제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

HTML 문서 생성기의 설계 및 구현

최지연*, 민수홍, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과
e-mail : {gratel, shmin, dscho}@ewha.ac.kr

Design and Implementation of HTML Document Generator using Dictionary based Pre-processor

Ji-Yeon Choi*, Su-Hong Min, Dong-Sub Cho
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

E-Mail 서비스는 WWW 시스템의 가장 기본적인 기능으로, 인터넷 기술이 발전하고 사용자가 기하급수적으로 증가함에 비례하여 e-Mail 사용자도 늘고있다. 그러나 기존의 e-Mail 은 HTML 의 텍스트 기반 구조를 통한 비동기적인 형태의 서비스를 계속 유지하고 있어, 이에 대해 좀더 동기적인 표현방법이 요구되고 있다. 따라서 본 논문에서는 동기적인 E-Mail 서비스에 초점을 맞추어, 단순히 정보를 제공받는 서비스가 아니라 사용자 위주로서의 E-Mail 서비스를 제안하고자 한다. 메일 내용에 대해 자동으로 필터링하여 단어마다의 색깔 지정과 하이퍼링크의 자동 생성으로 e-Mail 을 재편집할 수 있는 편리함을 제공하며, e-Mail 수신자가 원하는 정보를 쉽게 효율적으로 접근할 수 있도록 한다.

1. 서론

WWW 이란 거대한 인터넷상에 흩어져 있는 정보들을 효율적으로 표현하고 접근할 수 있도록 설계된 분산 하이퍼미디어 시스템이다[1]. 이때 제공되는 정보는 하이퍼텍스트(hypertext) 형태로 표현된다. 하이퍼텍스트란 일반적인 텍스트 데이터와 멀티미디어 데이터 그리고 다른 하이퍼텍스트에 대한 포인터로 구성되는 문서를 말한다[2][3][4]. HTML 은 1990 년 이후 웹 상에서 가장 많이 사용되는 문서 교환의 표준 형식으로서 1986 년에 제정된(ISO 8879) SGML(Standard Generalized Markup Language)를 바탕으로 정의된 하나의 응용이다. HTML 은 구조적 특성을 갖는 이유로 손쉬운 편집이 어려웠으나 웹이 전 세계 이용자들의 각광을 받으면서 HTML 을 이용한 문서편집이 활발해지고 빠르고 간편한 WYSIWYG(What You See Is What You Get) 방식의 HTML 편집기도 나오게 되었다.

그러나 HTML 은 텍스트 기반의 구조를 통한 비동기적인 형태의 서비스 유형으로 문서검색과 단순한 표현 방식을 제공한다. 따라서 다양한 형태의 동기적인 데이터 및 네트워크 자원을 효과적으로 교환 및 검색하기에는 한계를 가지고 있다[5].

특히 WWW 의 대표적인 기능이라 할 수 있는 E-Mail 서비스는 대부분의 인터넷 사용자들이 사용하는 기본적인 기능임에도 불구하고 단순한 텍스트 기반의 표현 방식을 사용하고 있는 실정이다.

E-Mail 사용자는 단순 텍스트로 짜여진 e-Mail 을 받고 메일의 내용을 일일이 읽어서 직접 확인해야 한다. 또한 메일의 내용에서 사용자가 어떤 단어에 대한 정보를 얻기 위해서는 별도의 창을 열어 검색 엔진을 사용하여 검색하고자 하는 단어를 입력한 후에 정보를 얻어야만 한다.

이러한 e-Mail 사용은 메일 수신자에게 효과적인 정보제공 측면에서는 많은 문제점을 갖고 있다. 이미 알고 있거나 기본적으로 제공될 수 있는 정보들을 사전식으로 등록하여 자동 검색으로 이 문제를 해결하는 것이 좋다.

본 논문에서는 이러한 불편함을 줄이기 위하여 사전기반 전처리기를 이용한 HTML 문서 생성기의 설계 및 구현에 대한 제안을 하고자 한다.

이 논문은 2003 년도 두뇌한국 21 사업에 의하여 지원되었음.

