5) 다음 연결 웹 문서 URL 추출 흐름도

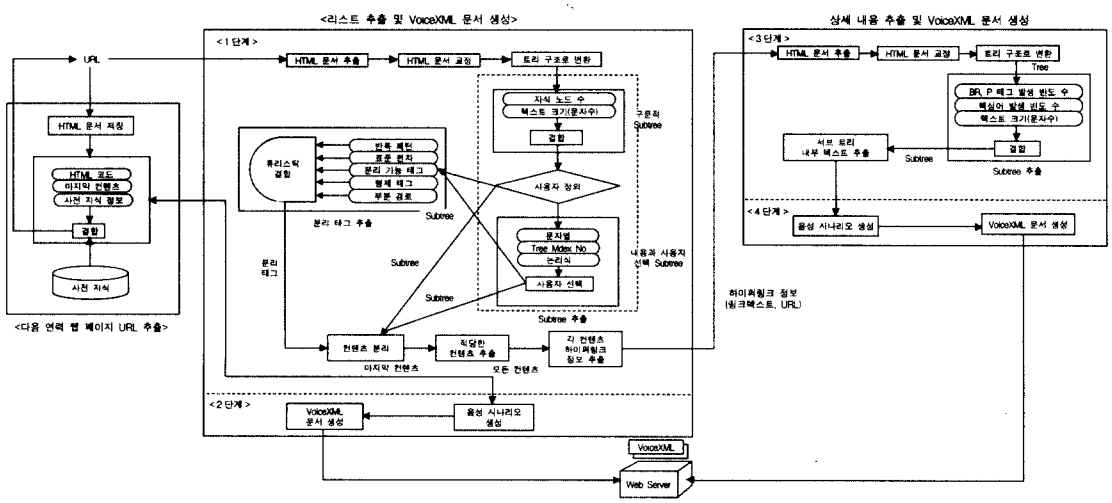
HTML 문서 구조에 대한 사전지식 정보 입력과정을 윈도우즈의 마법사와 같은 형태로 제시하였다. 사전지식 입력 마법사도 입력 스키마처럼 3 단계로 나눌 수 있다. 1 단계는 연결 링크의 형태 정보를 입력하고, 2 단계는 선택된 연결 링크 형태에 따라 텍스트일 경우는 텍스트 정보를, 이미지일 경우는 링크의 위치 정보를, 일련번호일 때에는 일련번호 정보를 입력한다. 마지막 3 단계는 추출할 최종 웹 문서 정보를 입력한다.

5. HTMLtoVXML 변환모듈 설계 및 구현

HTML 문서를 다른 마크업 문서로 변환하는 가장 간단한 방법은 HTML 문서의 각 태그들을 다른 마크업 언어의 비슷한 역할을 하는 태그와 1 : 1로 매칭 하여 변환하는 것이다. 이러한 방법을 이용한 변환은 HTML 문서를 무선 인터넷용 마크업 문서로 변환할 때 주로 적용된다[14]. 그러나 HTML 태그는 태그 내에 있는 컨텐츠의 시각적인 표현 방법만 기술할 뿐 의미를 제공하지 않기 때문에 태그별 1 : 1 매칭에 의한 변환은 자연스러운 정보 제공이 힘들다. 다른 방법으로는 사용자의 웹 문서 이동 경로를 기록하여 그 경로를 이용하여 원하는 컨텐츠를 쉽게 찾아 원하는 다른 마크업 언어로 자연스러운 정보 제공이 되도록 시나리오를 재구성하는 방법이다. 이때 웹 문서의 이동 경로는 HTML 문서의 트리 구조에서 html 또는 body 태그를 루트로 하는 절대경로로 기록하게 된다[15]. 이러한 방법의 문제점은 이동 경로에 해당되는 태그들의 삭제, 변경, 삽입 등으로 인해 전혀 원하지 않은 결과를 초래하게 된다. Embley, Buttler와 최훈일은 게시판이나 검색결과처럼 비슷한 형태의 내용이 나열되어 있는 멀티 컨텐츠를 문서의 구조적 특징을 이용하여 내용 단위로 분리하는 방법을 제시하였으나, 하나의 웹 문서만을 대상으로 하기 때문에 여러 웹 문서에 걸쳐 있는 연속적인 컨텐츠를 분리하기 위해서는 각 웹 문서를 하나하나 접속해야 하는 단점이 있다.

