e-Discovery를 위한 오픈 소스 기반의 ESI 분류 및 검색 도구 설계 및 구현

김헌*, 이태림**, 신상욱***
*부경대학교 컴퓨터멀티미디어학과전공
**부경대학교 대학원 정보보호협동과정
***부경대학교 IT융합응용공학과
e-mail:mybreathing@pknu.ac.kr

Design and Implementation of a tool for ESI Categorization and Search in e-Discovery

Hun Kim*, Tae-Rim Lee**, Sang-Uk Shin***
*Major of Computer and Multimedia Eng. Pukyong National University
**Dept of Information Security Graduate School, Pukyong National University
***Dept of IT Convergence and Application Eng, Pukyong National University

요 약

2006년 12월 1일 FRCP 개정안이 발표되면서 e-Discovery를 제정함으로 기업 시스템 내의 ESI에 대한 통합적인 관리가 필요하게 되었다. e-Discovery는 시스템 내의 ESI들을 모두 검토해야 함으로 비용과 시간을 줄이기 위한 e-Discovery 지원 도구가 필요하다. 하지만 국내의 e-Discovery 분야는 거의 시작단계이거나 이제 시작해야 하는 단계라고 볼 수 있으며, e-Discovery 지원 도구들은 외국산 s/w가 대부분이다. 본 논문에서는 e-Discovery 분야에서 가장 중요하다고 볼 수 있는 ESI 수집을 위해 이용되는 여러 e-Discovery 도구들의 기능을 분석하여 어떤 시스템에도 유연성있게 구축될 수 있는 개방적인 오픈 소스 기반의 ESI 수집 지원 도구의 설계를 제안한다.

1. 서론

정보화 시대라 일컫는 현대에는 다양한 형태의 디지털 기술 개발과 발전을 통해 업무를 비롯한 일상에서도 전자문서의 사용이 급증하게 되었다. 전자문서란 컴퓨터 기술을 이용하여 생성되거나 저장되는 정보들을 의미하며, 관련 분야의 기술에는 워드프로세서, 데이터베이스, 스프레드시트 등의 비즈니스 응용 프로그램, 이메일, 메신저, Blog 등의 인터넷 응용 프로그램, USB, CD 등 이동식 저장 매체 기술이 있다. 이외에도 수많은 관련 분야의 기술들이 새롭게 등장함에 따라 전자문서의 종류는 물론 사용 방식들도 나날이 다양해지고 있는 추세이다. 이런 흐름은 사법 제도에도 영향을 끼쳐 전자문서 형태의 증거가 재판에 많이 활용되고 있기에, 이들을 적절하게 확보하기 위한 기술 개발이 활발히 진행되고 있다[1].

특히 Discovery 제도는 소송 당사자가 상대방이나 제3자로부터 소송과 관련된 증거자료를 수집하기 위한 변론 전의 절차를 통칭하는 개념이다. 소송의 쟁점은 무엇인지, 어떠한 관련 자료가 존재하며 그 자료들을 증거로 활용할 수 있는지 평가하는 절차라고 할 수 있다. 이러한 Discovery 제도를 개선하여 제정된 e-Discovery는 앞서 언급한 현대의 흐름을 반영하고, 전자 문서의 특성상 자료의 종류나 범위가 광범위하고 유동적이며, 복제나 수정이 용이하다는 점을 고려한 제도이다. 이에 따라 2006년 12월 1일 미국의 FRCP(Federal Rules of Civil Procedures) 개정은 기존 증거 개시의 대상이 종이문서로 제한했던 것을 확대하여 ESI(Electronically Stored Information)를 포함시켰으며, ESI 종류에는 크게 소프트웨어 응용 프로그램을 통해 생산된 문서의 원본을 의미하는 Native 파일, 문서의 다양한 특성을 나타내주는 정보를 일컫는 메타데이터(Metadata), 문서 또는 파일시스템 전체에 대한 신속한 검토 분석을 가능하게 하기 위해 개발된 표준인 이미지 파일 등이 있다[2].

