기업의 전자증거개시 대응을 위한 예측 부호화(Predictive Coding) 도구 적용 방안*

A Study on Application of Predictive Coding Tool for Enterprise E-Discovery

유준상 (Jun Sang Yu)**
임진희 (Jin Hee Yim)***

초 록

해외에 진출한 국내기업의 소송 사례가 증가하면서 기업들의 전자증거개시제도의 대응에 대한 요구가 증가하고 있다. 영미법에서 유래된 제도인 전자증거개시제도는 절차 진행과정에서 여러 곳에 산재해 있는 전자적 정보들을 중 제한된 시간 내에 소송과 관련된 전자적 정보들을 찾아 증거자료로 검토하여 제출하는 제도이다. 이는 하루에도 수많은 전자기록이 생산되는 국내기업들의 기록관리가 잘 이루어지지 않고 있는 현실에서 제한된 시간 이내에 증거자료를 추리고 검토하여 제출하는 것은 쉽지 않은 일이다. 검토대상을 줄이고 검토과정을 효율적으로 진행하는 것은 소송에서 승소를 위한 가장 중요한 과제 중 하나이다. Predictive Coding은 전자증거개시 검토 과정에서 사용되는 도구로써 기계학습을 이용하여 기업들이 보유하고 있는 전자적 정보들의 검토를 도와주는 도구이다. Predictive Coding이 기존의 검색도구보다 효율성이 높고 잠재적으로 소송과 관련된 전자적 정보를 추려내는데 강점이 있다고 판단된다. 기업의 효율적인 검색도구의 선택과 지속적인 기록관리를 통해 검토비용의 시간적, 비용적 절감을 꾀할 수 있을 것으로 예상된다. 따라서 기업은 전자증거개시 제도에 대응하기 위해서 시간과 비용적 측면을 고려한 전문적인 Predictive Coding 솔루션의 도입과 기업 기록관리를 통해 가장 효과적인 방법을 모색해야 할 것이다.

ABSTRACT

As the domestic companies which have made inroads into foreign markets have more lawsuits, these companies' demands for responding to E-Discovery are also increasing. E-Discovery, derived from Anglo-American law, is the system to find electronic evidences related to lawsuits among scattered electronic data within limited time, to review them as evidences, and to submit them. It is not difficult to find, select, review, and submit evidences within limited time given the reality that the domestic companies do not manage their records even though lots of electronic records are produced everyday. To reduce items to be reviewed and proceed the process efficiently is one of the most important tasks to win a lawsuit. The *Predictive Coding* is a computer assisted review instrument used in reviewing process of E-Discovery, which is to help companies review their own electronic data using mechanical learning. *Predictive Coding* is more efficient than the previous computer assister review tools and has a merit to select electronic data related to lawsuit. Through companies' selection of efficient computer assisted review instrument and continuous records management, it is expected that time and cost for reviewing will be saved. Therefore, in for companies to respond to E-Discovery, it is required to seek the most effective method through introduction of the professional Predictive Coding solution and Business records management with consideration of time and cost.

키워드: 전자증거개시, 기계학습, 검색도구, 기업기록, 기록관리, 예측 부호화
E-Discovery, electronic discovery, machine learning, search tool, business record, predictive coding

---

* 본 연구는 명지대학교 기록정보과학전문대학원 석사학위논문을 축약한 것임.
** 명지대학교 기록정보과학전문대학원 기록관리전공 석사(yudrake@gmail.com) (제1저자)
*** 명지대학교 기록정보과학전문대학원 부교수, '인간과기억아카이브'(hmarchives.com) 아키비스트 (yimjhkr@mju.ac.kr) (교신저자)
■ 논문접수일자 : 2016년 11월 20일 ■ 최초심사일자 : 2016년 12월 2일 ■ 게재확정일자 : 2016년 12월 10일
■ 정보관리학회지, 33(4), 125-157, 2016. 〔http://dx.doi.org/10.3743/KOSIM.2016.33.4.125〕

1. 머리말

1.1 연구 배경 및 목적

영미법에서 유래된 증거개시(Discovery)제도는 재판이 개시되기 전 당사자 서로가 가진 증거와 서류를 상호 공개를 통해 쟁점을 정리하고 명확히 하는 절차를 통칭하는 개념으로, 민사소송이나 형사소송에 있어서 법원의 개입 없이 소송 당사자 간에 서로의 요청에 의해 소송과 관련된 정보를 공개하는 법 제도이다.

컴퓨터 기술의 발전으로 인해 기록물의 체계가 종이기록에서 전자기록으로 변함에 따라 종이로 생산된 증거들이 대부분을 이루었던 증거개시제도도 전자기록을 증거로 제출하게 되었다. 전자증거개시제도 상에서 증거로 제출되는 되는 전자기록은 '전자적 정보(Electronic Stored Information, ESI)[1]'라고 불린다. 기존의 종이기록과 다른 특성을 가진 전자적 정보를 다루기 위한 전자증거개시(Electronic Discovery, E-Discovery)제도가 제정되었다. 전자증거개시제도는 북미권을 중심으로 활발히 시행 중이며, 국내에도 도입이 논의된 바 있다.

현재 전 세계 전자증거개시 소송의 80% 이상이 미국에서 진행되고 있다. 과거에는 미국 내에서만 중요시 되었지만, 국가 간 경계를 허무는 글로벌 기업들의 증가로 인해 미국 외 국가에서도 미국 시장에서의 소송을 위해 전자증거개시를 준비하고 있다. 또 이미 해외에 진출한 국내기업들의 전자증거개시를 통한 소송 사례가 잇따르고 있는 것이 현실이다. 해외에 진출한 국내기업들도 소송에 대비한 전자증거개시제도의 준비가 필요하다.

전자증거개시 소송 사례들을 살펴본 결과 전자증거개시 과정 중 시간적, 비용적 요소의 가장 많은 비용을 차지하는 단계는 검토단계(Review)인 것으로 나타났다. 따라서 전자증거개시 중 방대한 양의 전자적 정보를 인간의 수동적인 검토보다 빠르고 정확한 검토를 도와주는 도구인 검색도구 선택의 중요성이 높아지고 있다. 이에 해외 유수기업들은 검토단계의 효율성 제고를 위해 Predictive Coding[2]으로 대표되는 지능형 검색도구를 도입하는 추세이다.

