효과적인 대용량 이메일 분류 및 아카이빙 시스템 설계 및 구현

Implementation and Design of Efficient Classification and Archiving System for Large Amount of Email

김 응 진, 문 지 혜, 정 호 영, 임 지 수, 송 석 일
한국교통대학교*

Kim eungjin, Moon jihye, Jung hoyoung, Lim jisu, Song seokil
Korea National University of Transportation*

요약

이 논문에서는 대용량의 이메일을 분류하여 아카이빙하는 시스템을 설계하고 구현한다. 이 논문에서 개발하는 이메일 아카이빙 시스템은 업무영역 별로 이메일을 분류하여 업무 관련 이메일에 대해 업무영역 카테고리별로 아카이빙을 수행한다. 분류의 정확도를 위해 온톨로지를 이용한 텀벡터의 확장 방법을 사용하였으며, 빠른 분류 및 아카이빙을 수행하도록 인메모리 기반의 분산 및 병렬 처리 프레임워크인 Spark을 기반으로 구현한다.

Ⅰ. 서론

이메일 아카이빙은 이메일 메시지를 보관하고 보호하여 시간이 지난 후에도 지속적으로 접근할수 있도록 하는 자동화된 절차를 말한다[1]. 이메일 아카이빙을 통해서 실수로 지워진 이메일을 복구하거나, 이메일과 첨부파일에 포함된 회사의 지식재산을 보호하거나 전자증거개시제도(eDiscovery)에 대응할 수 있다.

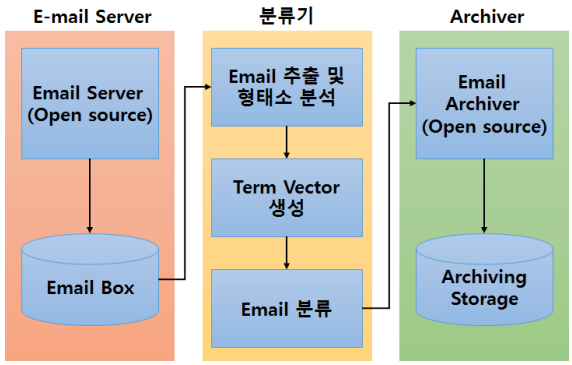
이 논문에서는 기업에서 업무 이메일을 아카이빙할 때 업무 영역에 따라 분류하여 아카이빙 할수 있도록 하며 개인 메일이나 광고성 메일은 아카이빙 하지 않도록 한다. 기존에 이메일을 자동으로 분류하는 여러 시스템이 존재한다. [2]에 따르면 최근에 개발된 이메일 분류 시스템은 Google 의 Inbox[3]. IBM의 Verse[4] 이다. 이 시스템들은 메일 사용 내역을 분석하여 수개의 메일함을 만들고 수신하는 메일을 각 메일함에 분류하여 저장한다.

[2] 에서는 Inbox 와 Verse의 분류기법에서 더 나아가 메일을 주제 중심으로 분류하는 방법을 제안하고 있다. 이 방법은 서버측에서 분류하는 것이 아니라 모바일 디바이스에서 분류할 수 있도록 하는 방법을 제안하고 있다.

이 논문에서는 업무 이메일을 업무 영역에 따라 분류하되, 각 업무 영역의 온톨로지를 구축하고 이를 이용하여 보다 정확한 분류가 되도록 하는 방법을 구현하였다. 또한, 대용량의 이메일을 빠르게 분류하기 위하여 인-메모리 병렬 및 분산처리 프레임워크인 Spark[5]을 사용하여 설계 및 구현한다. 또한, 아카이빙 된 업무 이메일을 빠르게 검색할 수 있도록 업무영역 및 이메일의 단어를 기반으로 하는 색인을 구축하도록 한다.

Ⅱ. 아카이빙 시스템 설계

이 논문에서 설계하고 구현하는 이메일 아카이빙 시스템의 전체적인 구조는 그림 1과 같다. 아카이빙 시스템은 크게 이메일 서버, 분류기, 아카이버로 구성된다. 분류기는 이메일 서버의 메일박스에서 주기적으로 이메일을 가져와서 이메일의 내용을 분석하고, 각각의 이메일을 카테고리 별로 분류하여 아카이버에 전달한다. 아카이버는 분류기의 결과를 바탕으로 아카이빙을 수행하고 빠른 검색을 위한 색인을 수행한다.



▶▶ 그림 1. 이메일 아카이빙 시스템 구조

본 논문에서 설계 및 구현한 이메일 아카이빙 시스템에서는 이메일 서버는 오픈소스로 지원되는 Zimbra[6]를 사용하였다. Zimbra는 자체 메일 박스에 메일 메시지를 저장하며 REST-API를 통해 저장된 이메일 메시지를 외부로 내 보낼 수 있다.

