WebPress2000
(웹프레스 2000)

정보처리전문가협회장상

1. S/W 명 : WebPress2000(웹프레스 2000)

2. 제작자 :
   회사 : ㈜리드텍코리아
   주소 : 서울 서초구 양재동 87-7 풍한빌딩 2 층, 4 층
   전화 : (02)3142-4800

3. 개 요

   1) 개발 배경
   ① 인터넷 사용 인구의 폭발적 증가
   ② 공공기관, 기업체 등의 인터넷 서비스의 확산
   ③ 대학 및 공공도서관의 전자도서관 서비스 확대
   ④ 원문정보 서비스의 느린 전송속도, 저장용량의 비대화, 사용의 불편함 등을 대체할 서비스 형식의 필요성 대두
   ⑤ 인터넷 기반의 지식 관련 시스템의 입력프로그램 필요
   (전자문서관리시스템, 지식관리시스템 등)
   ⑥ Contents 제작 공정의 개선 필요
   (수작업, 복수의 프로그램 사용에 따른 번거로움 등)

   2) *WebPress*2000 은 ...

*WebPress*2000 은 활자화되어 있는 인쇄물을 어떤 Web Authoring Tool 보다 쉽고 빠르게 인터넷 표준규격인 HTML(Hyper Text Mark-Up Language)파일로 변환시켜 주는 소프트웨어 입니다. 즉, 인쇄된 문서(논문, 학술지, 잡지, 일반문서 등)를 스캐닝(Scanning)하여 이미지의 문서구조를 분석한 후, 텍스트 영역은 OCR(Optical

Character Recognition:문자인식)을 통해 텍스트로, 이미지 영역은 JPEG 이미지로 처리하여 HTML 로 자동변환함으로써, 변환된 결과에 대해 추가로 편집이나 수정을 하지 않고도 인터넷에서 그대로 활용할 수 있도록 합니다.

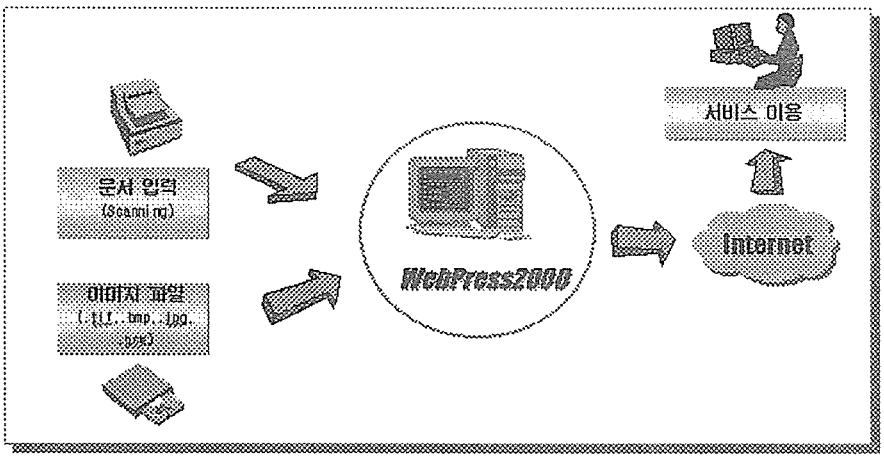

그림 1. *WebPress*2000 의 활용 개념

2. 기 능

1) 구 성

*WebPress*2000 은 통합 Web Authoring Tool 로서 7 개의 단위 기능으로 구성되어 있습니다.

①Image Acquire System

스캐너 컨트롤 / Image Format Converting / DIB Format 처리

②Image Clean-Up System

기울기 보정 / 이미지 회전 / 흑백반전 / 노이즈 제거 / 컬러이미지 속성 조정 (Hue, Saturation, Contrast, Brightness)