이러한 e-Discovery 관련 업무들을 수행하기 위한 지원 도구들은 다양한 ESI들에 대한 백업, 삭제, 검색 등 통합적인 관리를 수행하며, 소송 발생 시 빠른 대응을 가능하게 하기 위한 자동화된 디지털 데이터 관리 시스템이다. 현재 대부분의 e-Discovery 지원 도구들은 EDRM(Electronic Discover Reference Model)[3]를 참조하여 설계되고 만들어지고 있으며, EDRM이란 FRCP에서 명시하고 있는 e-Discovery 관련 요구사항들을 효과적으로 준수하기 위해 절차를 표준화하고 기능 명세를 작성한 참조 모델이다.

최근 기업 간의 민사 소송이 증가하면서 증거 산출을

* 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 20110006097), 그리고 지식경제부 및 한국산업기술평가관리원의 산업원천기술개발사업[10035157, 실시간 분석을 위한 디지털 포렌식 기술 개발]의 일환으로 수행된 연구임

위해 조직 내 모든 시스템에 존재하는 ESI를 검토해야 함에 따라 시간과 비용을 절감하기 위해서 e-Discovery 지원 도구 사용에 대한 필요성이 매우 부각되고 있다. 하지만 현재 상용화되어 있는 e-Discovery 도구들은 대부분 외산 솔루션들이기에 도입을 위한 비용 지출이 크며, 국내의 기업 환경을 고려하지 않은 제품들이기에 기능적으로 부족한 면이 많다. 이러한 비용과 시간을 최소화하기 위해서는 각자 기업의 시스템 환경에 맞는 도구가 필요할 것이며, 이를 개발하기 위한 노력이 시급하다.

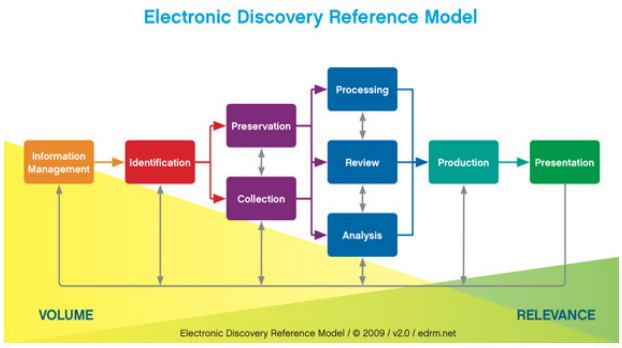
결국 e-Discovery 업무의 궁극적인 목표가 소송 쟁점에 대해 밀접한 관련이 있는 자료들을 검색하여, 증거화하는 것이라 보면 해당 업무에 소요되는 비용과 시간의 낭비를 최소화하는 방법은 효율적인 검색 기법을 마련하는 것이라 볼 수 있다. 또한 검색을 통해 수집된 ESI들을 연관성 정보를 활용하여 중요도에 따라 선별 검토할 수 있게 된다면 비용 절감에 탁월한 효과가 있을 수 있다.

이에 본 논문에서는 효율적인 ESI 검색 및 수집 업무를 수행할 수 있도록 클러스터링 기법을 기반으로 하여 자동화된 ESI 분류가 가능한 관리 도구를 설계하고 오픈 소스를 활용하여 이를 구현한다. 구현에 활용하는 오픈 소스는 Apache사의 Lucene[4]과 Mahout[5]이며, TREC의 문서 컬렉션의 일부를 이용하여 동작을 테스트하고, 도구의 발전 가능성에 대해 분석한다.

2. 관련연구

2.1 EDRM(Electronic Discover Reference Model)과 e-Discovery 지원도구

EDRM은 FRCP의 e-Discovery 요구사항들을 기반으로 효과적인 업무 수행을 위해 필요한 단계 별 기능, 관련자 및 도구의 역할들을 상세하게 기술한 참조 모델이며, 다음 그림 1과 같은 절차들을 포함하고 있다.



(그림 1) EDRM의 절차

EDRM은 e-Discovery 제품과 서비스를 개발하고, 선택 평가하기 위한 일반적이고 유연하며 확장 가능한 프레임워크를 제공해준다. 이는 여러 관련 조직들의 협의 하에 개발되었으므로 e-Discovery에 대한 일반적인 표준으로 공인되어 활용되고 있다[3].