본 논문에서 구현한 HTMLtoVXML 변환 에이전트는 (그림 6)과 같이 크게 리스트 추출 및 VXML 문서 생성, 상세 내용 추출 및 VXML 문서 생성, 다음 연결 웹 문서 URL 추출의 세 과정으로 나눌 수 있다. 리스트 추출 및 VXML 문서 생성 과정은 리스트 HTML 문서에서 리스트 컨텐츠를 추출하여 이를 리스트 시나리오에 맞는 VXML 문서를 생성하는 것이다. 이 부분의 1 단계에서 트리구조 변환 이후에 점선으로 둘러싼 부분이 컨텐츠 선택 부이며, 본 논문에서 제안한 순차적 혼성 컨텐츠 선택 알고리즘으로 구현되었다. 이 단계에서는 상대적으로 처리 속도가 빠른 2장에서 제안한 구조 특성을 이용한 컨텐츠 선택 방식이 적용되며, 원하는 내용이 아닐 경우에는 순차적으로 4장에서 제시된 다른 방식이 적용된다. 상세 내용 추출 및 VXML 문서 생성 과정은 리스트 컨텐츠에 대한 상세 내용 문서에서 상세 내용을 추출하여 이를 상세 내용 시나리오에 맞는 VXML 문서로 생성한다. 다음 연결 웹 문서 URL 추출 과정은 컨텐츠가 여러 웹 문서에 걸쳐있을 경우 리스트 HTML 문서 구조의 사전지식을 통해 다음 연결 웹 문서의 URL을 구해 리스트 추출 및 VXML 문서 생성 과정으로 그 값을 전달하여 한 번의 접속으로 모든 컨텐츠 추출이 가능하도록 한다. 이 부분은 기존의 연구에서는 제시되지 않았으며, 본 논문에서 제안한 연속된 웹 문서 인식 방식에 의하여 결정된 연결 법칙과 정보를 사용하여 연속된 웹 문서를 추출하고, 그 정보를 1단계로 넘겨 앞서의 과정을 반복적으로 수행하게 한다. (그림 6)에서 점선으로 표시된 컨텐츠 선택 블록과 실선으로 표시된 다음 연결 웹 페이지 URL 추출 블록은 반복적 접속 과정에서는 최초의 해석 결과를 이용하는 데이터 블록으로 대치되며, 반복적 접속은 시간적 계획기(scheduler)에 의하여 이루어진다.

컨텐츠를 내용 단위로 분리하기 위한 분리 태그를 추출하



(그림 6) HTMLtoVXML 변환 모듈 구성도

기 위해 먼저 컨텐츠 분리 후보 태그들을 추출해야 한다. 분리 후보 태그는 선택된 서브 트리의 루트 노드의 자식(Child) 노드를 컨텐츠 분리 후보 태그로 한다.

컨텐츠 분리 후보 태그가 하나일 경우에는 그 후보 태그가 바로 컨텐츠를 분리하기 위해 경계가 되는 분리 태그가 되지만 둘 이상일 경우에는 분리 후보 태그들 중 적절한 태그를 분리 태그로 선택을 해야 한다. 여러 개의 분리 후보 태그 중에서 적절한 분리 태그를 추출하는 방법은 HTML 문서를 작성할 때 나타나는 몇 가지 규칙을 기반으로 하는 휴리스틱 알고리즘을 이용하여 분리 태그를 추출한다. 본 논문에서 사용한 휴리스틱(Heuristic)은 각 분리 후보 태그들 사이의 글자 수에 대한 표준 편차, 각 컨텐츠 사이에는 일관성 있게 반복적으로 나타나는 태그 쌍들의 반복 패턴, 후보 태그 노드에서 임의의 다른 노드까지의 모든 경로를 목록으로 만들고, 각각의 식별된 경로의 발생횟수로 나타낸 부분 경로, 서브 트리에서 인접한 형제 태그 쌍의 발생횟수를 세어 내림차순으로 모든 태그 쌍들의 등급을 분류하는 형제 태그, 컨텐츠를 분리할 때 경험적으로 자주 사용되는 태그를 등급이 높은 순서대로 재배열하여 분리 후보 태그들의 등급을 분류하는 분리 가능 태그를 이용한다. 각 휴리스틱들은 특정 형태의 웹 문서에서 최적화되어 있고, 다른 휴리스틱들과 독립적이다. 그러므로 웹 문서의 올바른 컨텐츠 분리 태그를 추출할 가능성을 향상시키기 위해 각각의 독립적인 휴리스틱을 결합시키는 것을 고려한다.

