2003년도 대한전기학회 하계학술대회 논문집 2003.7.21-23

E-Mail 시스템의 첨부파일 형식별 자동분류 에이전트 설계

현영순*, 정옥란, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과

Agent for File Format based Classification of the Attached File in E-Mail System

Young-Soon Hyun*, Ok-Ran Jeong, Dong-Sub Cho
Dept. of Computer Science & Engineering, Ewha Womans University

Abstract - E-Mail 은 현재 가장 많이 쓰이고 있는 인터넷 서비스로서 최근 인터넷 사용자의 증가와 함께 그 사용자 또한 늘어나고 있다. 기존의 메일 기반 에이전트는 서버에 무분별하게 메일을 저장하는 방식이기 때문에, 첨부파일에 접근하기 위해서는 관리자가 수신된 메일을 일일이 확인해야 하는 번거로움을 가지고 있다. 대량의 메일을 수신하는 기업에서는 메일 뿐만 아니라 첨부되어오는 파일들에 대한 처리부담이 더욱 크다. 본 논문에서는 이를 보완하기 위해 도착한 메일의 내용을 분석하여 키워드를 추출하고, 폴더를 생성하여 카테고리별로 첨부된 파일을 분류해주는 자동분류 에이전트를 제안하고자 한다. 카테고리 별로 분류된 파일은 다시 형식별로 분류되도록 설계하였다. 이는 관리자의 업무부담을 줄이고, 첨부파일을 효과적으로 관리할 수 있는 장점이 있다.

1. 서 론

정보통신 기술의 발전으로 급속히 보급되고 있는 것이 인터넷이라고 할 수 있다. 이용자가 급증하고 있으며, 인터넷 이용의 한 방법인 e-mail의 이용도 활성화되고 있다. e-mail은 기존의 문서우편에 비해 그 편리성과 신속성으로 인해 인터넷의 활용측면에서 빼놓고 이야기 할 수 없는 중요한 수단이 되고 있다. E-Mail은 기존의 문서우편에 비해 논리적 공간에서 통신망을 이용해 빠른 속도로 다수에게 전달할 수 있다는 특징을 지니고 있다[1].

기존의 메일 에이전트 시스템에서는 수신된 메일을 무분별하게 서버에 적재하는 방식이었다. 그렇기 때문에 대량의 메일을 수신해야 하는 경우, 메일의 내용은 물론 메일에 첨부되어온 파일을 보기 위해서는 관리자가 일일이 확인하고 분류해야 하는 번거로움이 있었다. 수신된 메일을 적절하게 분류하여 목적에 맞는 폴더로 첨부파일을 자동으로 분류해주는 에이전트를 사용함으로써 관리자의 업무부담을 줄이고, 메일을 효과적으로 관리할 수 있을 것이다.

본 논문에서는 관리자의 업무부담을 줄이고 첨부파일의 효과적인 관리를 위해 메일의 내용을 읽어 keyword를 추출한 뒤 카테고리별로 폴더를 생성하고, 폴더 내에서 다시 파일 포맷 형식별로 폴더를 생성하여 해당 폴더별로 첨부파일을 분류 시켜주는 첨부파일 형식별 자동분류 에이전트에 대해 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 메일 시스템에 대한 기존의 연구에 대해 언급한다. 3절에서는 제안하는 E-Mail 시스템의 첨부파일 형식별 자동분류 에이전트의 전체 구성에 대해 설명하고, 4절에서는 본 논문의 결론과 추후 연구 계획에 대해 언급한다.

2. 본 론

2.1 관련 연구

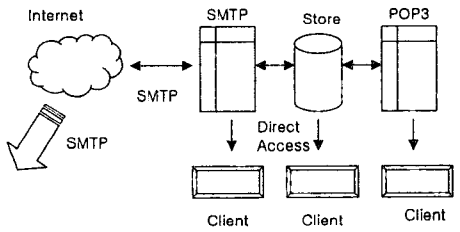
2.1.1 POP(Post Office Protocol)

POP는 RFC 1460에 기술되어 있으며, 메일서버에서 메일을 회수하는 프로토콜이다. 기존의 메일서버는 메일을 회수하기 위해서는 메일 서버에 로그인 한 후 메일 프로그램을 사용해서 메일을 확인하였으나, 현재 모든 Mail Agent(Netscape, Outlook Express, Eudora등)는 이 POP 및 SMTP를 사용하여 메일서버에 직접 로그인 하지 않고 메일을 확인할 수 있으며, 메일을 보낼 수 있다.

2.1.2 SMTP(Simple Mail Transfer Protocol)

인터넷 메일 시스템은 RFC 821 "Simple Mail Transfer Protocol"에 정의되어 있으며, 편지 배달 에이전트(IIS의 SMTP서버 같은 메일서버)가 다른 에이전트에게 편지를 전달하는데 사용하는 프로토콜이다.