대표적인 e-Discovery 도구는 다음과 같다.

- AccessData eDiscovery[6] : 포렌식적인 수집 방법으로 법 관련 기관에서의 높은 인지도를 자랑한다. 효율성과 사용자 편의를 강조한 인터페이스, 다중 연관성 모델링 지원 및 데이터 자동 분류기술, 인덱싱을 제공한다.
- EnCase eDiscovery[7] : 데이터 식별, 보존, 수집, 가공과 리뷰 플랫폼을 위한 로드 파일 산출 기능을 제공하지만 복잡한 인터페이스가 단점이다.
- Nuix[8] : 기업 내 모든 데이터를 인덱싱 하여 소송에 대비할 수 있으며, 새로운 정보는 실시간으로 인덱싱 한다. 인덱싱 작업 이후 빠르게 ESI 산출들을 위한 업무를 수행한다.
- Clearwell[9] : 빠른 설치와 간편한 사용이 장점이며, 높은 수준의 Culling rate를 보여준다.

2.2 문서 분류를 위한 기계 학습 기법

정보 검색 분야에서는 전자 문서 관리를 위한 문서 분류 기술 연구가 활발히 진행되고 있다. 특히 웹과 같은 대규모 검색 시스템에서는 새롭게 추가되는 웹 페이지들을 자동으로 수집·분류하여 사용자 질의에 대한 검색 대상 범위를 축소함으로써 성능 향상을 꾀하고, 검색 결과에 대한 사용자 만족도를 높이기 위해 분야 별로 분류된 결과를 제공하는 등의 다양한 기술들이 개발되었다. 기계 학습 기법의 종류는 크게 3가지이며, 감독자가 제공한 학습 문서 집합을 바탕으로 훈련을 통해 분류하는 교사 학습(Supervised Learning), 학습 문서 없이 문서를 구성하는 용어들의 벡터를 활용하여 유사도에 근거한 분류를 수행하는 비 교사 학습(Unsupervised Learning), 교사와 비 교사 학습 기법을 혼용하여 활용하는 반 교사 학습(Semi-supervised Learning)이 있다[10].

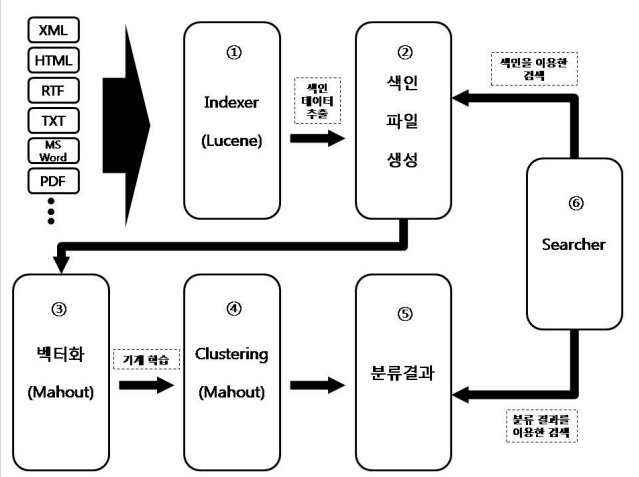
3. 오픈소스를 활용한 ESI 분류 및 검색 도구 설계 및 구현

3.1 ESI 분류 및 검색 도구 설계

기계 학습 기법들은 모두 자동화된 문서 분류를 위해 활용 가능하지만, 기업 내 모든 ESI들에 대해 분류 범주를 사전에 결정하고 학습 자료를 선정하는 일은 매우 어렵다. 이에 본 논문에서 설계를 위해 사용한 기법은 학습 자료 선정이 불필요한 비 교사 학습 방식의 클러스터링이며, ESI 분류를 위한 도구의 동작 흐름은 그림 2과 같다.

