제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

E-Book을 위한 교정 편집기

서현석, 하상호
순천향 대학교 정보기술공학부

e-mail : hyunseok321@hanmail.net, hsh@sch.ac.kr

The Correction Editor for E-Book

Hyun-Seok Seo, Sangho Ha
Dept. of Information Technology, SoonChunHyang University

요 약

e-book 은 책을 종이가 아닌 전자형태로 출판하는 것이다. 이러한 e-book 은 기존의 책들과는 달리 다양한 서비스를 제공할 수 있으며 가격이나 책의 재고관리와 같은 경제적인 부분에서도 소비자나 출판사입장에서 상당한 이득을 기대 할 수 있다. 일반적으로 e-book 은 OCR(문자 인식기)을 사용하여 생성되는데 OCR 의 문자인식률은 일반적으로 95%정도이며 인식된 텍스트 문서에 대한 교정이 필요하다. 그러나 기존의 워드 프로세서들은 간단한 수준의 교정기능만이 제공되고 있다. 본 논문에서는 교정을 보다 효율적으로 수행할 수 있는 교정 편집기를 설계하고,구현한다.

1. 서론

E-book 은 컴퓨터, PDA 와 같은 모바일 제품 그리고 인터넷의 확산으로 시간과 공간의 제약 없이 다양한 서비스를 제공할 수 있게 되었다. 또한 e-book 은 디지털시대 출판계에 닥친 가장 큰 변화이다. 미국 맥그로-힐사는 고객이 원하는 대로 현장에서 인쇄해 판매하는 "주문형서적(book-on-demand)"서비스를 도입하였다. 맥그로-힐의 자회사인 프리 미스사는 이스트 코닥사와 RR 도넬드사의 기술을 도입, 학생들을 대상으로 원하는 교재를 현장에서 인쇄해 판매하는 새로운 개념의 서적을 운영하고 있다. 이 서비스는 분당 90 매를 인쇄할 수 있는 고속의 프린터를 사용하고 있어 교재 한 권을 고객의 손에 넘기기까지 소요되는 시간은 몇 분에 불과하다. 같은 교수에게 수강하더라도 강의내용이 조금씩 변경될 경우 필요한 부분만을 다른 교재에서 발췌한 서적을 판매함으로써 불필요한 교재의 구입에 따르는 비용을 줄일 수 있어 학생들로부터 인기를 모으고 있다. 그리고 제록스사도 주문형서적시장에 진출을 서두르고 있다.

한편 국내에서도 다소 늦은 감이 있지만 전자출판으로 진출하는 회사들이 늘고 있으며 이에 대한 연구가 이루어 지고있다. 북스 웹폭스[6]은 우편이나 택배 대신 온라인으로 책을 읽을 수 있는 전자서점으로 정가의 절반 가격이하로 다운로드 받거나 html 형식으로 인터넷에서 볼 수 있다. 전자책 서비스 업체인 북토피아[8]는 전자책 서비스 사이트 e-book 을 통해 전자책 서비스를 하고 있다. 또한 북토피아는 북몰 사이트를 통해 종이책, 전자책, 두가지 형식의 서적을 동시에 공급하고 있다. 맞춤형 웹메일 서비스인 마이북토피아, 독자 및 출판사들에게 다양한 정보를 제공하는 플라자등을 순차적으로 종합적인 북 포털 사이트로 개편해나가고 있다. 그리고 초록배카툰즈가 바로북[7]이라는 이름으로 인터넷을 통한 전자책 서비스를 제공하고 있는 것을 비롯해 YES24, 리얼북[9], 온북[10]과 같은 온라인 전자책 서비스가 제공되고 있다. 이 같은 상황에서, 국립중앙도서관 대강당에서 출판계 및 도서관 관계자 400 여명이 참석한 가운데 전자책과 관련한 세미나가 이루어지기도 했다.