이 연구의 목적은 전자증거개시에 대응해야 하는 기업들이 전자기록을 효과적으로 검토하기 위해 대표적인 기계학습 검색도구인 Predictive Coding을 살펴보고, Predictive Coding 도구의 효과성을 높이기 위해 기업에서 전자기록 관리 시 고려해야 할 점에 대해서 논의할 것이다.

1.2 선행연구

전자증거개시에 관한 선행연구들을 기록학적, 법학적, 컴퓨터 공학적 관점에서 살펴보았다.

1) 전자적 정보(Electronic Stored Information)는 컴퓨터 하드웨어나 소프트웨어에 사용하기 위해 디지털 형태로 생성, 조작, 통신, 저장 및 사용되는 정보이다. 전자증거개시에서 취급하는 전자적 정보는 이메일과 웹페이지, 워드프로그램으로 작성된 문서, 오디오 파일, 비디오 파일, 컴퓨터의 데이터베이스, 그림 파일 등 컴퓨터에 저장된 모든 파일이 포함된다. 컴퓨터 외에도 서버나 랩탑 컴퓨터, 이동전화기, 하드 디스크 드라이브, 플래시 디스크, MP3 플레이어 등 저장장치에 저장된 모든 정보도 포함되며, 전자적 정보는 빠른 기술 발전과 변화에 대응하기 위하여 향후 개념 확장이 가능한 개방형 개념이다.

2) Predictive Coding은 제목에서처럼 '예측 부호화'라고도 쓰이나 본문에서는 기법이 아닌 기법이 적용되는 도구들을 통칭하는 말로 외국어 표기를 사용함.

현재 전자증거개시 분야는 기록학계에서 아직 생소한 분야로 관련 연구가 많이 부족한 실정이다. 전자증거개시 관련 논문은 대부분 전자증거개시제도와 절차를 다루는 법학계와 전자증거개시를 위한 기술들을 다루고 있는 컴퓨터 공학계에 분포해 있다.

김승범(2015)은 기록학 관점에서의 전자증거개시제도 초기 연구의 일환으로 전자증거개시제도의 개요와 한국을 포함한 각국 제도의 연혁을 살펴보고 전자적 정보 및 EDRM 등의 핵심 개념을 분석한 뒤 국내 기업과 기록관리자들이 전자증거개시 환경에 대응하기 위한 준비 요건을 밝히고, 기록관리표준에 포함되어야 할 필수 사항에 대한 시사점을 도출하였다. 이 연구는 기록학 관점에서 전자증거개시제도에 대하여 쓰인 첫 연구라는데 그 의의가 있다.

김일아(2016)는 전자증거개시에 능동적으로 대응하는 미국 기업 기록관리의 실무 현황 및 사례 연구를 바탕으로 기록관리 연구 현황을 수집 및 분석하여, 전자증거개시에서 기록관리가 필수적인 요건이라는 시사점을 도출하였다. 이 연구는 전자증거개시에 대응할 수 있는 핵심적인 기록관리 전략을 제시하는데 목적을 두고 있다.

탁희성(2011)은 우리나라는 현재 2007년 형사소송법 개정을 통해 증거개시제도가 이미 도입되어 있지만, 아직까지 전자적 정보의 특성을 반영한 관련 제도가 정착되지 않았음을 지적하고 있다.

이에 따라 전자증거개시에 대한 정의와 기존 증거개시와 차이점을 비교하고 전자증거개시의 대상으로 전자적 정보의 개념과 특성을 정리하였다. 다음으로 주요 외국의 전자증거개시 관련 입법 및 기술동향을 소개하면서, 미국에서는 기존 증거개시제도에서 증거개시의 대상을 종이문서에서 전자적 정보로 확대됨에 따라 전자적 정보의 특성에 맞는 전자증거개시 절차를 이미 적용하고 있다고 밝히고 있다. 따라서 우리나라에서도 전자적 정보로 이루어진 증거자료의 증가에 따른 사회적 변화에 부합할 수 있는 증거개시절차에 대한 입법적 개선방안을 제시하고, 전자적 정보의 특성을 고려한 증거개시 방식을 제시하고 있다.

김도훈(2014)은 미국 전자증거개시절차상 전자적 정보의 특성에 기인한 문제와 이에 따른 컴퓨터 지원 검토 도구들의 변화에 관해 정리하였다. 다음으로 가장 일반적으로 활용되고 있는 키워드 검색과 그 한계 그리고 이에 대한 법원의 평가를 검토하고, 새로운 대안으로 제시되고 있는 Predictive Coding에 대해 간략히 서술하였다. 마지막으로 새로운 기술에 대한 법학적 관점에서의 Predictive Coding의 시사점을 도출하고 있다.

김종호(2015)는 한국 기업과 다국적 기업 사이의 국제특허소송건수가 증가하고 있지만, 한국 기업들의 충분한 준비가 되어있지 않아 패소하는 경우가 다수 발생하고 있다고 지적하고 있다. 따라서 증거개시 및 전자증거개시 제도에 효과적으로 대처하며, 정보를 관리할 수 있는 노하우와 전략을 구체적 사례를 들어 증거개시제도를 도입하여 국내 전자소송시스템의 고도화를 이뤄야한다고 제시하고 있다.

위의 세 가지 연구는 법학적 관점에서 본 전자증거개시에 대한 연구들이다. 이밖에도 안정혜(2010), 전복만과 박지훈(2012) 등 법학계의 연구들은 제도에 대한 필요성을 언급하면서 이에 맞는 국내의 제도적 개선이나 기술적 적용이 필요하다고 제안하고 있다. 그러나 제도적인 측

면에서만 국내의 전자증거개시 도입을 강조하고 기업이 실질적으로 어떻게 대처해야 한다는 구체적인 방안이 없어 그 한계점이 있다.

김영수와 홍도원(2011a, 2011b)은 2006년 이후 전자증거개시의 대상이 되는 증거물의 범위에 전자적 정보를 포함하게 됨을 근거로, 전자증거개시에 대응하기 위한 시스템의 도입의 필요성을 강조하였다. 다음으로 EDRM이 제시한 전자증거개시 절차를 기반으로 한 세부 기능을 분석하고, 다양한 기능 중에서 소송 관련 전자적 정보의 양을 줄이는 컬링(Culling) 성능 향상을 위한 기술들을 간략히 제시하고 있다.