제안된 ESI 분류 및 검색 도구는 시스템에 존재하는 다양한 타입의 전자문서들을 대상으로 자동화된 문서 분류를 수행하며, 내용을 기반으로 한 검색 기능을 제공한다. 그림 2의 1번과 2번 과정은 이를 위한 도구의 첫 번째 과정인 원본 ESI들에 대한 색인 생성이다. 색인 생성을 담당하는 Indexer는 다양한 ESI 파일 포맷에 대하여 색인 데이터들을 추출 할 수 있어야 하며, 추출한 자료들의 용어들을 분석 할 수 있어야 한다. 추출된 색인 데이터들은 검색 및 자동분류를 수행하기 위해 파일로 존재한다. 그림

2의 3번 과정은 자동 분류를 위한 벡터화 과정이다. 벡터화 과정은 문서를 구성하고 있는 용어들을 기준으로 가중치를 적용하여 재 표현 하는 작업이며, 출현 빈도수와 같은 통계적 수치를 활용한다. 그림 2의 4번과 5번 과정은 클러스터링 알고리즘을 적용한다. 문서 클러스터링은 벡터화 된 데이터들을 이용하여 각 문서들의 거리 척도를 계산하고 유사도를 분석하여 문서들을 관련된 문서 군집으로 분류하여 결과를 보여준다. 위 과정을 모두 수행한 ESI 분류 및 검색 도구는 그림 2의 6번의 원본 ESI들에 대한 색인과 문서 클러스터링 분류 결과를 도출하여 사용자로 하여금 이를 이용한 검색이 가능하게 한다.



(그림 2) ESI 분류 및 검색 도구 시스템 구성

3.2 오픈 소스를 활용한 도구 구현

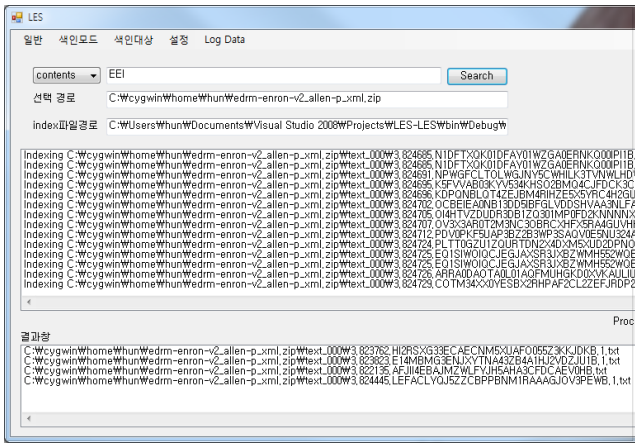
제안된 ESI 분류 및 검색 도구 시스템 구성을 이용하여 다양한 타입의 전자문서들을 색인 및 클러스터링 결과를 보여 줄 수 있는 Prototype을 구현하였다. 구현에 사용된 언어는 C#언어 이며 개발 도구로는 Visual Studio 2008을 이용하여 작업하였다. 설계된 ESI 분류 및 검색 도구의 각 과정 구현을 위하여 사용된 오픈소스는 다음과 같다.

- Apache Lucene : 확장 가능한 고성능 정보검색 오픈 소스이다. 검색 소프트웨어가 아닌 라이브러리로써 시스템의 문서 파일들을 Word 단위로 나누어 색인 작업과 Word를 분석하는 분석기 및 검색 기능을 가지고 있다. Lucene은 기본적으로 TXT 포맷을 색인할 수 있으며 그 이외의 문서 포맷은 별도의 파서를 이용하여 내용을 읽어 색인 작업을 수행 할 수 있다.
- Apache Mahout : 데이터를 분석해 샘플을 처리하고 추출하는 것 뿐만 아니라 데이터를 학습시켜서 많은 양의 문서를 자동으로 분류할 때 쓰이는 오픈소스 라이브러리이다. Lucene의 색인화된 정보들을 문서별로 벡터화 하여 Clustering기법을 이용한 문서 자동 분류에 이용하기 위해서 사용한다.