일반적으로 출판계에 큰 변화를 일으키고 있는 전자책은 다음과 같은 과정으로 제작된다. 종이책을 고

속 스캐너(Scanner)를 통해 스캔(Scan)하여 책에 대한 그래픽 파일을 생성후 OCR(Optical Character Reader: 문자 인식기)를 통해서 책을 텍스트 파일로 저장한다. 그러나 OCR 을 통해 만들어진 텍스트는 일 반적으로 95%의 정확도를 갖으며 텍스트의 5%는 정확하게 인식되지 않는 문제점이 있다. 따라서 정확하게 인식되지 않은 5%의 텍스트를 교정하는 것이 필요하다. 그러나 MS-WORD 와 아래 한글을 포함하여 기존의 워드 프로세서는 문서 편집이 목적이기 때문에 교정 기능이 충분하지 않다. 찾기, 바꾸기와 같은 단순한 수준의 교정기능만이 제공될 뿐이다. 그러므로 기존의 워드프로세서를 사용한 교정작업은 효율적이지 못하여 상당한 시간이 소요된다.

본 논문에서는 OCR 로 인식된 텍스트 문서에서의 교정작업을 빠르고 효과적으로 수행할 수 있는 교정 편집기를 설계하고 구현한다. 교정 편집기는 기본교정, 링크교정, 대/소문자 변환과 같은 기본적인 교정기능을 효과적으로 제공하며, 교정한 문자열은 마킹기능을 통하여 교정 가능성 있는 문자열과 이미 교정한 문자열을 표시하고, 교정 가능한 문자열은 효율적으로 교정할 수 있게 하고, 이미 교정한 문자열을 나중에 검토할 수 있게 하는 기능을 제공한다.

2. 교정 편집기

2.1 교정 편집기 구조

그림 1 은 교정 편집기의 전체 구조를 보여준다. 교정 편집기의 기반이 되는 문서편집 모듈은 MFC(Microsoft Foundation Class)[1]에서 제공하는 CrichEditCtrl 클래스에 기반하여 구성되어있으며, 워드패드 수준의 편집환경을 제공한다. 그리고 마킹 모듈은, 문서에서 교정기능을 사용하여 교정된 문자열과 동일한 모든 문자열은 파란색으로, 교정된 문자열은 붉은색으로 각각 표시한다.

교정기능은 크게 단일 교정과 그룹 교정으로 나눈다. 단일 교정은 한 문자열에 대하여 한번 교정하는 기능이며, 그룹 교정은 한 문자열 또는 여러 문자열에 대하여 반복적으로 교정할 수 있는 기능이다. 교정 기능을 통하여 교정을 하면 교정한 문자열에 대해 마킹을 수행하며, 단일 교정은 마킹 후 문서편집 모듈로 이동하여 문서를 편집할 수 있도록 하는 반면, 그룹교정은 해당 교정기능의 종료 전까지 다른 문자열들에 대해 교정과 마킹을 반복적으로 수행한다.

<그림 1> 교정 편집기 구조

2.2 교정 편집기 기능

교정 편집기는 기존의 워드 프로세서와 다르게 모든 교정 기능에 대해 팝업 메뉴를 제공하며, 기존의 워드프로세서에서 제공하지 않는 교정기능인 링크 교정, 파란색 검토, 붉은색 검토, 재 교정 등을 제공한다. 다음은 교정 편집기에서 제공하고 있는 단일 교정과 그룹 교정의 각 기능에 대해서 설명한다.

기본 교정
문서에서 선택된 문자열을 교정한 후 마킹을 한다. 또한 교정을 할 때 이전에 사용하였던 교정 문자열 리스트를 제공한다.

링크 교정
문서에서 선택된 문자열과 같은 문서의 모든 문자열을 링크 문자열 리스트로 만들어 링크 문자열 리스트에서 아무 문자열이나 선택하여 교정할 수 있으며 교정 후 마킹을 한다. 또한 교정을 할 때 링크 문자열이 전에 사용하였던 교정 문자열 리스트를 제공한다.

찾아 바꾸기
문서에서 문자열이나 문자열 패턴을 찾을 수 있도록 하며, 찾아진 문자열을 교정을 할 수 있다. 교정이 이루어지면 마킹을 한다. 또한 교정을 할 때 찾아진 문자열이 이전에 사용하였던 교정 문자열 리스트를 제공한다.

대문자 변환
선택된 문자열에 대하여 소문자 알파벳을 대문자로 변환하며 변환된 알파벳을 포함한 영어 단어의 교정 전 문자열과 같은 문서의 모든 문자열은 파란색으로 교정 후 대문자로 된 문자열은 붉은색으로 마킹한다.