이태림과 신상욱(2012)은 미국의 전자증거개시 제도와 이를 적용함에 있어서 표준화된 업무수행 절차 확립을 위해 지금까지도 활발한 연구가 진행되고 있는 EDRM과 Sedona Conference 프로젝트 등에 대한 분석을 바탕으로 전자증거개시의 절차별 필수 업무 사항을 도출한다. 다음에는 전자증거개시의 시간 및 비용절감을 위한 대체 기술로 기계 학습과, 대용량 데이터를 위한 분산 처리 기법을 소개하고, 전자증거개시 절차상에서 활용방안을 제시하고 있다.

이밖에도 채은선(2008), 천우성과 박대우(2011) 등 컴퓨터공학계의 연구들은 디지털 포렌식적 관점에서 본 전자증거개시제도의 과정 중 전자적 정보를 어떻게 수집, 관리, 제출하는지에 대한 기술적인 내용들과 시스템 구축관련 연구들이 대부분을 이루고 있어 그 한계점이 있다.

1.3 연구 범위 및 방법

본 연구에서는 문헌 연구를 통해 기업의 전자증거개시 대응을 위한 Predictive Coding 도구 적용 방안을 모색한다.

첫째, 전자증거개시를 진행하는데 중요한 참조모델과 사례인 Electronic Discovery Reference Model(EDRM)과 Electronic Discovery Best Practice(EDBP)에 대해 공식 홈페이지와 관련 문헌 등을 통해 조사한 다음 핵심 단계인 검토 단계에 대해 자세히 살펴본다.

둘째, 전자증거개시에 대응해야 하는 한국 기업들의 입장에서 필요한 검색도구의 하나로서 미국의 아키비스트 저널 ARMA에서 반복적으로 언급되고 있는 Predictive Coding 솔루션으로 대표되는 기계학습 검색도구의 원리를 살펴본다. 이어 기계학습에 따른 검색 원리를 키워드 검색, 전문검색 등 다른 검색도구의 원리와 어떻게 다른지 기술적 처리과정을 살펴본다.

셋째, Predictive Coding을 이용하여 소송을 진행했던 사례를 살펴본다. 이를 바탕으로 Predictive Coding을 이용한 전자증거개시 대응 전략과 기록관리를 통한 전자증거개시 대응 전략을 제시하였다.

2. 전자증거개시 프로세스의 검토단계 효율화 필요성

2.1 전자증거개시 참조모형의 검토단계

EDRM(Electronic Discovery Reference Model)이란 전자증거개시제도를 통하여 소송을 진행시 법정에 제출되는 전자적 정보의 무결성을 보장하기 위해 개발된 전자증거개시 표준 프로토콜이다. 프로토콜이란, '법이 정한 요건과 절차를 바탕으로 구체적인 절차를 이행하면 법이 정

한 요건과 절차를 가장 잘 준수할 수 있다는 경험적인 모델'로 이해하면 된다.

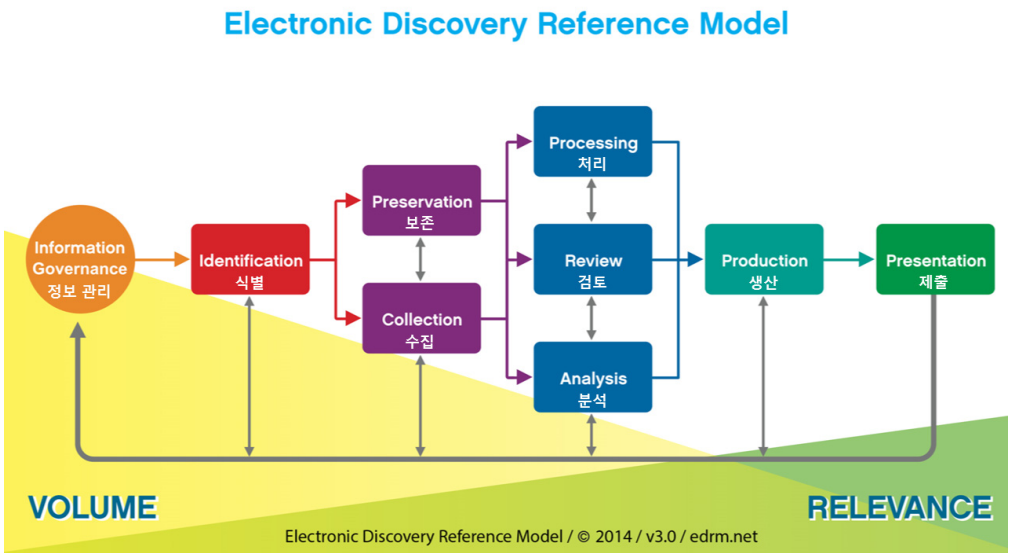
EDRM은 전자증거개시 및 정보 거버넌스(Information Governance)를 개선하고 표준과 가이드라인의 부족을 해결하기 위해 2005년 5월 출범했다. 이후 EDRM은 270개 이상의 조직들로 구성되어 있는데, 이중 전자증거개시 관련 174개의 서비스 및 소프트웨어 제공 업체들과 24개의 조직과 기업이 참여하여 지속적으로 개선 보완되고 있다. EDRM은 전자증거개시 소프트웨어와 서비스를 개발하고, 선택 및 평가하기 위한 일반적이고 유연하며 확장 가능한 프레임워크를 제공하고 있다. 이는 여러 관련 조직들의 협의 하에 개발되었으므로 공인된 전자증거개시에 대한 일반적인 표준으로 활용되고 있다.

EDRM은 〈그림 1〉과 같이 정보관리(Information Governance)부터 제출(Presentation)까지 총 아홉 단계로 구성되어 있으며 각 내용은 다음과 같다.

첫째, 정보관리는 소송이 시작되기 전에 조직에서 정보관리 정책에 의해 전자적 정보를 관리한다. 대부분의 정보가 전자적으로 저장되고 생산되는 환경에서 적절한 전자적 정보의 관리는 모든 조직에게 필수적이다. 정보 관리는 소송 목적으로 존재하지 않더라도 전자적 정보 요청에 대해 효과적인 대응을 위한 전제조건이라 할 수 있다.