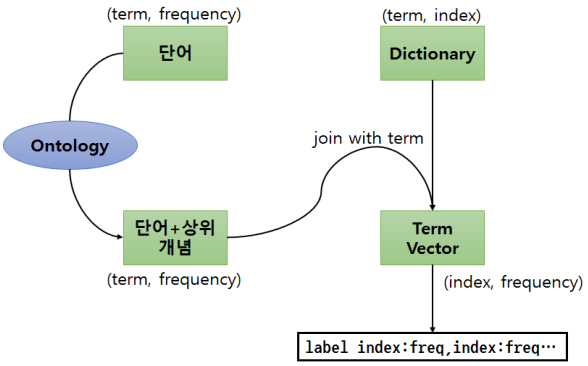
본 논문에서 구현하는 분류기는 Zimbra 로부터 주기

적으로 이메일을 가져와서 형태소 분석등의 전처리 과정을 거쳐 용어 벡터 (Term Vector)를 생성한다. 각 메일로부터 생성한 용어벡터를 이메일 분류 알고리즘에 입력하여 분류를 수행한다. Zimbra로부터 이메일을 가져올 때는 메일 표준 형식 중 하나인 eml 형식을 이용한다. eml 형식의 이메일 파일은 송신자, 수신자, 메일 제목, 메일 본문 등의 정보를 가지고 있다.

이 파일에서 이메일 분류를 위해 사용할 항목들을 추출하고 구조화 한다. 이 논문에서는 본문의 내용을 표현할 수 있는 단어들을 추출하기 위해 한국어 형태소 분석기 "은전한닢"을 사용한다. 이 형태소 분석기는 일본어 형태소 분석기인 "mecap"을 바탕으로 만들어졌고, 여러 가지 프로그래밍 언어를 통해 사용할 수 있는 라이브러리를 지원한다. 이 형태소 분석기를 이용하여 이메일 파일에서 추출한 항목들 중 본문 내용의 문장 성분 중 문장의 의미를 잘 표현할 수 있는 품사인 "명사"를 추출하여 분류기의 입력 데이터를 만들었다

생성된 텀벡터는 관리자에 의해서 구축된 각 업무영역의 온톨로지를 기반으로 확장된다. 다음 [그림 3]에는 이메일의 단어들이 온톨로지를 통해 상위 개념이 추가되고, Dictionary를 참조하여 확장된 텀벡터로 생성되는 과정을 보여준다. Dictionary는 이메일에 포함되어 있는 단어를 텀벡터에 매핑하기 위해 사용된다. Dictionary는 기존 이메일에서 추출한 단어들과 온톨로지에 정의되어 있는 모든 단어들을 가지고 있으며, 각각의 단어들이 서로 다른 색인을 가지고 있는 키-값 쌍 (Key-Value pair)의 집합이다.

이 논문에서는 이메일 분류 알고리즘으로 SVM (Support Vector Machine)을 사용하였다. SVM은 지도학습법으로 초기에 레이블링이 된 훈련 데이터 집합을 제공하여야 한다. 시스템 운영 초기에 아카이빙 관리자는 업무 영역 카테고리를 설정하여 일반 사용자에 배포하여야 하며 일반 사용자는 초기 n개의 이메일에 업무 영역 카테고리에 따라 레이블링을 수행하여야 한다. 레이블링된 훈련 데이터를 이용하여 SVM 알고리즘을 훈련하고 테스트하여 실제 분류에 적용한다. 구현에는 Spark 의 MLib에서 제공하는 SVM 라이브러리를 사용하였다.



▶▶ 그림 2. Dictionary를 이용한 단어 mapping

이 단계에서 온톨로지를 사용하여 텀벡터를 확장시키는데, 온톨로지는 트리 형태의 구조로 설계하였다. 각 노드의 부모 노드는 해당 단어의 상위 개념을 가지고 있는 형태이다. 이 상위 개념을 텀벡터에 추가함으로써 포괄적으로 같은 의미를 갖는 문서들을 같은 카테고리로 분류할 수 있다.

분류기에 의해 분류된 각각의 이메일은 아카이버에 전달이 되고 기존의 오픈소스 아카이버를 이용해서 저장관리한다. 다만, 검색의 효율성을 위해서 업무 영역 카테고리와 이메일에 포함된 주요 단어들을 이용한 색인을 구축하고 관리한다. 색인 구축 및 검색은 Lucene 과 Solr을 이용하여 구현한다.

Ⅲ. 결론

이 논문에서는 대용량 이메일 분류를 위한 아카이빙 시스템을 설계하고 구현하였다. 제안하는 아카이빙 시스템은 업무 영역에 대해 구축된 온톨로지를 기반으로 분류의 정확도를 높이는 접근법을 사용하였다. 또한, 대용량의 이메일에 대해서도 빠르게 분류 및 아카이빙을 수행하기 위해서 인-메모리 기반 분산 병렬처리 프레임워크인 Spark을 이용하여 구현하였따.

향후 연구에서는 개발한 이메일 분류 및 아카이빙 시스템의 분류 정확도와 아카이빙 된 이메일에 대한 검색 속도등에 대한 성능평가를 수행하고 최적화를 수행한다.

▌참 고 문 헌▐

[1] https://en.wikipedia.org/wiki/Email_archiving
[2] 백준기, 이용구, 이상근, "내장형 분류지능을 활용한 주제별 이메일 분류 기법", 한국정보과학회 2015 한국컴퓨터종합학술대회 논문집, pp. 1793-1795, 2015.
[3] Google Inbox. http://www.google.com/inbox
[4] IBM Verse. http://www-03.ibm.com/software/products/en/ibm-verse
[5] Zaharia, Matei, et al. "Spark: Cluster Computing with Working Sets." HotCloud 10, 2010
[6] Zimbra, https://www.zimbra.com/