Lucene을 이용하여 첫 번째 과정인 Indexer를 구현 하였다. 색인 대상의 문서들의 내용을 Word 단위로 나눠 Word Term과 Filename Term의 두 가지 정보를 색인 파일에 저장 한다. 문서들의 Word들 중 검색 용어로 사용하지 않는 단어들은 Lucene의 Index 분석기를 통하여 불용어로 처리 했다. 생성된 색인 파일을 이용하여 두 번째 과정인 벡터화 작업은 Mahout의 라이브러리중 Mahout vector driver를 이용하여 색인 파일을 벡터화 하고 Mahout Clustering의 입력 파일로 만든다. 벡터화 과정에서는 용어에 대한 가중치 적용을 위해 TF-IDF(Term Frequency - Inverse Document Frequency)를 이용하였다. TF-IDF는 단어의 빈도수와 문서의 빈도수의 역수의 곱을 이용 하여 각각의 문서들의 가중치 값을 정하는 방법이다. 벡터화를 통해 생성된 결과 파일은 세 번째 과정인 클러스터링에 활용된다. Mahout Core에 있는 K-means 클러스터링 알고리즘은 비 교사 학습 방식의 클러스터링 기법으로 다양한 분야에서 데이터 분류를 위해 많이 사용되는 기법이다. K-means 클러스터링 알고리즘을 적용하면 ESI들은 각각의 클러스터들로 유사도에 따라 관련 문서들이 분류되며, 최종 문서 분류 결과와 앞서 생성된 Lucene 색인 파일은 Lucene index Searcher를 이용하여 그 내용을 검색하고 확인할 수 있다.

3.3 구현 결과 테스트

실험에 사용하기 위해 EDRM에서 다운로드 받은 데이터는 E-mail을 TXT로 변환한 약 1000개의 예제 파일들이며, 이는 TREC의 Legal Track의 Learning Task를 수행하기 위해 제공되는 문서 컬렉션의 일부분이다. 그림 3은 예제 데이터를 색인하고, 색인된 문서 내용 중 "EEI"라는 단어를 검색 했을 때 나타나는 결과이다.



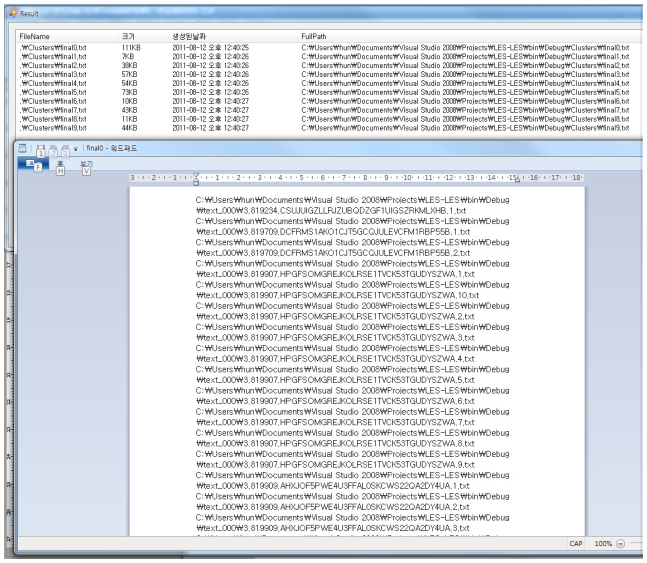
(그림 3) EDRM E-mail Data indexing test

그림 4의 선택경로는 색인 작업을 수행 할 데이터의 경로를 나타내는 것이며, index파일경로에 색인파일이 저장된다. 중앙에 있는 리스트박스에 현재 색인 진행 중인 데이터들의 경로들과 파일 이름을 보여주며, 진행 상황을 나타낸다. 색인 생성이 끝난 뒤에 상단에 있는 검색 창에 키워드를 검색하면 색인파일의 내용을 검색한 뒤 하단에 있는 결과 창에 키워드가 들어 있는 파일의 경로를 나타

낸다.

루씬으로 색인된 파일들을 이용하여 클러스터링을 하기 위해서는 문서로부터 추출된 텍스트 색인들을 벡터화 시켜줘야 한다. 벡터화 작업은 Mahout 라이브러리의 Utils에 있는 Drivers Class를 이용하면 2개의 출력 결과물을 얻을 수 있는데 하나는 색인된 단어들의 Document frequency를 나타내는 단어사전과 또 다른 하나는 벡터화된 정보를 가지고 있는 결과물이다. 단어사전에 기록된 내용은 각각 Word, Document frequency, index number를 나타낸다.