둘째, 식별은 소송이 예상되는 시점에서 수집해야 할 전자적 정보의 범위를 결정하고 전자적 정보의 데이터 지도를 작성한다. 이를 통해 증거개시에 있어서 활용해야 하는 전자적 정보의 범위를 결정하게 된다.

셋째, 보존은 식별과정을 거친 소송과 관련 있는 전자적 정보를 우연히 또는 고의로 삭제 및 변경되지 않도록 보존과정을 거쳐 해당 전자



〈그림 1〉 EDRM 프로세스 모델

적 정보를 수집할 수 있도록 준비하는 과정이다.

넷째, 수집은 하드 디스크, 외부 저장장치, 네트워크 스토리지 등 기업이나 조직의 데이터 시스템으로부터 식별하여 보존해 오던 소송 관련 전자적 정보들을 검토하고 분석하기 위해 직접 추출하게 된다. 증거원본에 대한 무결성 보장을 위해 디지털 포렌식 관점에서 활용 가능한 데이터 수집기법들이 주로 사용된다.

다섯째, 처리는 전자적 정보의 크기를 줄이고 분석과 검토를 위해 적절하게 변환하는 것을 목적으로 한다. 기존의 식별, 보존, 수집단계를 거친 데이터들을 대상으로 메타데이터를 기록하거나 소송에 직접적으로 활용할 수 있고 검토할 가치가 있는 데이터를 선택함으로써 '어떠한 데이터가 각 프로젝트에서 검토와 제출의 대상에 적절한 것인지 정확하게 식별하는 것'이다.

여섯째, 검토는 전자증거개시과정에서 소송에 대응할 문서를 식별하고, 비밀유지 특권이 있는 문서를 보호하기 위한 중요한 절차이다. 효율적이고 비용대비 효과적인 방식으로 문서들을 논리적 집합으로 조직화하여 문서 내용의 이해를 돕는 것을 목적으로 한다. 검토단계에서 조직의 법무담당자는 소송에 관련된 사실에 더 정확하게 접근할 수 있게 되고, 법률 전략을 실행할 수 있게 되며, 수집된 정보들의 형태를 기반으로 전략을 발전시킬 수 있게 된다. 검토의 범위를 정하고, 감독과 검토자를 관리할 절차를 설정하며, 적절한 전자증거개시 서비스 업체 및 소프트웨어, 기반을 선택하는 순서대로 진행하게 된다.

일곱째, 분석은 EDRM의 한 단계가 아니라 전체에 걸쳐 진행된다. 법무팀에서 검증된 데이터를 기반으로 신뢰할 수 있는 방법을 통해 전략과 범위에 대한 결정을 목적으로 한다. 크게 '내용 및 메타데이터 분석(Content/Metadata Analysis)'과 '절차분석(Process Analysis)'으로 구분된다.

여덟째, 생산은 전자적 정보를 효율적이고 사용가능한 포맷으로 준비하고 생산하기 위해여 비용, 위험, 오류를 감소시키고, 협의된 사양과 일정을 준수함을 목적으로 한다. 소송당사자들은 전자증거개시 초기단계에서 협의하여 생산 포맷을 결정하는데, 이를 위해 전자적 정보의 성격이나 유형 같은 요소들에 대한 이해도가 필요하다.

아홉째, 제출은 최종 산출한 전자적 정보를 증거로써 제출하기 위한 절차로, 정보를 제공받게 되는 당사자의 특성을 고려하여 증거 공개방법과 포맷 등을 결정하고 소송당사자들을 설득하는 과정이 포함된다.

2.2 전자증거개시 선진실무의 검토단계

EDBP(E-Discovery Best Practice)는 로펌과 기업의 법무 담당 부서에서 사용하기 위한 우수 사례의 모델을 제공하기 위해 만들어졌다. EDRM이 제시된 이후 전자증거개시를 이용한 재판이 여러 차례 진행되었다. 이에 따라 전자적 정보에 대한 법률적 해석의 변화와 기술적인 특징의 세부적인 이해의 필요성이 지속적으로 대두되면서, 법무전문가들을 중심으로 전자증거개시 절차에 EDRM보다 더 법률적인 내용이 반영된 구체적인 표준이 필요하다는 여론이 형성되었다. 따라서 EDRM이 추구하는 전자증거개시의 포괄적인 법률적, 기술적 가이드라인에 대한 제시와는 달리 법률적인 분야에 더 세부

적인 절차와 의미를 제시하기 위한 EDBP라는 모델이 탄생하였다.

EDBP는 연방민사소송규칙에 명시된 법률조항을 중심으로 모델을 구성하였고, 모델에서 제시하는 각각의 절차에 대한 절차에 대한 법률적 해석과 판례를 근거로 세부적인 방향을 제시한다. 하지만 EDBP는 오직 교육 자료로 활용하기를 권장하면서 EDBP 역시 EDRM 모델을 기초로 아이디어를 도출한 참조 모델임을 밝히고 있다.

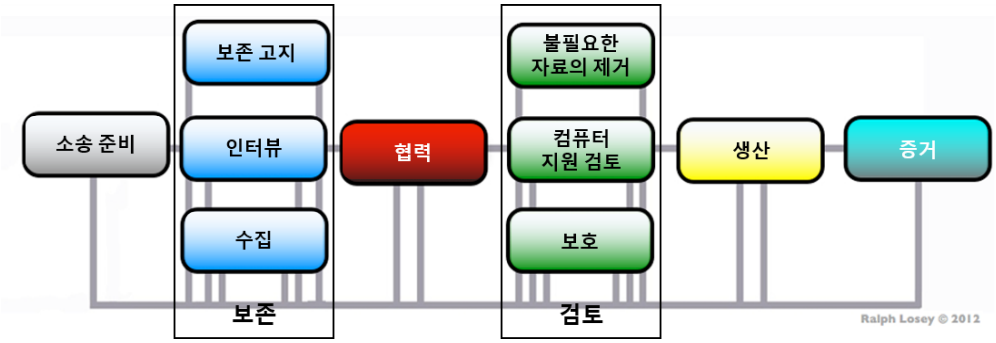
EDBP는 〈그림 2〉와 같이 소송준비단계(Litigation Readiness)부터 증거(Evidence)단계까지 총 여섯 단계로 이루어져 있으며, 각 내용은 다음과 같다.

