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

E-Mail 시스템의 첨부파일 자동분류 에이전트 설계

현영순*, 정옥란, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과
e-mail : {toyloveys, orchung, dscho}@ewha.ac.kr

Agent for Classifying the Attached File in E-Mail System

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

요 약

최근 인터넷 사용자의 증가와 함께 e-mail 사용자 또한 늘어나고 있다. 기존의 메일 기반 에이전트는 서버에 무분별하게 메일을 저장하는 방식이기 때문에 수신된 메일을 관리자가 일일이 읽어보아야 하는 비효율적인 시스템이다. 본 논문에서는 도착한 메일의 내용을 분석하여 메일과 메일에 첨부되어 온 파일을 자동으로 분류, 전달하는 시스템을 제안하고자 한다. 이는 대량의 메일을 수신한 경우 관리자의 업무부담을 줄이고, 메일을 효과적으로 관리할 수 있는 장점이 있다.

1. 서론

인터넷 사용자가 증가하면서 전자우편(e-mail)이 현대인의 필수 커뮤니케이션으로 자리잡아 가고 있다. 사용자들은 전자우편을 통해 단순한 텍스트 뿐만 아니라 여러 종류의 다른 문서, 이미지 등을 첨부해 송수신 하기도 한다. 또한 전자우편은 개인적인 목적뿐만 아니라, 광고, 판매, 서비스 등의 특정목적에 이용되고 있는 추세이다[1].

대량의 메일을 수신하는 기업이나 학교의 경우, 한꺼번에 많은 메일을 처리 및 관리하기 위해서 메일 문서 및 첨부파일에 대한 자동분류가 필요하다. 기존의 메일 기반 시스템은 수신된 메일을 무분별하게 서버에 저장하는 방식이기 때문에 수신된 메일을 회수하기 위해서는 관리자가 서버프로그램에 접속하여 자신의 PC 로 가져오고, 가져온 메일을 하나하나 읽고 분류하여 해당 부류로 전달해야 하는 부담이 있다.

본 논문에서는 이러한 단점을 보완하기 위해 메일의 내용을 분석하여 메일과 메일에 첨부되어 오는 파일들을 자동으로 분류, 기업이나 학교로 들어오는 대량의 전자우편을 해당 부류로 라우팅 시키는 자동분류 시스템을 제안하고자 한다.

자동분류 시스템을 사용한다면 관리자의 업무부담을 줄이고, 메일을 효과적으로 관리할 수 있으며 메일을 통해 들어온 질의등에 대한 답변시간 또한 줄일 수 있을 것이다.

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

2. 관련 연구

2.1 일반적인 전자우편 시스템

전자우편 시스템은 전자우편 클라이언트 프로그램으로 SMTP(Simple Mail Transfer Protocol)서버에 편지를 보낸다. 전자우편 메시지 수신인의 도메인이 서버가 있는 도메인과 같으면 아스키 텍스트를 서버에 저장하고, 자기 도메인이 아닌 경우에는 해당 서버에 전달한다. 전자우편 메시지가 수신자의 SMTP 서버에 도착하면 수신자에게 할당된 사서함에 저장된다. 수신자는 아웃룩 익스프레스 같은 우편 클라이언트에서

2003 년도 한국학술진흥재단의 지원에 의하여 연구되었음

POP3(Post Office Protocol) 프로토콜로 서버에 접속해서 자신에게 온 편지를 가져간다. 이 과정에서 포트 25 를 사용하는 SMTP 로 편지를 서버에게 전달하고, 포트 110 을 사용하는 POP3 프로토콜로 편지를 클라이언트까지 배달한다.

일부 시스템(주로 유닉스 플랫폼)에서는 SMTP 편지 저장 공간에 직접 접근할 수 있다. 즉, 사용자가 SMTP 서버에 접속해서 SMTP 의 사서함으로부터 전자우편을 직접 읽을 수 있는 것이다. 그러나, 사용자의 관점에서 보면 아웃룩 익스프레스처럼 일단 사서함에서 클라이언트의 우편함으로 편지를 옮긴 후 읽는 편이 더 편리하다[2].