소문자 변환
선택된 문자열에 대하여 대문자 알파벳을 소문자로 변환하며 변환된 알파벳을 포함한 영어 단어의 교정 전 문자열과 같은 문서의 모든 문자열은 파란색으로 교정 후 소문자로 된 문자열은 붉은색으로 마킹한다.

파란색 검토

선택한 파란색 문자열이 있으면 부분 검토가 수행되며, 그렇지 않은 경우에는 문서의 모든 파란색 문자열에 대해 검토를 수행한다. 검토를 수행하면서 교정을 할 수 있으며 교정이 이루어지면 마킹을 한다. 또한 교정을 할 때 검토되고 있는 파란색 문자열이 이전에 사용하였던 교정 문자열 리스트를 제공한다.

붉은색 검토

특정 붉은색 문자열을 선택하여 선택한 문자열에 대한 검토를 하거나 문서의 모든 붉은색 문자열에 대한 검토를 수행할 수 있다. 검토를 수행하면서 교정을 할 수 있으며 교정이 이루어지면 마킹을 한다. 또한 교정을 할 때 검토되고 있는 붉은색 문자열이 교정 전의 문자열 리스트를 제공하여 교정 전 문자열로 복구할 수 있다.

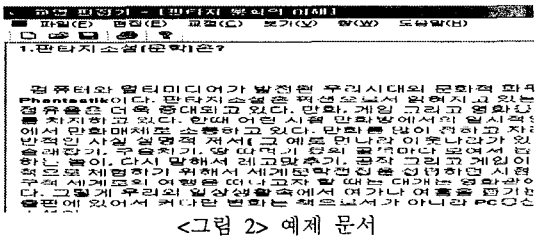
재 교정

선택된 문자열이 붉은색으로 마킹된 문자열에 대해서 교정 바로 전의 문자열을 보여준 후 다시 교정을 할 수 있게 한다. 재 교정이 이루어진 문자열은 다시 붉은색으로 마킹을 한다. 붉은색 검토를 간단하게 만든 기능으로 검토작업을 빠르게 할 수 있게 한다.

3. 구현 및 적용

교정 편집기는 Windows XP 환경에서 Visual C++를 사용하였다. 문서 편집 창은 MFC 의 CRichEditCtrl 클래스를 사용하여 구현하였으며, 워드 패드 수준의 편집환경을 제공한다. 그리고 마킹과 교정 기능들은 각각의 기능 단위 클래스로 구현하였다.

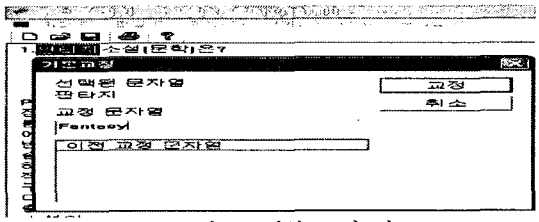
다음은 교정 편집기의 핵심 기능인 기본 교정, 링크 교정, 마킹, 파란색 검토, 붉은색 검토기능을 예제 문서 적용한다. 그림 2 는 교정 편집하기 위해 불러온 예제 문서이다.



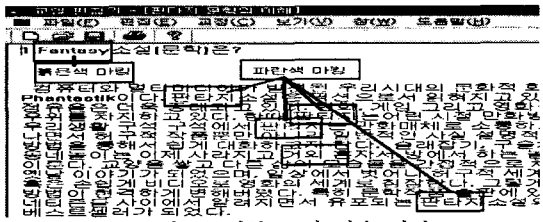
<그림 2> 예제 문서

그림 3 은 기본 교정을 불러온 문서에 적용하는 예이다. 기본 교정의 선택된 문자열 항목은 현재 문서에서 교정하려는 문자열을 나타내며, 교정 문자열은 선택된 문자열을 바꾸려는 문자열을 입력 받는다. 그리고 이전 교정 문자열은 교정하려는 문자열에 대하여 이전에 교정한 문자열 리스트로, 교정 문자열을 입력하는 대신 이전 교정 문자열 항목에서 선택할 수 있다. 1 줄의 "판타지"문자열이 선택되었으며 이를 "Fantasy"로 교정한다.