첫째, 소송준비는 EDBP의 다음 절차인 보존고지(Hold Notices), 인터뷰(Interview), 수집(Collections), 협력(Cooperation)의 네 가지 단계를 준비하기 위해 설계되었다. 소송준비 이후에 이어지는 네 가지 단계는 거의 동시에 일어나기 때문에 이전 단계인 소송 준비의 결과에 따라 가능성 여부나 성과의 정도가 결정된다.

둘째, 보존은 보존과 관련된 법률 조치로 보존고지와 인터뷰, 수집이 있다. 법률 조치는 거의 동시에 이루어지는데 이는 관련자와의 인터뷰에서 나오는 인터뷰에서 나오는 각종 정보에 대한 분석을 바탕으로 해당 조직이나 기관이 보유하고 있는 전자적 정보 중 소송과 관련되어 있거나 장차 관련될 수 있는 것들을 식별하는 절차를 통해서 이루어진다.

셋째, 협력은 기존의 보존·수집한 전자적 정보를 소송에 활용할 수 있도록 한다. 협력은 소송을 준비하는 내부적인 의미의 협력을 말하는 것이 아니라 소송 상대방과의 협력을 의미한다. 협력 절차에서 논의가 되어야 할 내용으로 소송과 관련 있는 전자적 정보에 대한 연관성과 청취, 생산에 대한 요구, 정보의 생산 형태, 협력 대상이 아닌 문서의 기준 등을 제시한다.

넷째, 검토는 전자증거개시 과정 진행 중 소송과 관련하여 대상이 될 자료를 선별하고, 확인하는 과정이다. 일반적으로 전자증거개시 전체 소송비용의 60~80%의 비중을 차지한다(E-Discovery Team, 2013). EDBP는 검토 단계에서 이루어지는 절차를 불필요한 자료의 제거(Culling)와 컴퓨터 지원 검토(Computer Assisted Review, CAR), 보호(Protections) 세 가지로 나누어 제시하고 있다. 다만, 보호에 대한 절차



〈그림 2〉 EDBP 프로세스 모델

는 명시되어 있지 않다. 나머지 절차를 살펴보면 다음과 같다.

불필요한 자료의 제거는 법률적 검색과 검토 단계에서 파일의 제거를 의미한다. 불필요한 자료를 제거하는 방법은 다양하지만 가장 우선하여 고려하고, 효과적인 것으로 제시되는 것은 소송 관련 인원에 의해서 이루어지는 방법이다. 이외에도 소송과 관련하여 정보를 생산한 기간과 보존한 기간 등 기간을 설정하여 불필요한 정보를 제거하기도 하고, 동일한 파일을 구분하여 중복되는 파일을 제거하기도 하며, 키워드 검색을 통해 소송의 증거와는 관련이 적은 파일을 구분하여 제거하기도 한다. 이러한 절차는 크게 어려워 보이지 않을 수도 있지만 불필요한 자료는 수집이 많이 되면 될수록 검토 비용 상승의 결과를 가져오기 때문에 상당히 중요한 작업이라고 볼 수 있다.

컴퓨터 지원 검토는 전자적 정보의 종류와 양이 기하급수적으로 늘어남에 따라 사람에 의해서 직접 검토되어지는 방식의 한계를 극복하고자 도입된 방법이다. EDBP는 컴퓨터 지원 검토를 전자적 정보의 검색과 문서의 코딩, 변호사가 최종검토를 할 소송과 가장 밀접하게 관련된 전자적 정보를 선별하고, 불필요한 데이터의 제거가 가능한 소프트웨어를 사용하여 검토하는 것으로 설명한다. 또한 변호사에 의해서 직접 전 과정의 검토가 이루어지는 방식은 컴퓨터 지원 검토를 사용하는 것보다 비효율적이고, 비용이 더 많이 든다고 제시한다.

다섯째, 생산은 EDBP에 명확한 절차가 제시되어 있지 않다. 다만, 전자문서생산에 대한 세도나 원칙과 미국캘리포니아북부 지방법원이 제정한 전자적 정보 체크리스트를 참고하도록 명시되어 있다.

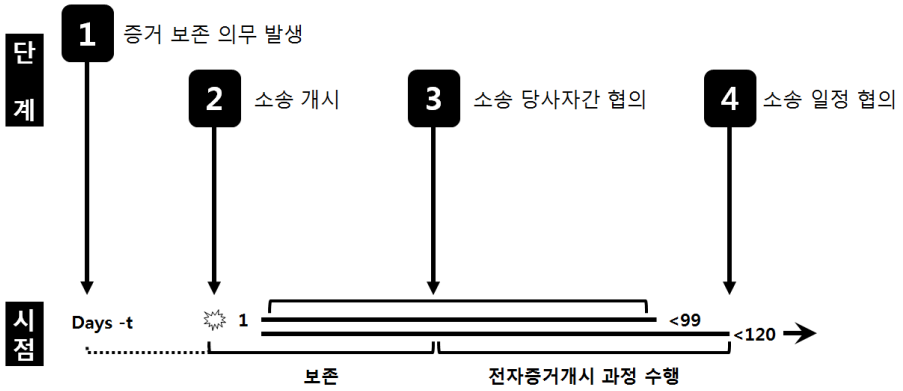
여섯째, 증거 역시 EDBP에 명확한 절차가 제시되어 있지 않다. 증거를 최종 제출하기 위한 프레젠테이션을 위해 전자적 정보를 찾는 작업은 전자증거개시 서비스 제공 업체와 컨설턴트가 하는 일이다. 따라서 변호사가 EDBP 전 프로세스에서 충분히 감시하고 참여하여 최종 결과물에 대해 신뢰성을 높일 수 있도록 해야 한다.