다섯 가지의 휴리스틱에 대한 최상의 결합 상태를 결정하기 위해, 스탠포드 확신도 이론을 이용한다[16] 5개의 각 휴리스틱을 결합하는 방법에는 총 26가지가 있으나 5개의 휴리스틱을 모두 결합하여 분리 태그를 추출하는 것이 가장 정확한 방법이다.

분리 태그를 기준으로 컨텐츠를 내용 단위로 분리할 때 각 컨텐츠의 시작점과 끝점을 추출해야 한다. 즉, 분리 태그의 시작 태그와 바로 다음에 나타나는 분리 태그의 시작 태그를 컨텐츠의 시작점과 끝점으로 할 것인지, 아니면 분리 태그 이전의 내용부터 바로 다음에 나타나는 분리 태그의 시작 태그를 시작점과 끝점으로 할 것인지를 결정하여 각 컨텐츠의 내용을 분리해야 한다. 본 논문에서 제안하는 분리 태그를 기준으로 컨텐츠를 분리하는 과정은 다음과 같다.

```
1. 최소 서브 트리의 루트 노드의 첫 번째 자식 노드 N을 구한다.
2. repeat
     if (N == STYLE 태그 or SCRIPT 태그 or !(주석) 태그)
        or ( N == BR 태그 and BR 태그 != 분리 태그)
        or ( N == HR 태그 and HR 태그 != 분리 태그)
        or ( N == P 태그 and P 태그 != 분리 태그)
     then
        N := 다음 형제 노드
        continue /* if문 다시 수행 */
     else
        repeat문 종료
     end if
   until end-of-최소 서브 트리
3. if N == 분리 태그 then
     컨텐츠 := 첫 번째 분리 태그의 시작 태그와 바로 다음에 나
               타나는 분리 태그의 시작 태그 사이의 내용
   else
     컨텐츠 := 첫 번째 분리 태그 이전의 내용부터 다음에 나타나
               는 분리 태그의 시작 태그 사이의 내용
   end if
```

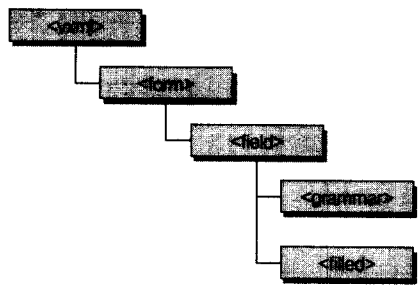
분리 태그를 기준으로 컨텐츠의 내용을 분리하면 적절하지 못한 내용들도 포함될 수 있다. 그래서 분리된 컨텐츠에서 적절한 컨텐츠만을 추출하여야 한다. 적절한 컨텐츠를 추출하는 방법은 컨텐츠가 나열되어 있는 웹 페이지의 각 컨텐츠의 텍스트는 숫자나 날짜, 문자들이 일정한 형태로 나타나게 되는데 이런 특징을 이용하여 적절한 컨텐츠를 추출한다. 본 논문에서 제안하는 적절한 컨텐츠 추출 방법은 다음과 같다.

```
1. repeat /* 모든 분리된 컨텐츠를 대상으로 함 */
     분리된 각 컨텐츠의 텍스트 항목의 형태를 숫자형, 날짜형, 문
     자형으로 구분하여 각 컨텐츠 의 텍스트 항목 형태 정보를 F
     배열에 저장
   until end-of-분리된 컨텐츠
2. F 배열에 저장된 형태 정보 중 같은 형태의 개수를 구함.
3. 개수가 가장 많은 형태 정보를 구하여 이 형태 정보를 갖는 컨텐
   츠를 추출
```

컨텐츠 추출과정을 통해 추출된 컨텐츠는 음성 인터페이스를 통해 사용자에게 정보를 제공하게 된다. 음성적 정보 제공 방법은 시각적 정보 제공 방법보다 사용자가 한번에 이해할 수 있는 정보의 양이 제한적이기 때문에 추출된 컨텐츠의 양에 따라 N개의 음성 시나리오가 생성된다.