③문서영상 해석 **System**

문서 구조 해석 / 영역 좌표 정보 추출 / 영역 속성(Text, Table, Picture) 분류

④문자인식 **Engine**

Text 영역 및 Table 의 Text 의 문자인식 / 다중 폰트의 문자인식

⑤**HTML Code Generator**

HTML 문서의 자동 생성 / Auto Tagging

⑥**Text Editor**

Text 편집 / 맞춤법 기능 / 문자열 교정

⑦**Web Editor**

HTML 편집

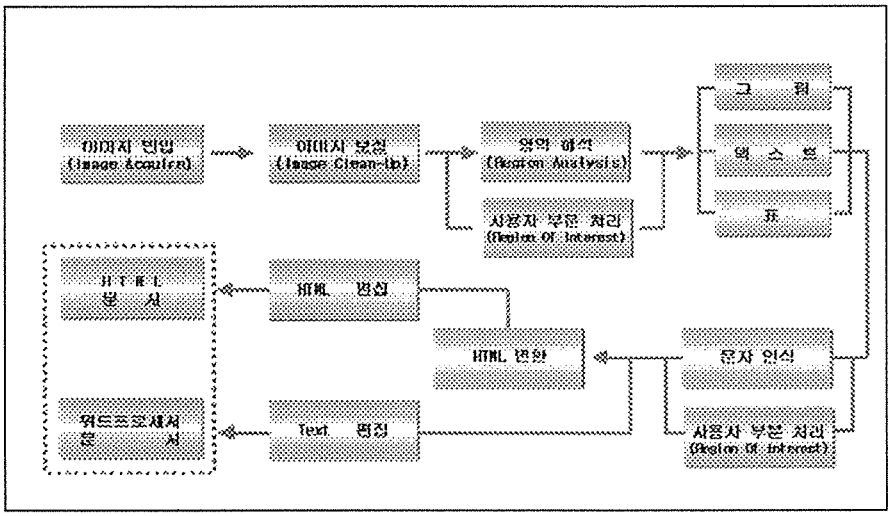

그림 2. *WebPress2000* 의 기능구성 및 처리흐름

2) 기 능

① 문서 스캐닝

활자로 되어 있는 인쇄물을 스캐너를 구동하여 비트맵이미지로 바꿉니다. 스캐너의 자동급지기능을 이용하면 대

량의 문서를 스캐닝할 수 있습니다. 컬러이미지 스캐닝도 지원됩니다.

② 이미지 입력

TIFF, BMP, JPG, PCX 등의 이미지 파일을 *WebPress*2000으로 Loading 합니다.

③ 이미지 Clean-Up

이미지의 상태를 최적화하기 위한 기울기 보정, 노이즈 제거, 흑백반전, 지우기, 이미지 회전, 컬러이미지 속성 조정 등의 기능이 지원됩니다.

④ 영역 해석

Loading 된 문서의 구조를 해석하여 텍스트, 테이블, 그림 등 세 가지의 영역속성으로 구분합니다.

⑤ 문자인식

텍스트영역과 테이블 영역 내의 이미지를 문자인식합니다. 원하는 영역을 지정하여 문자인식을 할 수도 있으며, 한글, 영어, 한자, 특수 기호 외에 독어, 불어, 서반어 등 12 개국어에 대한 문자인식이 가능합니다. 그리고 다양한 폰트가 혼용된 문서도 문자인식합니다.

⑥ HTML 변환

복잡한 구조의 문서 영상을 정확히 해석하고 영역별 좌표 정보를 추출한 후, HTML 규약에 따른 Auto Tagging 을 하여 원본과 동일한 구조의 HTML 문서로 변환합니다.

⑦ Text 편집

문자인식된 텍스트 결과를 편집하거나 교정할 수 있으며, 다른 워드프로세서 파일 포맷(.txt, .doc, .hwp 등)으로도 저장할 수 있습니다.

⑧ HTML 편집

WebPress2000 에서 제공하는 웹에디터를 이용하면 변환되어진 HTML 문서를 사용자의 의도에 맞게 편집하거나 수정할 수 있

습니다.

⑨ **Multi-Page**

대량의 문서를 처리할 때, 각 페이지를 Thumbnail 로 확인하면서 페이지의 이동이나 삭제 등의 기능을 이용하여 페이지 단위의 편집을 할 수 있습니다.

⑩ 일괄작업

대량의 문서를 쉽게 HTML 로 변환하는 과정입니다. 일괄작업이라는 개념을 이용하여 HTML 로 변환할 이미지 파일들을 대량으로 지정하여 연속처리를 하면 사용자가 개입하지 않고도 대량의 문서가 자동으로 HTML 로 변환됩니다.