2.2 웹 기반 전자우편 시스템의 동작원리

현재 많은 회사에서 일반 사용자들에게 무료로 서비스하고 있는 웹 기반 전자우편 시스템의 동작원리를 나타내면 다음과 같다. 사용자는 일반 웹 브라우저를 통해 서버에 접속한 후 보내고자 할 내용을 입력하고 POST 방식으로 서버에 전송하면 서버는 전송받은 메시지를 CGI 프로그램에서 파싱한다. 파싱된 메시지는 연속적인 8-bit 를 4 개의 ASCII 문자로 변화시키는 Base64 인코딩 작업을 거친 후, *sendmail* 프로그램을 구동해서 목적지 서버로 전송하게 된다. 목적지 서버에 도착한 메시지는 서버에 저장되어 있다가 수신자가 웹 브라우저를 통해 서버에 접속해서 요청을 하게 되면 반대과정을 거쳐 수신자의 브라우저에 보여지게 된다[3][4][5].

웹 기반 전자우편 시스템은 사용자가 특정회사의 전자우편 클라이언트 프로그램을 구입하지 않아도 브라우저만 있는 환경이면 전자우편을 송수신할 수 있도록 함으로써 사용자에게 매우 편리한 환경을 제공한다. 이러한 환경은 기업이나 학교 등의 인트라넷 환경에 적용될 수 있으며 공동의 사용자 인터페이스를 제공하므로 업무 수행능력이나 소속감 등을 높이는 데에도 기여할 수 있다.

웹 기반 전자우편 시스템을 통하여 수신된 메일들은 다양한 내용을 포함한다. 기업에서 인트라넷 환경에 웹 기반 전자우편 시스템을 적용하여 사용할 경우 기업은 매우 많은 양의 메일을 수신할 것이고 그 내용 또한 방대할 것이다. 사용자가 보내는 방대한 양의 메일을 데이터베이스에 저장하기만 한다면 관리자의 입장에서는 메일들을 일일이 읽어보아야만 하는 번거로운 작업이 된다. 따라서 수신된 메일을 적절하게 분류하여 목적에 맞는 부서로 메일을 자동으로 보내주는 에이전트가 필요하다.

3. E-Mail 시스템의 첨부파일 자동분류 에이전트의 설계

3.1 첨부파일 분류 과정

기존의 메일 기반 시스템은 수신된 메일들을 분류작업 없이 서버에 모두 저장하는 방식이다. 따라서 적재된 메일을 회수하고 분류하기 위해서는 관리자가 수신된 메일을 직접 일일이 읽고 분류하여 관련된 폴더로 전달해야 하는 번거로움이 있다.

시간과 노력을 절감하고 대량의 메일들을 효과적으로 관리하기 위해서 수신된 메일을 자동으로 분류, 전달해주는 시스템의 중요성이 증가하고 있다.

E-Mail 시스템의 첨부파일 자동분류 에이전트는 서버에 저장되어 있는 메일을 관리자가 POP3 서비스를 이용하여 관리자의 PC 로 가져온다. 가져온 메일의 내용을 텍스트 형태로 읽어 본문에 포함된 특정단어를 검색, 추출한다. 분류를 통해 메일의 카테고리를 결정한 후에는 이미 구성되어져 있는 메일 데이터베이스의 각 카테고리 테이블에 메일을 저장한다. 만약 메일을 통해 수업의 과제물을 제출한다고 보자. 학생들이 보낸 메일의 내용을 자동분류 시스템이 읽어 들여 제목이나 특정단어를 추출한다. 과목별로 분류하여 저장하기를 원한다면 메일의 내용에서 그 부분의 단어를 추출하고 분류하여 첨부되어온 과제를 각 카테고리 테이블에 저장한다. 사용자는 자신의 첨부파일을 전자우편으로 관리자에게 보내기만 하면 에이전트를 통해 자동으로 분류가 되고 메일 데이터베이스의 각 테이블에 저장되게 된다. 저장된 메일과 첨부파일은 각각의 담당 폴더로 보내지고 첨부파일 처리에 대한 확인 메일 또한 자동으로 전송자에게 보내지도록 한다. 이는 기존의 메일 기반 시스템과는 달리 관리자의 수작업이 필요 없는 시스템이므로 시간과 노력을 절감하고 수신된 메일을 효과적으로 관리할 수 있는 장점이 있다.

3.2 에이전트 전체 구조

본 논문에서 제안하는 E-Mail 시스템의 첨부파일 자동분류 에이전트의 전체 구성도는 다음 (그림 1)과 같다.