본 논문은 2 장에서는 메일 서비스에 대한 기존의 연구에 대해 언급한다. 3 장에서는 제안하는 HTML 문서 생성기의 구현과 실행에 대해 설명을 하고, 4 장에서는 본 논문의 결론을 맺고, 향후 연구 계획에 대해 언급한다.

2. 관련연구

2.1 멀티미디어 메일 서비스 관련 기술(MIME)

전자메일은 인터네트를 포함해 거의 모든 컴퓨터 네트워크상에서 가장 폭 넓게 사용되는 서비스로 전자메일의 인터넷 표준인 RFC 822(Crocker)은 현재 인터넷 뿐만 아니라 전세계적으로 가장 널리 활용되고 있다. 그러나 전자메일을 통해 주고받을 수 있는 정보형태는 7bit 의 ASCII 데이터만을 기본으로 하고 있어 오디오, 비디오등과 같은 멀티미디어 데이터를 주고 받을 수 있기를 원하는 메일 사용자의 요구를 충족시키지 못하고 있다 [6].

MIME(Multipurpose Internet Mail Extensions)은 이러한 멀티미디어 메일 전송의 필요성에 따라 부각된 인터넷 표준으로 1992 년 6 월 IETF 에 RFC 로 제안된 이래 급속히 발전하여 지난 93 년 9 월 RFC 1521 과 1522 라는 두개의 draft 표준으로 IETF 에 의해 승인되었다. MIME 은 RFC 822 의 확장표준으로 multipart / multimedia 메시지를 정의하는 표준화된 방식을 제공하고 있으며 현재의 RFC 822 과의 완벽한 호환을 유지하고 있다. 이를 위해 MIME 표준은 현재의 RFC 822 스타일의 헤더를 이용하여 서로 다른 메시지 타입과 멀티미디어 바디부들을 정의하고 있다

MIME 을 지원하는 메일 시스템의 핵심은 MIME UA 라 할 수 있다. MIME UA 는 복잡하게 구성된 multipart 메시지를 처리할 수 있어야 한다. MIME UA 는 MIME parser 와 MIME dispatcher 를 갖는다. Parser 는 수신된 메시지의 헤더 정보로부터 메시지의 구성 요소들을 해석하는 기능을 가지며, dispatcher 는 해석된 각 구성요소별로 디스플레이를 위한 특정의 viewer 들을 호출한다. MIME 형태의 메일 메시지를 작성하고자 할 경우에 MIME UA 는 메일 전송 시스템 과의 통신을 담당하는 MIME Message Builder 와 MIME 메시지 구성 성분들을 하나로 모으는 Message Designer 를 이용한다.

Message Designer 는 확장성을 고려하여 설계되어 있으며, 따라서 메일 이용자가 원하는 모든 형태의 작업을 지원하는 composition agent 의 추가가 가능하다. Interactive Multimedia E-Mail 은 클라이언트/서버환경의 메일 시스템에서 메일 이용자가 서버에 존재하는 특정 메일 메시지 중 자신이 보고 싶은 부분만을 선택하여 볼 수 있도록 하는 것으로 이는 MIME 메시지의 전체구성이 담겨있는 헤더정보만을 클라이언트측에 우선적으로 송신하고 실제의 메시지 내용에 대해서는 사용자의 요구에 따라 전송함으로써 가능해진다[7].

2.2 Sendmail 의 동작환경

- MTA(Mail Transfer Agent) : 인터넷상에 있는 하나의 컴퓨터로부터 다른 컴퓨터(메일서버)로 전자 메일을 전송하는 서버 프로그램.
- 인터넷 메일 시스템은 동격 서버들을 가진 분산 클라이언트/서버 시스템이다. 클라이언트는 서버와 통신하여 메일을 송수신하고, 서버들이 서로 통신한다. 클라이언트가 나가는 메시지를 직접 서버(MTA)로 보내면, 그 서버는 메시지를 수신자의 우편함으로 배달하거나 혹은 그것을 전달(forwarding)할 다른 MTA 로 보낸다. 이러한 시스템은 서버들을 계층적으로 배열함으로써 높은 확장성을 갖도록 하기 위한 것이다.
- 인터넷 핵심에서 대부분의 메일을 처리하는 가장 대중적인 무료 MTA 인 sendmail 은 원래는 단순한 텍스트메시지를 위한 메시지 주소 번역 기능을 제공하면서, 새롭게 연결된 네트워크에서 사용중인 다수의 메일 시스템을 통합하기 위해 작성되었다.