본 논문에서는 추출된 컨텐츠의 평균 문자수가 100개 미만이면 4개의 컨텐츠로 하나의 음성 시나리오를 구성하고 100개 이상이면 2개의 컨텐츠로 하나의 음성 시나리오를 구성하도록 하여 N개의 시나리오를 생성하도록 하였다.

각 음성 시나리오는 VXML 1.0 형식에 맞춰 VXML 문서를 생성한다. 생성되는 VXML 문서의 기본 구조는 (그림 7)과 같다.




```
<?xml version = "1.0"?>
<vxml version = "1.0">
<form>
<field>
   리스트 내용 프롬프트
   메뉴 프롬프트
<grammar> 음성 메뉴 항목 </grammar>
```

```
<filled>
    메뉴 선택에 따른 조건문
</filled>
</field>
</form>
</vxml>
```

(그림 7) 생성되는 리스트 VXML 문서의 기본 구조

6. 실험 및 결과

본 논문에서 구현한 HTMLtoVXML 변환 에이전트의 성능을 시험하기 위하여 일반적인 접근 빈도가 높고 본 논문에서 변환 대상으로 규정한 게시판 유형, 리스트 유형과 검색 결과 유형을 갖는 구인/구직 사이트, 신문 사이트, 검색 사이트를 시험 대상 사이트로 결정하였다. 각각의 유형에 대하여 시험 대상으로 선정한 구체적인 사이트는 <표 1>과 같고, 제시한 3가지 유형을 포함하여 통합적으로 시험한 사이트는 <표 2>와 같다.

〈표 1〉 시험 대상 웹 사이트

구분	웹 사이트 명	웹사이트 URL
구인/구직 (게시판 유형)	인크루트	www.incruit.com
	리크루트	www.recruit.co.kr
	헬로잡	www.hellojob.com
	잡엑스	www.jobex.co.kr
	데브피아	www.devpia.com
신문 (리스트 유형)	동아일보	www.donga.com
	중앙일보	www.joins.com
	스포츠투데이	www.sportstoday.co.kr
	국민일보	www.kukminilbo.co.kr
	한겨레 신문	www.hani.co.kr
검색 (검색결과 유형)	MSDN	msdn.microsoft.com
	한미르	www.hanmir.com
	알타비스타	www.altavista.co.kr
	구글	www.google.com
	네이버	www.naver.com

〈표 2〉 신뢰도 시험 대상 웹 사이트

웹 사이트 명	웹 사이트 URL
한국아이	www.hankooki.com
한국일보	www.hankooki.com/hankook.htm
일간스포츠	www.hankooki.com/dailysports.htm
서울경제	www.hankooki.com/sed.htm
Korea Times	www.hankooki.com/sed.htm

시험 방법은 GIT에서 적용한 방식과 동일하며, 적용한 알고리즘의 성공률을 두 단계로 계산한다[13] 첫째, 각 웹 사이트에 대하여 적용된 페이지 중에서 제시된 알고리즘을 적용하여 최고 점수를 받은 태그가 정확한 분리 태그인 페이지의 비율을 계산한다. 둘째, 이 각각의 발견적 알고리즘과 통합된 알고리즘에 대한 성공률을 결정하기 위하여 전체 웹 사이트에 대하여 평균을 구한다.

<표 1>에 열거된 웹 사이트에 대한 분리 성공률은 약 200개의 웹 페이지에 대하여 적용한 결과, HTML을 문법에 맞지 않게 서술한 페이지를 제외하면 100%의 성공률을 보였다. 그 이유는 열거된 웹 사이트가 표로 구성되어 있는 경우가 많고, 하나의 분리 후보 태그만을 갖는 경우가 많아 휴리스틱 알고리즘을 적용할 필요가 없는 페이지가 많았기 때문이다.