(그림 1)은 전자우편 배달과정을 보여준다.



(그림 1) 전자우편의 전송

2.1.3 MIME(Multipurpose Internet Message Extensions)

이진파일 추가와 멀티미디어 지원을 추가하는 프로토콜로서 RFC 1521, 1522에 기술되어 있다. 물론, 기존의 RFC 822 메시지 포맷과 호환되고, 이진파일 전송, 메시지유형의 결정, 새로운 문자집합, 미래를 위한 성장 지원 등을 포함하고 있다.

2.1.4 메시지 구조(MIME)

RFC822 메시지 구조에 추가한 헤더 필드로 수신자는 메시지가 MIME 구조인지 확인할 수 있으며, MIME으로 해석하게 된다[2].

메일 헤더의 각 구성요소는 다음과 같다.

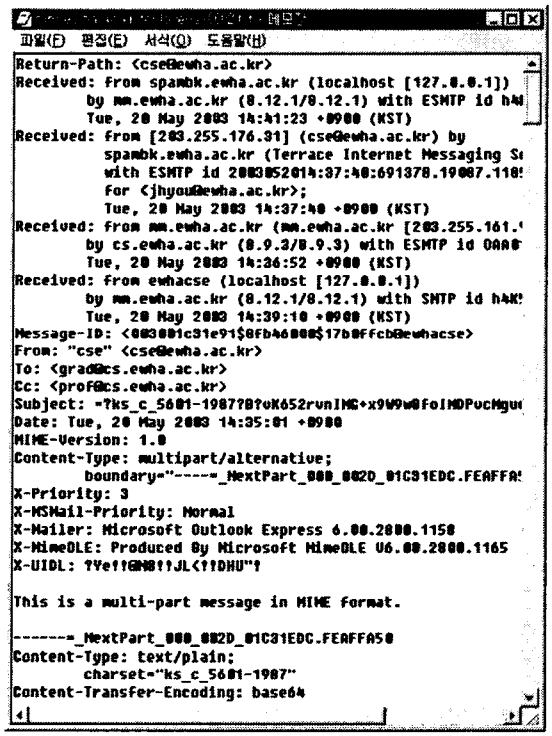
"Return-Path"는 수신자에게 메시지를 배달하는 마지막 메일 서버에서 덧붙이는 필드로 송신자로 메일을 다시 반환할 주소와 경로 등의 명확한 정보를 포함하고 있다.

"Received"는 편지가 배달되는 경로를 나타내며 "Received"가 한 줄 이상 나타나는 경우에는 메일 보내고 받는 서버 이외의 다른 서버들을 통과해온 것을 의미한다.

"From"은 보내는 사람의 주소를 의미한다.

"Reply-To"필드는 메시지를 처음 받은 메일 서버에서 추가하는 정보로 수신자가 메시지에 대한 답변을 회신할

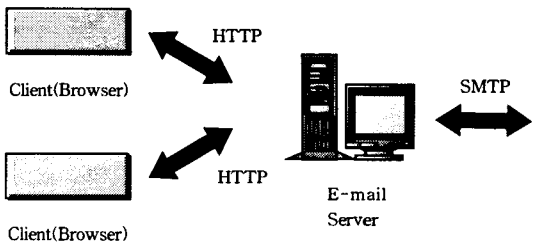
주소가 된다. 만약에 송신자가 이 필드를 비워둔다면 "From"필드의 주소로 회신하게 된다.
"To"는 수신자의 주소를 나타내며 "Subject"는 이메일의 제목, "X-mailer"는 송신자가 사용한 메일 클라이언트 프로그램, "Data"는 이메일이 보내진 날짜를 의미한다.
"Message-Id"는 해당 이메일에 지정된 식별 번호로 메일 서버가 메시지를 외부로 보내며 붙이는 일련번호로 해당 메시지가 어떤 컴퓨터에서 보내졌는지 알 수 있다.
"Content-Type"은 메일 본문이 어떤 형태인지 알려주는데 text/plain은 일반 문자열을 사용한 본문이고, 일반 문자열과 여러 인코딩 방식이 섞여 있을 경우에, multipart/mixed는 일반 문자열과 파일을 첨부하였을 때, multipart/alternative는 같은 내용이 일반 문자열과 HTML로 반복하여 선택하여 읽을 수 있는 경우, multipart/related는 HTML 형식의 메시지를 보내며 배경그림을 첨부했을 때 사용된다.
"Content-Transfer-Encoding"은 본문이 인코딩된 방식을 표시하는데, 한글 메시지에 8비트라고 표시되어 있으면 인코딩 없이 본문을 그대로 보낸 것이고 BASE64라든가 Quoted printable이라고 적혀 있으면 그런 방식으로 인코딩 했다는 의미이다[3].