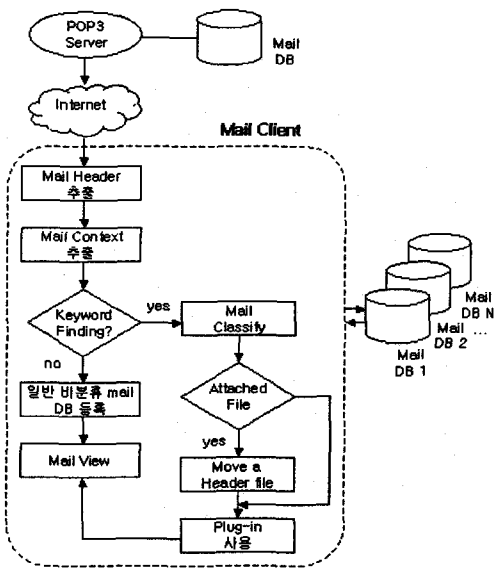

그림 1. 에이전트 전체 구성도

각 과정은 다음과 같은 단계를 거친다.

모든 mail message 에 대해

Begin

1. Search for header
2. Search for context using predefined keywords
3. Extract attached files
4. Move attached files to file folders
5. Using SMTP, send e-mail to mail sender
6. View results on client screen

End

본 연구에서는 데이터베이스에의 저장을 위해 특정 단어(Keyword) 검색 방법을 이용하기 때문에 매칭 테이블을 결정하기 위한 키워드를 지정해 주어야 한다. 에이전트를 어떻게 설계할 지에 따라 키워드의 내용이 달라진다. 특정단어의 추출은 Loop 문을 이용하여 본문 text 를 검색하도록 한다.

3.3 메일 및 첨부파일 자동분류 시스템

사용자가 e-mail 을 통해 전송한 메일은 한 곳의 서버에 저장되고 저장된 메일과 첨부파일은 Keyword 검색을 통하여 데이터베이스 내의 매칭되는 테이블에 저장된다. 이때 일정한 시간이 경과하면 새롭게 수신된 메일이 있는지 확인하고 새로운 메일이 있다면 메일 데이터베이스에 저장되도록 한다.

에이전트는 테이블에 저장된 메일과 첨부파일을 관련 폴더에 전달하고, 전송자에게 SMTP 서비스를 이용하여 처리확인 메일을 보낸다.

다음 (그림 2)는 새로 수신된 메일이 있을 때, 수신된 메일을 확인해주는 화면이다.

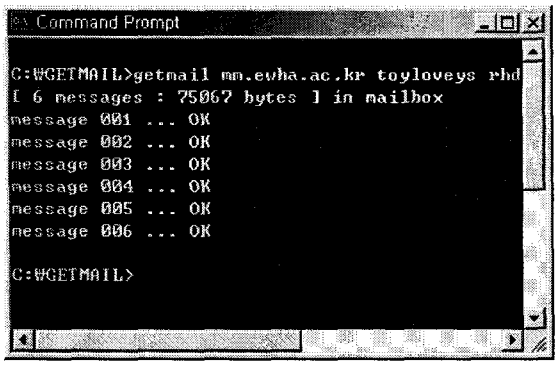

그림 2. 메일서버에 수신된 메일의 수

다음 (그림 3)은 수신된 메일을 텍스트 파일로 변환한 화면이다. 변환된 파일에서 각각의 필드에 넣을 데이터를 찾아 변수에 저장한다. 그리고 난 후 본문의 내용에서 키워드를 검색하여 데이터를 각각의 테이블에 저장한다.

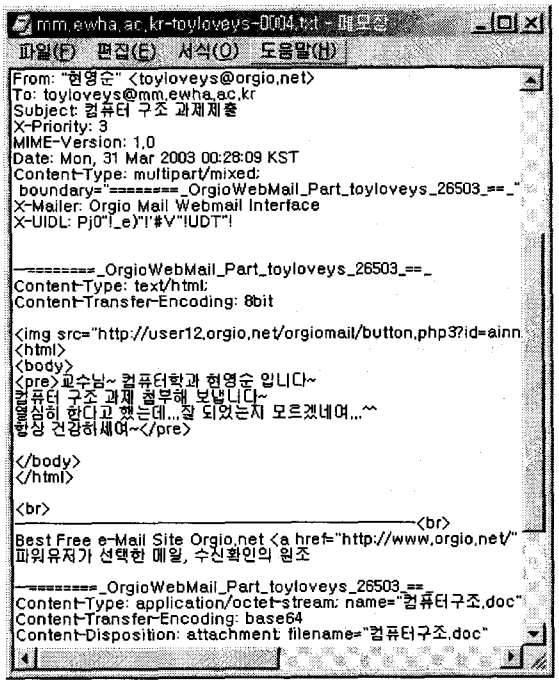

그림 3. 수신된 메일을 텍스트로 변환

다음 (그림 4)는 각 폴더에 분류되어 저장된 화면이다.