항상 최종 작업정보를 기억하고 있으므로 예기치 않게 작업이 중단된 경우에도 재실행을 하면 중단된 이후부터 작업이 계속됩니다.

3) 입출력

① 입력

- 활자로 된 인쇄물
- BMP, JPG, PCX, TIFF 등으로 된 이미지 파일

② 출력

- HTML 파일
- TXT 파일(텍스트)
- JPG 파일(그림)
- MS-WORD(.doc)파일
- 아래아한글(.hwp)파일

3. 처리 과정

1) 이미지 입력 및 Image Clean-Up

스캐너에서 스캐닝을 거친 이미지와 이미지파일을 로딩(Loading)하고, 이미지의 자동 기울기 보정, 노이즈 제거, 지우기, 흑백반전, 이미지의 90°, 180°, 270° 회전 등과 같은 이미지프로세싱을 하는 과정이며, 컬러 이미지 속성의 변경도 가능합니다. 지원되는 이미지 파일은 BMP, JPEG, TIFF, PCX 포맷등입니다.

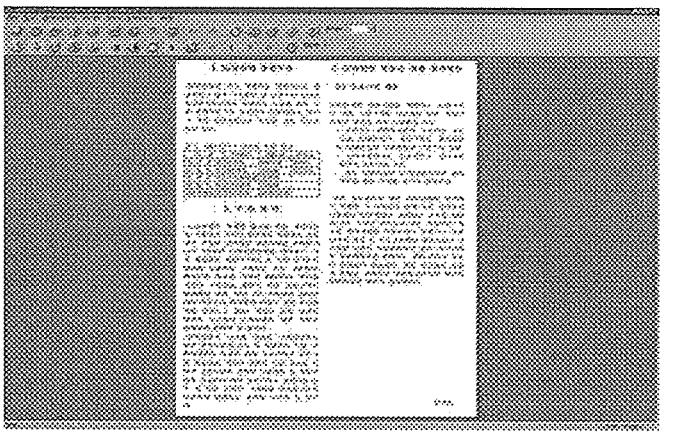

그림 3. 이미지 입력 단계

2) 문서영상해석 (영역해석)

이미지 전처리과정을 거친 메모리 상의 이미지에 대하여 인식영역을 분석하는 단계로서 텍스트(Text),테이블(Table), 그림(Picture)등의 3가지 영역으로 분류됩니다. 분류한 각 영역의 좌표값은 원본과 동일한 구조를 가진 HTML 문서를 만들 때 활용됩니다.

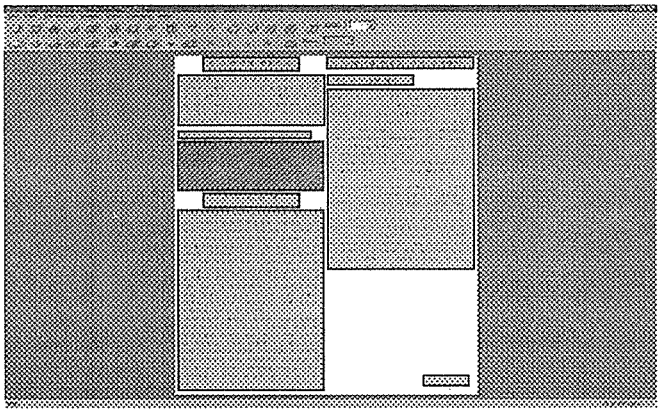

그림 4. 영역해석 단계

3) 문자인식(Optical Character Recognition)

문서영상의 텍스트영역은 OCR을 통해 텍스트로 변환됩니다. OCR은 한글, 한자, 영문 및 특수기호를 지원하며, 서로 다른 문자폰트가 혼합된 문서의 인식도 가능합니다. 인식결과는 내장된 텍스트에디터(Text Editor)에서 맞춤법검사 및 사용자 사전 등의 기능을 이용하여 수정, 편집이 가능합니다.