2.3 디렉토리 시스템의 일반적 구성

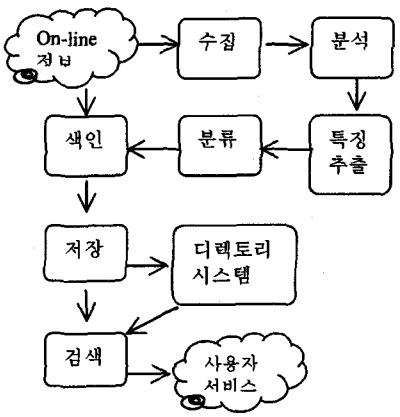

그림 1. 디렉토리 시스템의 구성

2.4 e-Mail 전송 프로토콜

SMTP	• TCP/IP 계열의 표준화된 통신 규약 • 메일 서버들 사이의 전송 • 메일 서버와 메일 클라이언트 사이의 전송 • (UNIX 환경) • 메일을 사용자의 시스템에 저장하지 않는다. • 메일 저장소에 따로 저장. • 사용자의 요구시 가져온다.
POP3	• 메일 서버와 메일 클라이언트 사이의 전송 • (pc 환경) • 전송될 메일을 사용자의 시스템에 저장. • 사용자의 메일 조작이 용이
IMAP	• 서버에 기반을 둔 메일 저장 • 메일의 머리말만을 복사 • POP3 의 기능을 확장하여 대체할 목적으로 개발

표 1. e-Mail 전송 프로토콜

3. 사전기반 HTML 문서 생성기의 설계 및 구현

기존의 메일 서비스는 사용자의 관점을 고려하지 않은 단순히 전달 목적만을 가진 텍스트기반 메일 서비스이다. 사용자는 자신이 필요한 내용인지 아닌지를 알기 위해서 메일의 내용을 일일이 읽어서 판단을 해야하고, 메일 내용 안의 어떤 단어에 대한 정보를 얻기 위해서는 Back 버튼을 이용하거나 새 창을 띄워서 정보를 얻어야만 하기 때문에 기존의 메일 서비스는 비효율적이라고 할 수 있다. 따라서 사용자의 이중적인 작업을 피하고 사용자로 하여금 원하는 정보를 한눈에 알아보고 그에 대한 정보를 한번에 얻을 수 있도록 하는 좀더 동적인 메일 서비스가 필요하다.

보다 편리한 메일 서비스를 위해 사용자는 강조하고 싶은 단어와 단어정보를 그룹으로 묶어 사전에 추가한다. 메일이 도착하면 사전을 통해 메일은 자동으로 필터링되며, 이때 사용자가 사전에 미리 등록해 두었던 단어들은 그룹별로 색이 다르게 변하며, 단어에 대한 정보가 링크로 표시된다. 이는 텍스트 위주의 메일 서비스를 사용자의 필요에 따라 동적으로 작용하게 하여 사용자에게 편리함을 주고 이중적인 시간과 부담을 줄일 수 있다.

3.1 TXT2HTML 전처리기의 구성

보내고자 하는 e-Mail 의 내용을 HTML 로 변환해주는 전처리기는 사용목적에 따라 두 가지로 나눌 수 있다. Sendmail 과정에서 메일 본문을 편집한 후 전송 전에 자동으로 HTML 문서를 생성하는 경우와 수신자가 POP3 서버에서 메일을 가져올 때 본인 결정에 따라 HTML 문서로 변환하여 내용을 보는 경우가 있다. 각 경우에 따른 처리 동작 과정은 그림과 같다.