벡터화된 결과물을 이용하여 클러스터링 알고리즘 중 하나인 k-means 알고리즘에 적용한다. Mahout의 kMeansDriver Class는 벡터화된 결과물을 입력 데이터로 사용하여 연관성 있는 단어와 문서들로 군집화 시킨다. 클러스터링의 결과물은 각각의 군집화 된 클러스터들의 Top Term들을 보여주며, 이 때 Top Term은 문서를 구성하고 있는 용어들 중 가중치가 높은 용어들을 의미한다. 분류된 군집의 Top Term들을 이용하여 군집에 대한 정보를 획득할 수 있으며, 다음 그림 4는 분류된 군집과 이에 속하는 문서에 대한 절대 경로를 시각화 한 것이다.



(그림 4) 분류된 문서들의 절대 경로

그림 4의 상단에 있는 리스트박스는 관련 있는 문서들로 분류된 클러스터들의 리스트를 보여준다. 각 클러스터들은 관련된 문서의 경로를 가지고 있으며, 하단에 있는 워드패드는 그 내용을 나타낸 것이다.

소송에 연관성 있는 ESI들을 검색함에 있어서 클러스터링 결과물을 이용한다면, 검색 용어와 군집에 대한 비교를 통해 검색 영역을 축소시킬 수 있게 되어 검색 작업에 소요되는 시간을 줄일 수 있을 것이다. 또한 시스템에 추가되는 새로운 전자문서들은 색인 작업 수행 후 벡터화를 적용하고, 이미 생성된 분류 군집과 비교하여 자동으로 분류 결과를 얻을 수 있을 것이다. 또한 검색으로 확보한 ESI들을 증거로써 얼마나 가치가 있는지 검토를 수행하는 업무에서도 문서 간 유사도 정보를 선별 기준으로 삼아 연관성 높은 ESI들만을 검토 가능하게 활용할 수 있을 것이다..

5. 결론 및 향후 과제

본 논문에서는 효율적인 ESI 검색 및 수집 업무를 수행할 수 있도록 클러스터링 기법을 기반으로 하여 자동화된 ESI 분류가 가능한 관리 도구를 설계하고 오픈 소스를 이용하여 이를 구현 하였다. Lucene을 이용하여 색인 대상들을 분석하고 색인 파일로 생성하였고, Mahout를 이용하여 생성된 색인 파일들에 대한 벡터화 작업을 수행하게 하였으며, 문서 간 유사도에 따라 자동화 된 분류를 수행하기 위해 K-means 클러스터링을 활용하였다. 또한 분류 기능을 테스트하기 위해 문서 집합을 선정하여, 도구의 동작을 확인하였다.

현재 본 도구는 TXT 형태의 문서들에 대해서만 색인화 가능하고, 분류하며 검색하는 기능만 가지고 있다. 향후에는 보다 더 다양한 ESI 타입들을 고려하여 시스템에서 폭넓게 사용되고 있는 포맷들에 대한 문서 색인 기능을 추가할 예정이며, 분류 된 결과를 보다 더 사용자가 분석하기 쉬운 형태로 시각화하는 방법을 모색하고자 한다.

참고문헌

[1] 한국EMC 컨설팅, CEO e-디스커버리를 고민하다, 전자신문사, 2011

[2] 김영수, 신상욱, 홍도원, “ESI 관점에서의 e-Discovery”, 정보통신산업진흥원 주간 기술동향, 2010

[3] EDRM Framework Guides, available from: http://edrm.net/resources/guides/edrm-framework-guides; 2010

[4] Apache Lucene, http://lucene.apache.org/

[5] Apache Mahout, http://mahout.apache.org/

[6] AccessData eDiscovery, http://accessdata.com/

[7] Guidance Software, EnCase Enterprise, http://www.guidancesoftware.com

[8] Nuix, http://www.nuix.com/

[9] Clearwell, http://www.clearwellsystems.com/

[10] Fabrizio Sebastiani, “Machine Learning in Automated Text Categorization”, Journal ACM Computing Surveys, Volume 34 Issue 1, March 2002