그림 4 는 기본 교정을 적용한 결과다. 교정한 결과 "판타지"는 "Fantasy"로 교정되었고 교정한 "Fantasy" 붉은색으로, 문서의 모든 "판타지"는 파란색으로 마킹하였다.



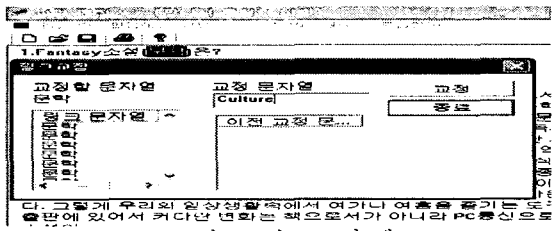
<그림 3> 기본 교정 예



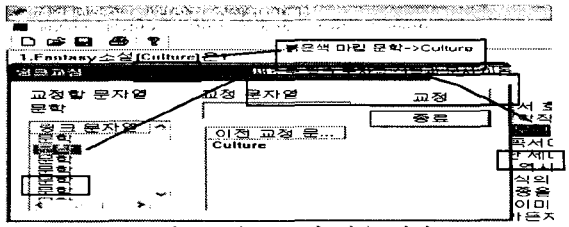
<그림 4> 기본 교정 적용 결과

그림 5 는 링크교정을 그림 4 의 결과 문서에 적용한 예로 1 줄의 "문학"을 선택하여 "Culture"로 교정한다. 링크교정의 교정할 문자열 항목은 현재 교정하려는 문자열을 나타내며, 링크 문자열 항목은 교정하려는 문자열과 동일한 모든 문자열을 그 위치와 함께 나타내며, 선택을 하면 선택된 문자열의 위치로 문서가 이동한다. 그리고 기본 교정에서와 같이 교정 문자열과 이전 교정 문자열 항목을 제공한다.

그림 6 은 링크 교정을 한 결과 "문학"은 "Culture"로 교정되었으며 교정한 문자열인 "Culture"는 붉은색으로, 문서의 모든 "문학"은 파란색으로 마킹하며, 이전 교정 문자열에 "Culture"가 추가되었다. 그리고 링크교정은 종료할 때까지 교정을 한 후에도 다른 링크 문자열을 선택하여 교정하고 마킹을 한다.



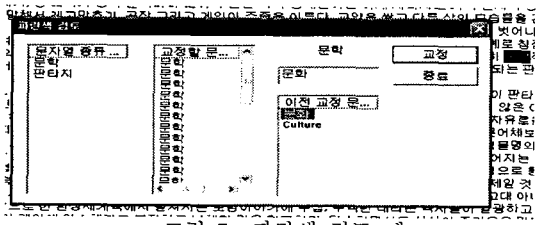
<그림 5> 링크 교정 예



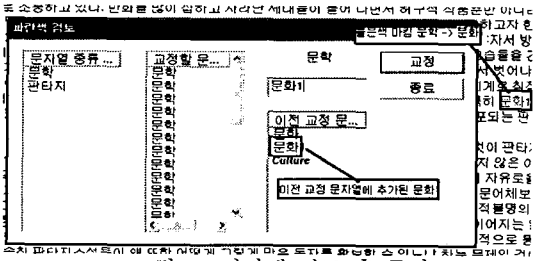
<그림 6> 링크 교정 적용 결과

그림 7 은 그림 6 에서의 결과 문서에 대해 파란색 검토를 하는 예이다. 문자열 종류 "문학"을 선택하여 교정 문자열은 특정한 위치의 "문학"을 선택하여 "문화 1"로 교정 한다. 파란색 검토의 문자열 종류 항목은 파란색으로 표시된 모든 문자열의 종류를 나타내며, 선택을 할 수 있다. 문자열 종류가 선택되면 선택된 문자열 종류에 대한 모든 문자열을 나타내는 교정할 문자열에서 특정한 위치의 문자열을 선택할 수 있으며 선택된 문자열은 교정 문자열을 입력 받아 교정할 수 있다. 그리고 기본 교정과 같이 교정 문자열과 이전 교정 문자열 항목을 제공한다.