〈표 3〉 각 알고리즘의 확률적 순위와 결합 알고리즘의 신뢰도

휴리스틱 알고리즘	순위 1	2	3	4
분리가능태그	0.85	0.1	0.15	0.0
반복패턴	0.65	0.0	0.15	0.0
부분경로	0.95	0.0	0.0	0.0
형제태그	0.80	0.2	0.0	0.0
표준편차	0.70	0.2	0.1	0.0
결합알고리즘	0.99	0.02	0.01	0.0

따라서 좀 더 다양한 웹 페이지 저작 기법을 사용한 웹 페이지를 대상으로 분리 태그 추출 신뢰도를 시험하였다. <표 3>은 <표 2>에서 제시한 웹 사이트에서 약 200개의 웹 페이지를 대상으로 각각의 휴리스틱 알고리즘에 대한 확률적 순위와 휴리스틱 결합 알고리즘의 신뢰도를 표시하였다. GIT의 결합 알고리즘 신뢰도가 94%인 것과 비교해서 높은 성공률을 보인 이유는 적용한 웹 페이지의 유형이 GIT에 비하여 적은 것으로 보인다. GIT의 경우는 웹 페이지에 대한 정보가 없어 정확한 비교는 불가능하다. 시험 결과는 규칙성을 갖는 표나 탐색 결과의 링크 나열과 같은 컨텐츠에 대해서는 구현된 변환기가 매우 안정적이라는 것을 보여주며, 좀 더 다양한 방식으로 표현한 웹 페이지의 경우도 통합 휴리스틱 알고리즘을 적용하였을 때 99%의 신뢰도를 보여 거의 모든 웹 페이지가 변환이 되는 것을 확인할 수 있다. <표 3>은 시험에 사용한 웹 페이지를 대상으로 분리 후보 태그들에 대해 각 휴리스틱 알고리즘을 적용한 결과에 대한 신뢰도와 각 휴리스틱을 결합했을 때에 대한 신뢰도를 나타낸다.

〈표 4〉 각 웹 사이트의 사전지식 형태

웹 사이트 명	사전지식 형태
인크루트	일련번호, 이미지
리크루트	일련번호, 이미지
헬로잡	일련번호, 이미지
잡엑스	일련번호, 텍스트
데브피아	일련번호
동아일보	이미지
중앙일보	이미지
국민일보	이미지
한겨레 신문	일련번호
MSDN	일련번호
한미르	일련번호
알타비스타	일련번호
구글	일련번호, 텍스트
네이버	일련번호

사전지식을 이용한 다음 연결 웹 문서 추출에 관한 시험은 <표 1>의 웹 사이트를 대상으로 하였다. 다음 연결 웹 문서에 대한 사전지식 형태는 <표 4>에 나타난 것처럼 일련번호 형태가 가장 많았으며, 일련번호와 텍스트 또는 일련번호와 이미지처럼 두 가지 형태가 혼합된 형태도 많이 나타났다. 그리고 사전지식을 이용한 다음 연결 웹 문서 추출 결과는 100%의 성공률을 보였다.

7. 결 론

본 논문은 HTML로 작성된 기존의 컨텐츠를 이동 단말기 및 전화를 통하여 음성으로 제공하기 위하여 VXML을 사용하는 변환 시스템에 관한 것이다. VXML 문서로 변환하는데 드는 비용과 시간을 줄이기 위해 HTML 문서를 자동으로 VXML 문서로 변환하는 HTMLtoVXML 변환 에이전트를 설계하고 구현하였다.

HTML 문서에 표현된 정보의 의미를 알 수 없는 HTML의 제약성 때문에 변환 가능한 HTML 문서의 유형을 제안고, 구조적 특성을 이용하는 컨텐츠 선택 방식의 견고성문제와 사용자 의도를 반영하지 않는 문제점을 해결하기 위하여 순차적 혼성 컨텐츠 선택 방식을 제안하고 구현하여 그 실효성을 입증하였다. 문서의 구조 분석을 통해 컨텐츠를 분리, 추출하여 음성 시나리오에 맞게 VXML 문서로 변환하였다. 또한, 여러 웹 문서에 걸쳐 있는 동일한 형태 컨텐츠는 웹 문서 링크 연결 구조에 대한 사전지식을 통해 한번에 모든 컨텐츠를 추출하여 좀더 실용적이고 유용한 음성 시나리오의 생성을 가능하게 하였다.

개발된 변환기를 사용하여 국내의 약 400여개의 웹 페이지를 대상으로 변환기의 기능과 성능을 시험하였다. 적용결과 HTML 문법을 정확하게 적용한 모든 웹 페이지를 정확하게 VXML 문서로 변환할 수 있었다. 따라서 기능의 유효성, 실용성 및 신뢰성이 있다고 판단된다.