(그림 2) MIME 메시지 구조

2.1.5 웹 기반 전자우편 시스템

웹 기반 전자우편 시스템은 계정과 서비스를 제공하는 서버와 사용자와의 인터페이스 역할을 담당한다. 브라우저를 통해서 전자우편 서비스를 제공하는 웹 사이트에 접속한 사용자는 ID와 암호를 가지고 자신의 정보를 관리하게 된다. 또한 사용자는 서버에 자신의 전자우편 주소를 가지게 되며 이를 이용하여 다른 사용자와 전자우편을 주고 받을 수 있다. 전자우편 클라이언트 역할을 하는 브라우저는 사용자로부터 데이터를 입력받으며, 이는 HTTP(HyperText Transmission Protocol) 프로토콜을 이용하여 서버에 존재하는 CGI(Common Gateway Interface)프로그램에게 전달된다. CGI프로그램은 전달받은 사용자 데이터를 수정하여 실제로 전달 가능한 전자우편의 형태로 만들며 SMTP프로토콜을 이용하여 수신자에게 전달된다. (그림 3)는 웹 기반 전자우편 시스템의 동작을 보여준다.

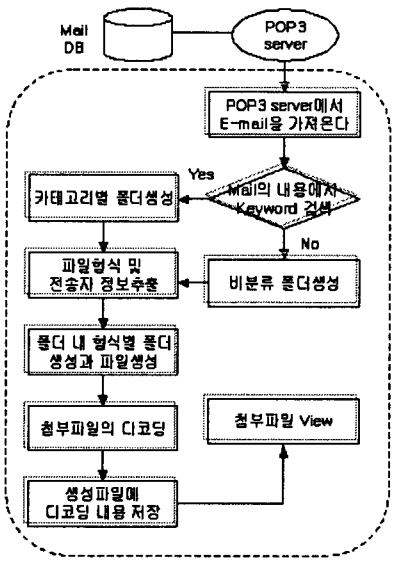


(그림 3) 웹 기반 전자우편 시스템의 동작

2.2 첨부파일의 형식별 자동분류 에이전트

2.2.1 전체구조

본 논문에서 제시한 E-Mail 시스템의 첨부파일 형식별 자동분류 에이전트의 구조는 (그림 4)과 같다.
에이전트는 관리자의 PC로 가져온 메일의 내용에서 keyword를 추출하여 첨부된 파일을 카테고리별과 형식별로 생성한 폴더 내에 자동으로 분류하여 준다.
본 논문에서 제시한 첨부파일 형식별 자동분류 에이전트 시스템을 이용하면 대량의 첨부파일에 대한 관리를 수월하게 할 수 있다.

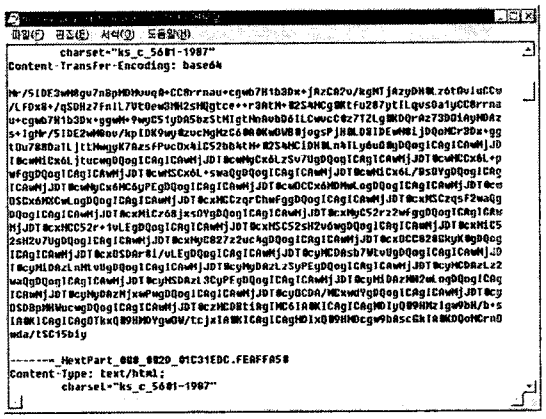


(그림 4) 에이전트 전체구성도

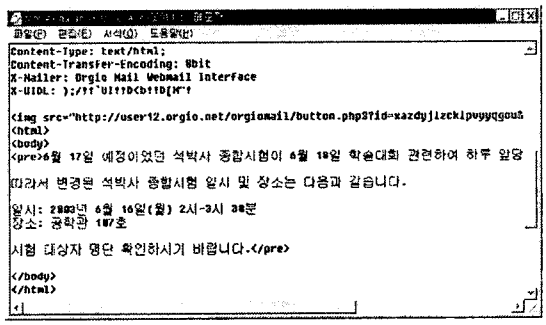
2.2.2 동작원리

본 논문에서 제시한 첨부파일 형식별 자동분류 에이전트의 동작원리는 다음과 같다.
전송자가 보낸 메일과 첨부파일은 한 곳의 메일서버에 저장된다.
서버에 저장되어 있는 메일을 관리자가 POP3 서비스를 이용하여 관리자의 PC로 가져온다. 메일은 텍스트 형식