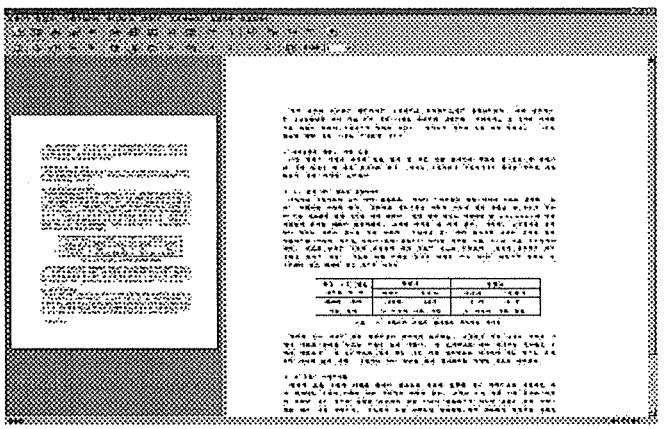

그림 5. 문자인식 단계

4) HTML 변환

원본 이미지의 구조와 동일한 모습의 HTML로 변환하는 과정입니다. 사용자 설정에 의해 전처리 단계(Image Processing) 및 문서영역해석, OCR처리, HTML변환과정을 일괄처리할 수 있으며, 일괄자동처리의 경우 단계별 진행상태를 사용자가 페이지단위로 확인할 수 있습니다.

(A4 Size의 문서를 기준으로 이미지 파일로부터 HTML변환까지의 과정은 약 7초 미만의 시간이 소요됩니다.)

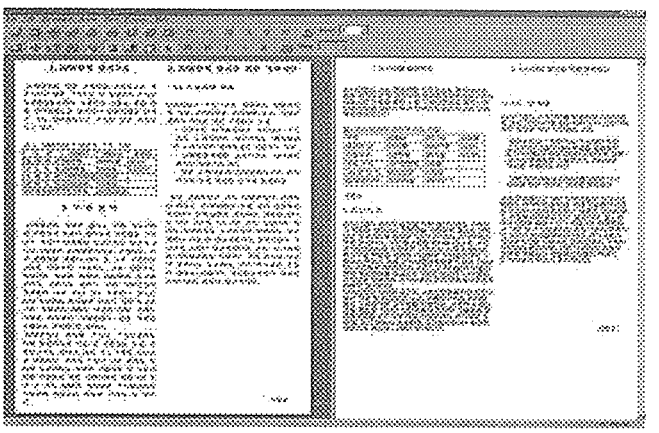

그림 6. HTML 변환 단계

5) HTML 편집

*WebPress*2000은 HTML로 변환된 문서를 사용자의 목적에 맞게 수정,편집할 수 있도록 Web Editor를 내장하고 있습니다. Web Editor를 사용하여 사용자는 HTML문서를 특정 웹 페이지에 링크(Link)시키거나, 필요에 따라 새로운 레이아웃으로 편집하여 DB에 저장할 수 있습니다.

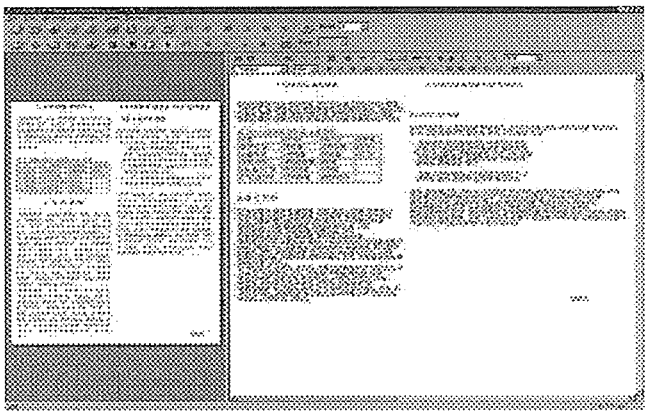

그림 7. HTML 편집 단계

4. 기대 효과

① 인쇄물의 내용을 손쉽게 HTML 문서로 자동변환함으로써 자료의 활용도를 극대화합니다.

- Internet 문서(HTML 문서)의 제작시간을 대폭 단축시킬 수 있습니다.
- 각종 교육기관의 전자 교과서 및 교안을 제작할 때 효율적으로 활용할 수 있습니다.
- 원본과 동일한 구조로 변환함으로써 사용자가 원본자료와 같이 편하고 쉽게 자료를 이해할 수 있도록 합니다.

② 일괄작업 기능을 활용하여, 대량의 문서를 신속하고 용이하게 작업합니다.