몇몇 HTML 문법을 준수하지 않은 웹 페이지, 특히 시작 태그에 대한 종료 태그가 없는 경우는 제대로 변환되지가 않는데, 이를 해결하기 위해서는 HTML 문서를 XML 형식을 준수한 XHTML 문서로 변환하여 이를 VXML 문서로 변환하면 된다. 현재 구현된 변환기는 HTML 문서들이 갖는 많은 문제점에 의해 변환 가능한 대상 HTML 문서의 유형을 제한하였는데, 변환 대상을 확장하기 위해서는 좀더 지능적인 방법과의 결합이 요구된다고 판단된다.

참 고 문 헌

[1] Asakawa, "Annotation-Based Transcoding for Nonvisual Web Access," Pro. of ASSETS'00, pp.172-179, Nov., 2000.

[2] Stuart Goose, Mike Newman, Claus Schmidt, Laurent Hue, "Enhancing Web accessibility via the Vox Portal and a Web-hosted dynamic HTML<->VoxML converter," WWW9, Vol.33, No.1-6, pp.583-592, June, 2000.

[3] Mohan, R., Smith, J. & Li, C.-S., "Adapting multimedia internet content for universal access," IEEE Transactions on Multimedia, Vol.1, No.1, pp.104-114, March, 1999.

[4] Asakawa, et al, "User Interface of a Homepage Reader," Pro. of ASSET'98, pp.149-156, April, 1998.

[5] D. W. Embley, Y. S. Jiang and Y.-K. Ng, "Record-boundary discovery in Web documents," Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD'99), pp.467-478, May, 1999.

[6] David Buttler, Ling Liu, Calton Pu, "A Fully Automated Extraction System for the World Wide Web," IEEE ICDCS-21, April, 2001.

[7] 최훈일, 장영건, "HTMLtoVoiceXML 변환기 설계 및 구현", 한국정보과학회논문지 : 컴퓨팅의 실제, 제7권 제6호, pp. 559-568, 2001.

[8] Anita W. Huang, "A Semantic Transcoding System to Adapt Web Services for Users with Disabilities," Pro. of ASSETS'00, pp.156-163, Nov., 2000.

[9] Carole Goble, Sean Bechhofer et al, "Conceptual Open Hypermedia = The Semantic Web?," Proceedings of the 2nd Int. Workshop on the Semantic Web, May, 2001.

[10] Hori M., Kondoh G., Ono K., Hirose S. and Singhal S. "Annotation-based web content transcoding," In Proc. of WWW9, pp.197-211, May, 2000.

[11] Bruce Krulwich, "Automating the internet : agents as user surrogates," IEEE Internet Computing, Vol.1, No.4, pp.34-38, July-August, 1997.

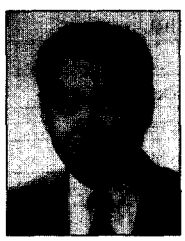
[12] Henry Lieberman, Bonnie A. Nardi, David Wright, "Training Agents to Recognize Text by Example," Proceedings of Agents'99, May, 1999.

[13] W3C, "Voice eXtensible Markup Language(Voice XML) version 1.0," http://www.w3.org/TR/voicexml, W3C Note 05, May, 2000.

[14] 이정은, 장지산, 김민수, 김성찬, 신동규, 신동일 "HTMLtoWML 변환기 시스템의 설계 및 구현", 2001년 한국정보과학회 춘계학술대회논문집, 제28권 제1호(A), pp.184-186, 2001.

[15] Juliana Freire, Bharat Kumar, Daniel Lieuwen, "WebViews : Accessing Personalized Web Content and Services," WWW10, pp.576-586, May, 2001.

[16] G. F. Luger, W. A. Stubblefield, "Artificial Intelli- gence : Structures and Strategies for Complex Problem Solving," Third Edition. Addison Wesley Longman, Inc., 1997.



장 영 건

e-mail : ygjang@chongju.ac.kr

1980년 인하대학교 전자공학과 학사

1979년~1983년 국방과학연구소 연구원

1983년~1994년 대우중공업 중앙연구소 책임연구원

1991년 인하대학교 전자공학과 석사(정보공학)

1995년 인하대학교 전자공학과 박사(정보공학)

1995년~1996년 고등기술연구원 책임연구원

1996년~현재 청주대학교 컴퓨터정보공학과 부교수

2002년~현재 UC Davis visiting Professor

관심분야 : HCI, CTI, 음성정보처리를 이용한 웹 프로그래밍, 재활공학