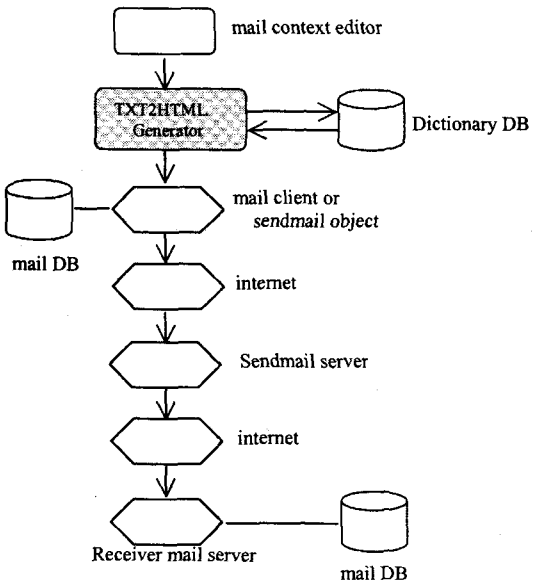

그림 2. SMTP 용 TXT2HTML Generator

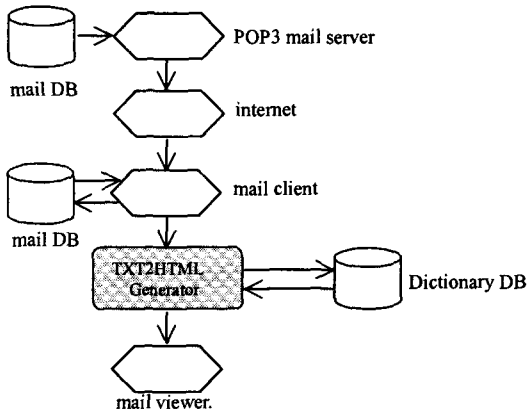

그림 3. POP3 용 TXT2HTML generator

3.2 Dictionary DB 의 구성

Dictionary DB 는 메일의 전송자, 수신자 모두가 개인적 용도로 사용할 수 있는 구조를 가져야 한다. 일반적 공통항목은 표준화 시켜놓고 개인적 정보 부분은 별도 추가, 삭제, 수정이 가능하도록 하였다. 그림 4 는 Dictionary DB 가 갖는 요소들이다.

[공통요소]	[개인적요소]
사전식 단어 검색	주소록 검색(URL, e-Mail)
시사용어 검색	즐겨찾기 검색
주요 공공기관 URL 검색	개인 소속기관 정보검색
주요 인물 e-Mail 검색	이미지 정보 검색
세계 각국 주요 URL 검색	음성, 비디오정보검색
주요 상품 URL 검색	웹 정보 검색

그림 4. Dictionary DB 검색 요소

3.3 전체 구조

웹 문서에서 가장 널리 사용되는 HTML 은 HYPER TEXT 형식의 문서로서 일반적인 문서에 비해 사용자에게 보다 더 많은 표현의 영역을 제공한다. 이를 통해 검색자가 원하는 내용을 보다 더 효율적이고 알아보기 쉽게 표현할 수가 있다. 예를 들어 단어에 색이 들어간 경우 사람들은 그 단어를 다른 단어들보다 더 강하게 인식하게 된다. 이러한 기능은 단어들을 색으로 표현함으로써 단어의 중요도가 차별화 될 수 있도록 한다.

전체적인 구성을 살펴보면, 먼저 각 단어들을 그룹별로 나누고 원래의 문서를 수정하는 전처리 과정과 색을 변경하고 하이퍼링크를 생성하는 과정으로 나누어진다.

전처리 과정에서는 메일 내용을 읽어들여 각 단어가 어떤 단어에 속하는지를 파악한다. 파악 후에 어떤 단어가 어떤 그룹에 속한다고 판별되었을 때, HTML 문서를 읽어들여 각 단어별 그룹에 맞게 색이 변경되고 링크가 걸리도록 설계하였다.

본 논문에서 제안하는 사전기반 HTML 문서 생성과정은 다음과 같다.