2.3 검토단계의 효율화 필요성

전자증거개시 과정에서 승소를 위한 핵심 쟁점은 제한된 시간 내에 최소한의 비용으로 증거자료를 제출하는 것이다. 제한된 시간 이내에 승소하기 위한 충분한 증거를 확보하지 못하거나 막대한 소송비용을 지출하고도 패소하여 판매 및 소득의 근원인 원천 기술과 상표권에 대한 권리 상실을 한다면 기업에게 치명적인 타격이 될 수 있다. 따라서 승소와 관련된 중요 쟁점을 전자증거개시절차 진행에 따른 시간적 요소와 비용적 요소로 나눌 수 있다.

첫째, 시간적 요소는 실제 소송 절차 흐름에 따라 중요도가 달라진다. 미국 연방민사소송규칙 상의 일반적인 소송 절차에 있어서 법에 명시된 주요 마감시점(Dead Line)은 〈그림 3〉과 같이 시간에 흐름에 따라 네 단계로 나누어진다.

〈표 1〉에 따르면, 실제 소송 관련 정보를 수집하고 증거로 생산하여 제출하는데 소요할 수 있는 기간은 소송 발생일로부터 120일 이내이다. 소송 당사자 간의 주요 쟁점과 증거 제출 양식 협의, 소송 일정을 확인하는 등 많은 업무들이 포함되어 있으므로, 실제 소송과 관련된 전



〈그림 3〉 전자증거개시 절차와 마감시점(Dead Line)

출처: Volinino와 Redpath(2010)의 그림 재구성.

〈표 1〉 마감시점 단계별 세부 내용

단계	시점	내용
1단계	Day -t	소송 발생 이전 기업 내 정보들에 대해 보존의 의무를 수행하는 단계
2단계	Day 1	소장 접수로 인한 소송 발생
3단계	By Day 99	소송 당사자 간 협의를 통해 공개되어야 할 증거 범위와 형식, 시간 및 비용 등을 결정하는 단계
4단계	By Day 120	소송에 대한 재판 스케쥴을 결정하는 단계

출처: 이태림과 신상욱(2012)의 내용 재구성.

자적 정보의 검색 및 증거로서 적합성 검토 등에 투자할 수 있는 시간은 길지 않음을 알 수 있다(이태림, 신상욱, 2012).

법이 정한 기한 내에 증거를 마련하지 못한다면 패소는 자명한 일이므로 업무 수행에 있어서 시간은 아주 중요한 요소임을 의미한다. 또한 소송은 언제든지 발생 가능하며 기업이 다루는 정보의 양이 나날이 대규모화 되고 있는 상황을 고려해보면, 소송과 상관없이 평상시에도 업무와 연관된 정보들을 체계적으로 분류하고 관리하여, 정보 검색에 소요되는 시간을 줄임으로써 신속한 대응을 가능하게 해야 한다.

둘째, 전자적 정보의 양이 기하급수적으로 증가함에 따라 비용적 요소의 중요도가 상승하고 있다. 2006년 미국에서 전자증거개시제도를 의무화하여 소송 당사자들이 전자적 정보를 증거자료로 제출하기 시작했다. 그런데 전자증거개시 한 건을 수행하는데 평균 150만 달러가 소요되는 등 많은 비용 문제가 발생하고 있다.

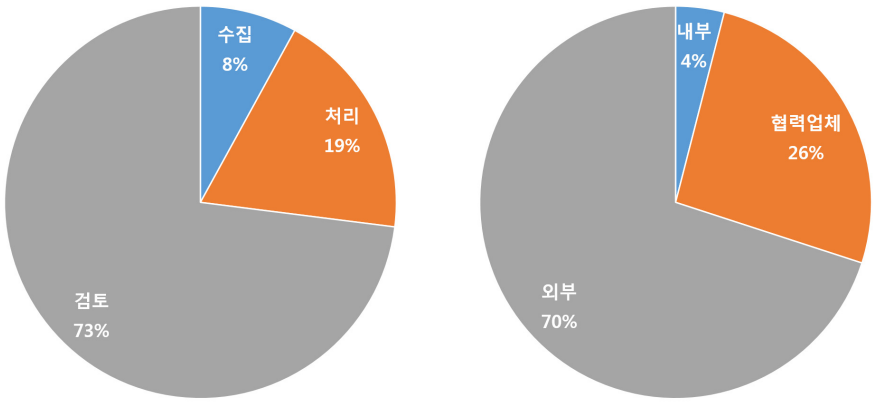
2008년 가트너(Gartner, Inc.)의 보고서에 따르면 미국의 많은 로펌에서 주니어 변호사의 시급이 $200에서 시작하는 것을 근거로, 법률 전문가의 전자적 정보의 검토비용이 1GB당 $18,750로 비용을 추정하였다. 소송 증거를 제출하기 전 검토단계에서 변호사들의 전자적 정보를 직접 수작업으로 검토를 해야 하기 때문에

많은 비용이 예상 되는 것으로 보인다. 또, 전자증거개시 한 건당 평균 150만 달러의 소송비용이 소모되는 현실에서 기업 콘텐츠 아카이빙 솔루션에 적용할 정책과 전략을 제대로 수립하지 못한 기업은 효과적인 콘텐츠 아카이빙 솔루션을 구축한 기업에 비해 전자증거개시 과정에서 30%가 넘는 비용을 더 지출하게 될 것이라고 밝혔다(Debra, 2008).

〈그림 4〉에 따르면, 2012년 Rand Corporation에서 전자정보개시프로젝트 중 비용이 가장 많이 소요되는 부분은 업무 절차적인 측면에서 수집(8%)이나 처리(19%)가 아니라 검토(73%) 단계에서 가장 많은 비용을 소모하고 있다고 발표했다(Pace & Zakaras, 2012). 전자증거개시 과정에서 검토과정은 가장 많은 비용 지출을 차지하고 있다. 이는 방대한 자료 중 '불필요한 자료의 제거 및 처리'를 수행하여 얻어진 결과를 대상으로 다시 관련성, 대응성, 면책특권 등에 대한 검토가 필요하고, 그와 같은 검토는 여전히 법률전문가가 수동적으로 진행하기 때문에 많은 비용이 소모되고 대부분의 비용을 차지하는 것으로 보인다.