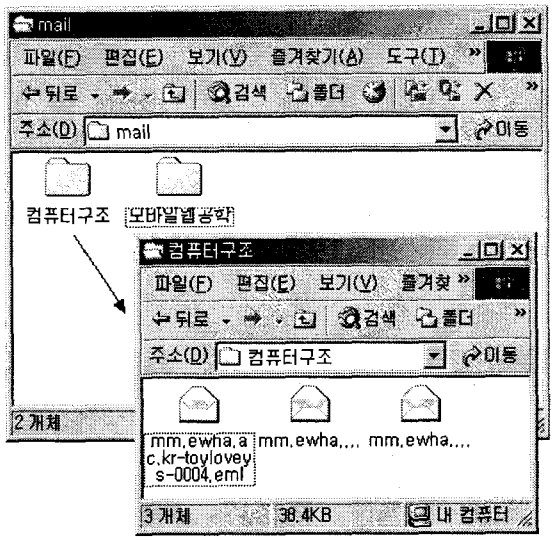

그림 4. 항목별 분류된 메일들

4. 결론 및 추후연구

인터넷 사용자의 증가와 더불어 e-mail 사용자 또한

증가하고 있다. E-mail 은 전화나 FAX 를 대체하는 차세대 커뮤니케이션으로 현대인의 필수 커뮤니케이션으로 자리잡아가고 있다. 사용자들은 e-mail 을 통해 단순한 텍스트 뿐만 아니라 여러 종류의 다른 문서, 이미지 등을 첨부해 송수신하기도 한다.

기존의 메일 기반 에이전트는 서버에 무분별하게 메일을 적재하는 방식이기 때문에 수신된 메일을 관리자가 일일이 읽어보아야 하는 비효율적인 시스템이다.

본 논문에서는 도착한 메일의 내용을 분석하여 메일과 첨부 파일을 자동으로 분류, 전달, 처리확인 메일까지 전송하도록 하는 시스템을 제안하였다. 본 논문에서 제안한 에이전트를 사용할 경우 기존의 시스템들에서 나타나는 단점인 관리자의 메일관리에 대한 번거로움을 줄이고, 메일을 보다 효율적으로 관리할 수 있다는 장점이 있다. 또한 메일을 통해 들어오는 질의등에 대한 답변 시간을 단축하고, 분류·저장된 문서들에 대해서는 통계 결과도 기대할 수 있을 것이다.

이 논문의 한계는 메일을 분류하기 위한 방법으로 매칭 테이블을 결정하기 위한 키워드를 미리 지정해 주는 데 있다. 이는 미리 지정해놓은 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 문서 분류에 대한 연구를 할 것이다 [6][7].

참고문헌

[1] 강영순, 이용배, 김태현, 조숙현, 맹성현, "전자우편 문서의 효율적인 분류를 위한 전처리," 정보과학회 2002 년 춘계학술대회, 2002.4.

[2] 임양원, 권기훈, 임한규, "서비스 엔진을 이용한 웹 기반 메일 에이전트 시스템의 설계 및 구현," 한국정보처리학회 논문지 제 7 권 제 2 호, 2002.2.

[3] 박동욱, 박재희, 김진상, 김일민. "PGP 방식을 이용한 웹 기반 전자우편 보안 시스템," 한국정보처리학회 논문지 C, 2001.2.

[4] Sol, S. and Berznieks, G., CGI/PERL : Web Scripts. M&T Books, 1997.

[5] Stallings, W, Network and Internetwork Security : Principles and Practice. Prentice Hall, 1995.

[6] 임형근, 장덕성, "색인어 연관성을 이용한 의료정보문서 분류에 관한 연구," 한국정보처리학회 논문지 B, 2002.10.

[7] G. H. John, and P. Langley, "Estimating continuous distributions in Bayesian classifiers," Proc. 11th Conf. On Uncertainty in Artificial Intelligence, Nontreal Canada, pp338-345, 1995.