으로 읽어 들이게 되며 한 줄씩 본문에 포함된 keyword를 검색, 추출한다.
메일의 Content-Transfer-Encoding이 8bit인 경우에는 인코딩 없이 본문을 그대로 전송한 것이므로 특별한 디코딩 작업 없이 keyword를 추출한다.
Content-Transfer-Encoding이 Base64라든가 Quoted printable이라고 적혀 있으면 그런 방식으로 인코딩 했다는 의미이므로 디코딩 작업이 필요하게 된다.
본 논문에서는 Base64로 인코딩 되어 온 경우에 디코딩할 수 있는 알고리즘을 구현하였다.
(그림 5)는 Base64로 인코딩 되어 온 메일 본문의 화면을, (그림 6)은 인코딩 없이 전송된 메일 본문의 화면을 보여준다.



(그림 5) Base64로 인코딩 되어 전송된 본문



(그림 6) 인코딩 없이 전송된 본문

미리 지정해놓은 keyword가 발견되면 관리자의 PC에 keyword 관련 폴더를 생성한다. 이때, 메일의 내용 중에서 첨부되어 온 파일의 형식을 검색, 추출하여 관련 폴더 내에 형식별 폴더를 재생성하고 전송자의 이름과 전송시간, 첨부파일의 이름을 추출하여 "전송자의 이름_전송시간_첨부파일명.포맷형식" 명으로 파일을 생성한다.
첨부되어 오는 파일은 Base64로 인코딩 되어 오므로 따로 추출하여 디코딩 한다.
마지막으로, 디코딩 한 내용을 형식폴더 내 미리 생성해 놓았던 파일에 저장한다.
keyword를 찾을 수 없으면 비분류 폴더를 생성하고, 파일의 형식 검색의 과정부터는 keyword 발견시와 동일하게 수행한다.

2.2.3 분류과정

① 서버에 저장되어 있는 메일을 관리자가 POP3 서비스를 이용하여 관리자의 PC로 가져온다.
② 가져온 메일의 내용을 한 줄씩 텍스트 형태로 읽어 본문에 포함된 keyword를 검색, 추출한다.
③ keyword가 발견되면 관리자의 PC에 각 카테고리별로 폴더를 생성한다.
④ 메일의 내용 중에서 첨부되어 온 파일의 형식을 검색, 추출하여 관련 폴더 내에 폴더를 재생성하고 전송자의 이름과 전송시간, 첨부파일의 이름을 추출하여 "전송자의 이름_전송시간_첨부파일명.포맷형식"명으로 파일을 생성한다.
⑤ 메일의 내용 중에서 Base64로 인코딩 되어 첨부된 파일만을 따로 추출하고 디코딩 한다.
⑥ 디코딩한 내용을 미리 생성해놓은 파일에 저장한다.

3. 결 론

정보통신기술의 발전과 인터넷의 보급 확산으로 종래의 우편제도를 통한 의사소통방식 외에 전자우편에 대한 의존이 증가하고 있다. 하지만 전자메일의 무분별한 수신은 전자메일 사용자의 검색시간 증가로 인한 업무 손실을 발생시킨다. 수신된 메일을 적절하게 분류하여 목적에 맞는 부서로 첨부파일을 자동으로 분류해주는 에이전트를 사용함으로써 관리자의 업무부담을 줄이고, 메일을 효과적으로 관리할 수 있을 것이다.
이 논문의 한계는 첨부파일을 분류하기 위한 방법으로 키워드를 미리 지정해 주는 데 있다. 이는 미리 지정해 놓은 키워드를 포함하지 않은 문서에 대해서는 분류가 어렵다는 단점을 가지고 있다. 따라서 성능 향상을 위하여 문서들의 특징을 전혀 모르는 상황에서도 문서 내용에서 공통된 패턴을 발견하고 문서를 분류할 수 있는 기계학습 기법을 이용한 자동 문서분류에 대한 연구가 필요할 것이다[4][5].

[참 고 문 헌]

[1] 김한섭, 배수정, 연현정, 황윤철, 이상호, "개인정보보호를 위한 전자메일 주소 추출 방지 기법", 정보과학회 추계학술대회, 2002. 10.
[2] 임양원, 권기훈, 임한규, "서비스 엔진을 이용한 웹 기반 메일 에이전트 시스템의 설계 및 구현", 한국정보처리학회 논문지 제7권 제2호, 2002. 2.
[3] 정옥란, 조동섭 "유사성을 이용한 효율적인 메일 그룹핑 전략", 한국정보처리학회 춘계학술발표논문집 제 9권 1호, 2002.
[4] C. Buckley, G. Salton and J. Allan "The Effect of Adding Relevance Information in a Relevance Feedback Environment, "Proc. 17th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.292-298, 1994.
[5] G. Salton, and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York. 1983.