출처별 수집 비용에서는 내부(4%), 협력 업체(26%), 외부(70%)의 수집 비용이 발생하는 것으로 나타났다. 이는 내부나 협력 업체의 전자적 정보는 상대적으로 외부에 산재해있는 전자적 정보 보다 수집에 훨씬 용이한 점에서 비용의 차이가 나타나는 것으로 판단된다. 기업기록관리의 측면에서 보았을 때, 기업 내부나 협력업체보다 외부의 전자적 정보를 어떻게 수집, 관리하는 문제가 비용적 측면에서 더 중요한 것으로 보인다.

앞서 전자증거개시 참조모형과 전자증거개시 선진실무를 살펴보고 가장 핵심적이며 자원소모가 큰 과정이 검토단계인 것으로 나타났다. 따라서 업무를 위해 매일 많은 양의 정보를 처리하고 생산하는 기업들은 검토단계에서의 효율적인 전자증거개시를 위해서 검토단계에서의



업무 절차 출처

〈그림 4〉 업무 절차 / 출처별 전자증거개시 소모비용 비율

출처: Pace와 Zakaras(2012)의 차트 재구성.

시간적·비용적 자원의 소모를 감소시킬 수 있는 새로운 검색도구의 필요성이 대두되고 있다.

3. 검토를 위한 검색도구 현황

3.1 검토를 위한 기존 검색도구의 종류

전자증거개시진행 과정 중 검토단계에서 시간과 비용의 소모를 줄이는 것이 주요 쟁점으로 부각되고 있다. 이를 위해 기업에서는 검토단계의 효율적인 검색도구를 선택하는 것이 중요하다.

2010년을 기점으로 기존의 수동적 검토에서 컴퓨터 지원 검토로 검토과정의 변화가 이루어지면서 검색 도구에 대한 신뢰성과 비용 절감 등의 효과에 대한 의문이 제기 되었다. 그러나 2009년 미국표준기술연구소(National Institute of Standards and Technology: NIST)가 후원하는 텍스트 검색 컨퍼런스(Text Retrieval Conference)가 기존의 수동적 검토와 컴퓨터 지원검토를 비교 분석한 결과 컴퓨터 지원 검토가 수동적 검토에 비해 비용이 적게 들면서 더 나은 결과를 제시할 수 있는 것으로 나타났다(김도훈, 2014).

검토단계에서 사용되는 검색도구들은 소송과 관련 있는 전자적 정보를 걸러내기 위한 일종의 거름망이라고 볼 수 있다. 소송과 관련 있는 전자적 정보를 추려 내어 법무관계자가 관련성, 대응성, 면책특권에 대해서 평가를 하게 된다. 검색도구에는 유형별, 알고리즘별로 다양한 종류의 검색도구들이 있으나 본문에서는 전자증거개시과정에서 주로 사용되는 키워드 검색(Keyword Search), 개념 검색(Concept Search), 디스커션 스레딩(Discussion Threading), 클러스터링(Clustering), 중복 식별(Near-duplicate Identification)에 대해서 살펴보겠다.

첫째, 키워드 검색은 가장 일반적인 검색방식으로 사용자에 의해 지정된 하나 이상의 단어를 포함하는 문서를 찾는 검색도구이다. 대표적으로 키워드검색은 색인 검색(Index Search)과 전문 검색(Full-text Search) 두 가지로 나뉜다. 색인 검색은 문서에서 중요한 키워드가 되는 단어들을 색인어로 등록한다. 색인(indexing)이란 개개의 전자적 정보의 특성을 표현하는 데이터 요소를 추출하여 각 전자적 정보를 표현하는 작업을 말하며, 색인 결과 추출된 데이터 요소를 색인어(index term) 또는 메타데이터(metadata)라고 부른다. 보통 웹 검색 엔진이나 온라인 데이터베이스 서비스에서 각 전자적 정보에 대한 색인 결과 색인 데이터베이스가 생산되며, 이 데이터베이스는 검색을 위한 도구로 사용된다. 전문 검색은 색인 검색과는 달리 자동색인 프로그램이 특정 키워드가 아닌 전자적 정보의 본문 전체 내용 중 적합한 키워드를 색인어로 추출하여 데이터베이스에 저장한다.

색인 검색과 전문검색은 색인어를 추출하는 주체가 누구인가에 대한 차이점에 의해 구분된다. 색인 검색은 생산자나 관리자가 색인어를 추출하기 때문에 사람의 주관적인 관점이 키워드에서 드러나게 된다. 추출하는 사람의 색인어에 대한 이해나 색인어 선정 당시의 관점에 의해 그 품질이 좌우된다. 전문 검색은 자동색인 프로그램에 의해 색인어가 추출되어 색인 검색에 비교적 객관적이지만 검색 알고리즘별로 검색성능이 달라 전자적 정보의 질의어 특성에 따라 색인 전략을 세우는 것이 중요하다.

둘째, 개념 검색은 개념적으로 유사한 성격을 가진 전자적 정보들에 대한 구조화되지 않은 텍스트를 검색하는데 사용되는 검색도구이다. 개념 검색은 대규모의 구조화되지 않은 텍스트와 디지털 컬렉션을 처리하는데 유용하고, 기존 키워드 검색의 한계점을 개선하기 위해 개발되었다. 전자증거개시에서 개념 검색은 키워드 검색보다 관련성이 높은 결과를 도출할 수 있는 안정적이고 효율적인 검색방법으로 인정되고 있다.