- 전자도서관(Digital Library), 전자문서관리시스템(EDMS) 등의 지식관리시스템에서 데이타베이스의 제작툴(Tool)로 활용하면 생산성을 크게 향상시킬 수 있습니다.

③ 인터넷을 통하여 전송되는 자료의 크기를 감소시켜 네트웍 부하를 획기적으로 경감하는 효과를 얻을 수 있습니다.

- 이미지를 텍스트로 변환했을 때 파일크기는 약 10 분의 1 로

감소됩니다.

(현재 전자도서관에서 서비스되는 원문정보는 대부분 이미지 형태로 되어 있슴)

④ 도서관,연구소 등 학술연구기관의 정보서비스를 획기적으로 향상하여 국가정보화에 기여하게 됩니다.

⑤각종 공공기관에서 인터넷을 통한 대민정보서비스에 효과적으로 활용하여 국민의 정보공유 기회를 확대합니다.

⑥HTML, XML 변환에 관한 처리기능을 확보하여 전자문서의 표준화 동향에 능동적으로 대처할 수 있습니다.

⑦*WebPress*2000 에 적용된 요소기술을 Component 로 활용함으로써 전표 및 정형화된 문서의 인식, 필기체 인식, 맹인 낭독시스템, 번역시스템 등 다양한 응용 시스템에 적용할 수 있습니다.

5. 주요 요소기술

① 문서영상 처리 및 이해기술 (자료의 입력 및 전처리)

-Format Conversion 기술 및 다양한 Image Format (PCX, GIF, TIF, BMP) 해석기술

-Raw Data와 관련된 정보의 추출 및 다양한 형식의 영상 Format 생성

-Raw Data Image와 관련 정보로부터 DIB Image 생성 및 DIB Format 처리

-스캐너를 제어하기 위한 범용 드라이버와 비호환부문에 대한 제어

-문서구조를 분석하기 위해 인식된 이미지의 상태를 최적화하는 기술

② 영상 해석 기술

- 효과적인 이미지 추출을 위한 원본 이미지의 사이즈 분할 및

통합

- Text, Image, Table의 영역 분리
- 문서 Layout의 해석 및 영역간 연관정보 추출
- 영역별 좌표와 특성 정보의 내장화

③ 문자인식 기술

- 다중 폰트 인식을 위한 폰트 독립적인 인식기의 설계 및 구현
- 적용 Algorithm : Hybrid Approach (구조적 방법 + 통계적 방법)
- 방대한 양의 문자를 빠른 속도로 인식하기 위한 다단계 분류기 개발
- 인공지능(Artificial Intelligence)에 의한 폰트 학습 기능

④ HTML Generation 기술

- 문서구조 및 내용에 대한 Tag 자동 생성

6. 기 타

1) 특허 출원

① 대용량 인쇄체 문자인식을 위한 특징 추출방법 (국제/국내 특허 접수)

② 다국어문서 인식에서 개별문자 추출방법 및 인식시스템 (97.12.10(97-67558)출원)

③ 다중 폰트 문자인식을 위한 폰트 분류 방법 (97.12.10 (97-67557)출원)

2) 모범적인 産研 협동프로젝트

*WebPress*2000 의 개발은 적극적인 의지로 핵심 요소기술을 (1998 년 11 월 "디지털영상 해석기술" 에 관한 기술이전계약 체결) 이전한 한국전자통신연구원(ETRI)과 이전받은 요소기술을 기반으로 획기적인 S/W 의 개발을 추진한 (주)한국

문서공학의 모범적인 산연협동 프로젝트의 결과입니다.

3) 정부 지원사업의 알찬 성과

*WebPress*2000 의 개발은 1999 년 3 월 중소기업청의 엄정한 기술 평가에 따라 기술혁신개발사업("디지털 영상처리에 의한 복합문서의 HTML 자동변환 기술")의 개발과제로 선정되어 연구개발비 일부를 지원받았습니다. 국가 경쟁력 및 산업 발전에 이바지할 수 있는 제품을 정책적으로 지원하는 정부의 의지에 힘입어, 당사 연구개발진은 고양된 의욕을 가지고 활용성 있는 양질의 제품을 개발하였습니다.