그림 8 은 파란색 검토를 적용한 결과로, 교정된 "문화 1"가 붉은색으로 마킹이 되었으며, 이전 교정 문자열에 "문화 1"이 추가 되었다. 그리고 파란색 검토를 종료할 때까지 다른 파란색 문자열들을 검토할 수 있다.



<그림 7> 파란색 검토 예

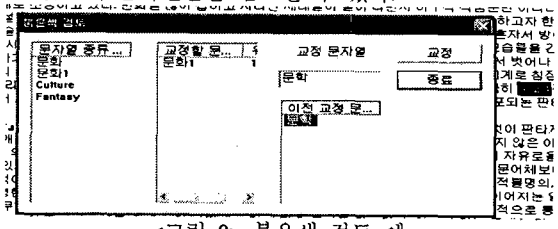


<그림 8> 파란색 검토 후 문서

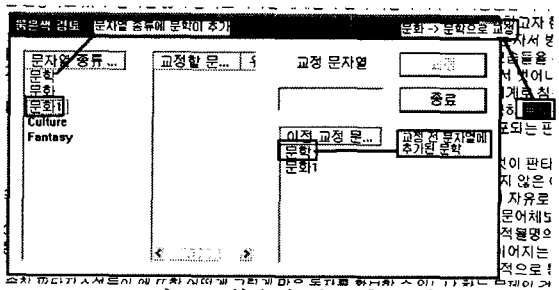
그림 9 는 그림 8 에서의 결과 문서에 대해 붉은색 검토를 하는 예다. 문자열 종류로 "문학 1"을, 교정할 문자열로 특정한 위치의 "문화 1"을 교정 문자열로는 교정 전 문자열에서 "문학"을 선택하여 특정위치의 "문화 1"을 "문학"으로 교정 한다.

붉은색 검토에서도 파란색 검토와 같이 문자열 종류 항목을 제공하지만, 붉은색으로 표시된 모든 문자열의 종류를 나타낸다는 것이 다르다. 그리고 교정할 문자열항목은 문자열 종류 항목이 선택되었을 때 선택된 문자열 종류에 대한 모든 문자열을 나타내며, 그 위치를 함께 보여준다. 교정할 문자열을 선택하면 선택된 문자열이 있는 위치로 문서가 이동한다. 그리고 교정 전 문자열 항목은 교정하려는 문자열에 대하여 교정 전의 문자열 리스트를 나타내며, 교정할 때 교정 문자열을 입력 받는 대신 선택할 수 있다.

그림 10 은 붉은색 검토 적용 결과로 교정된 "문학"은
붉은색으로 마킹 되었으며, 붉은색 문자열의 종류로 "문학"이 추가 되었으며, 교정 전 문자열에 "문화 1"이 추가되었다. 그리고 붉은색 검토를 종료할 때까지 다른 붉은색 문자열들을 검토할 수 있다.



<그림 9> 붉은색 검토 예



<그림 10> 붉은색 검토 후 문서

4. 결론

본 논문에서는 E-Book 생성과정 중에 95%의 인식률을 갖는 OCR 로 인식된 텍스트를 빠르고 효율적으로 교정할 수 있는 교정 편집기를 설계하고 구현하였다. 논문의 교정 편집기는 기존의 워드 프로세서가 제공하지 않는 링크 교정, 파란색 검토, 빨간색 검토 교정 기능들을 제공하고 있으며, 이러한 교정 기능을 사용하여 교정을 빠르고 효과적으로 수행할 수 있다. 또한 논문의 교정기능을 기존의 워드 프로세서에 포함 시켜 교정 기능을 한층 더 강화 시킬 수 있다고 본다. 향후 연구는 기존의 워드 프로세서에서 제공하고 있는 맞춤법 검사와 한자한글변환, 일어변환과 같은 다른 나라의 문자로의 변환 기능을 포함시킬 계획이다.

참고문헌

[1] 김용성, "Visual C++ 6 완벽 가이드"
[2] Charles Petzold Paul Yao, "Programming Windows95"
[3] http://www.codeguru.com
[4] http://www.tipsoft.com
[5] http://www.devpia.com
[6] http://books.webfox.co.kr
[7] http://www.barobook.co.kr
[8] http://www.booktopia.com
[9] http://www.realbook.co.kr
[10] http://www.onbook.co.kr