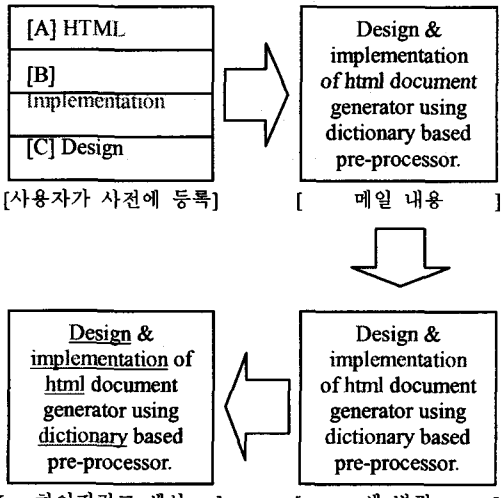

그림 5. 사전기반 HTML 문서 생성 과정

따라서 사용자가 미리 단어를 사전에 추가하여 두고 메일이 도착할 때마다 자동으로 HTML 문서가 변경되도록 한다면 사용자가 메일 내용을 읽을 때도 편하고, 상대적으로 적은 시간과 노력이 필요할 것이다.

3.4 구현

```
// colourize the keywords ----------------------------------------
 void Engine::colourKeys(int index, string key, string
cssclass) {
      if(abortColour(index)) {
      return;
      }
    buffer.insert(index, "<a href=\"http://www.daum.net\">
    <font CLASS=" + cssclass + ">");
    buffer.insert(index+key.size()+50, "</font></a>");
 }
```

그림 6. Colorize & hyperlink the keywords

```
#include "langtext.h"

LangText::LangText() {
        init_switches();
        doStrings    = No;
        doNumbers    = No;
        doKeywords   = No;
//      doCaseKeys   = No;
        doRemComnt   = No;
}
void LangText::fill() {
        // no keywords
}
```

그림 7. 변환되기 이전의 문서

```
#include "langtext.h"

LangText::LangText() {
        init_switches();
        doStrings    = No;
        doNumbers    = No;
        doKeywords   = No;
//      doCaseKeys   = No;
        doRemComnt   = No;
}
void LangText::fill() {
        // no keywords
}
```

그림 8. 변환된 이후의 문서

4. 결론 및 향후 연구과제

본 논문에서 제안하는 사전기반 전처리기를 이용한 HTML 문서 생성기는 모든 메일 내용을 읽어들여 동일한 내용의 다른 웹 문서로 변환시키는 기능을 함으로써 각 사용자들의 환경설정(사전에 단어추가)에 따라서 메일의 중요도가 측정될 수 있다.

향후 연구 과제는 메일에서 사전을 기반으로 HTML 문서를 재생산하여 색 지정 기능과 하이퍼 텍스트 기능을 제공하는 데에만 그치지 않고, 더 나아가 사전에 등록된 정보를 바탕으로 소리, 동영상 및 가상현실 등의 멀티미디어 개체들을 표현할 수 있도록 하이퍼 미디어 기능을 지원하는 메일 시스템을 구축하는 데에 있다.

참고문헌

[1] w3 team at CERN, The WWW Book, 1995.

[2] Michael J. Hanah, General HTML Syntax http://WWW.sandia.gov/sci_computer/html_ref.html

[3] T.Berners-Lee, HyperTEext Markup Language. Rfc866 1995.

[4] J.K. Cohen, Elements of HTML Style http://bookweb.csis.uci.edu/staff./StyleGuide.html

[5] http://www.javaworld.com/javaworld/jw-12-1996/jw-12-jack.html

[6] 이봉환, 박문호, 이하욱, 주기호, 이찬도, 이남준, 심영진, " IMAP 프로토콜을 이용한 멀티미디어 메일 시스템 " , 한국정보처리학회 논문지 제 4 권 제 5 호 1997. 5.

[7] 멀티미디어 전자메일 MIME http://physics.hallym.ac.kr/course/96/st96-2/hw/ans/ans8/yhshin.html