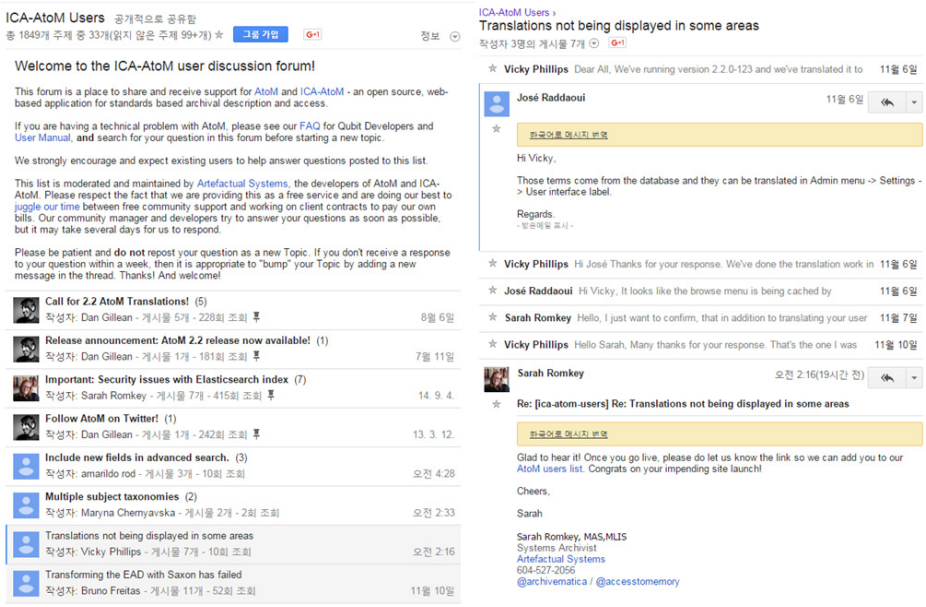
보통 인간의 언어를 처리하는 모든 컴퓨터 시스템의 주요 쟁점은 '다의어를 어떻게 처리할 것인가'에 관한 것이다. 예를 들어, 영어 단어 'fire'는 '연소(a combustion activity)', '해고(to terminate employment)', '발사(to launch)', '일어서게 하다(to excite)'의 네 가지 뜻을 의미한다. 영어권에서 가장 많이 사용되는 상위 200개의 다의어들 중 동사는 12개의 의미를 가지고 있고 명사는 8개의 의미를 가지고 있다. 또, 상위 2,000개 다의어들 중 동사는 8개의 의미, 명사는 5개 이상의 의미를 가지고 있다. 키워드 검색은 키워드의 동의어 또는 다의어 효과에 따라 이용자의 의도와는 달리 관계없는 검색결과를 제공하는 경우가 발생한다. 개념 검색은 단어의 실제 의미를 도출하기 위해 '어의 중의성 해소(Word-sense disambiguation, WSD)'를 사용하여 이러한 문제를 극복할 수 있다.

셋째, 디스커션 스레딩은 변호사들이 수동적으로 검토를 할 시에 업무상 주고받은 이메일, 메시지와 같은 전자적 정보들을 주제별로 묶어 시간의 흐름 순으로 나열하여 보여주는 뷰어 형태의 기술이다.

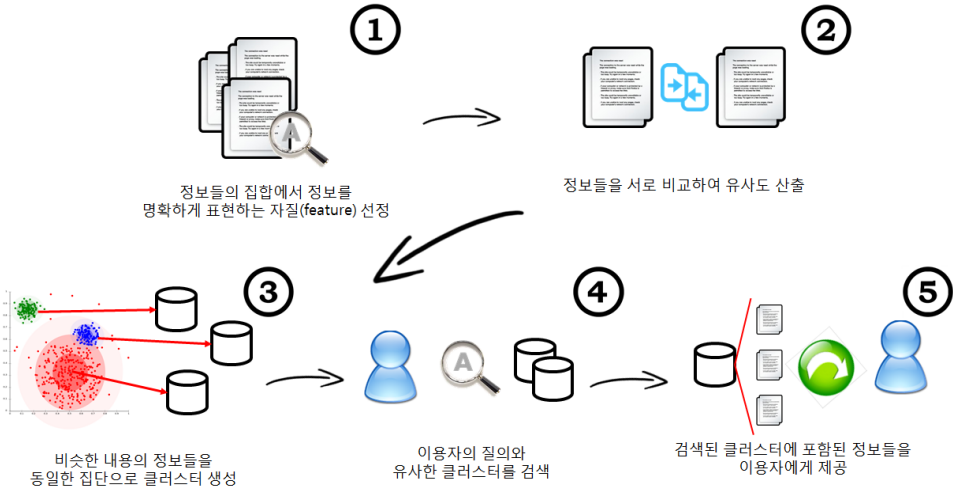
대화 스레딩(Conversation threading)이라고도 불리는 이 방식은 알고리즘을 이용하여 상호 동적으로 연결된 관련문서들의 전체적인 대화를 시간적 순서의 스레드로 정렬한다. 예를 들어, 송수신된 이메일을 하나의 스레드로 그룹화한다면 시간의 흐름 순이나 관련 있는 메시지의 묶음 등 다양한 방식으로 나타낼 수 있다. 이러한 표현 방법은 시간에 흐름에 따른 스레드 대화 내용을 한눈에 파악할 수 있다.

〈그림 5〉는 ICA-AtoM 사용자 토론 그룹의 스레드 뷰(Thread view)화면이다. 최상이 수준에서 여러 게시물과 함께 토론할 수 있고 제목과 날짜, 답변 등이 시간 순으로 정렬된 것을 볼 수 있다. 이와 같이 검토단계에서 최종적으로 검토를 하는 변호사들을 위해 업무상 주고받은 메시지나 이메일을 스레드 뷰 형식으로 시간의 순서와 대화 주제에 따라 정렬해서 보여준다. 변호사들은 이를 이용해 대화의 흐름을 쉽고 빠르게 파악 하는데 도움을 얻을 수 있다.

넷째, 클러스터링은 검토과정에서 대규모의 데이터 중 유사한 성격의 전자적 정보를 한데 묶어 정렬하는 데이터 처리방식으로 수집된 전자적 정보를 처리할 때 주로 쓰인다. 클러스터링의 사전적 의미는 '유사성과 같은 어떤 개념을 바탕으로 데이터를 몇 개의 집단으로 분류하는 방법'이다. 검색도구로서 클러스터링은 각 전자적 정보를 표현하는 자질(feature)들을 비교하여 정보 간 유사성을 측정한 다음 비슷한 내용의 정보들을 동일한 집단에 속하도록 군집화하는 기법이다. 클러스터링 결과 생성된 유사한 정보들의 집단, 즉 정보 클러스터들은 각 클러스터를 대표하는 클러스터 센트로이드(Cluster centroid)를 갖게 되며, 검색 시 이용자의 질의와 클러스터 센트로이드가 비교되어 질의에 가장 적합한 클러스터가 검색된다.



〈그림 5〉 디스커션 스레딩 - 사용자 토론 그룹의 스레드 뷰

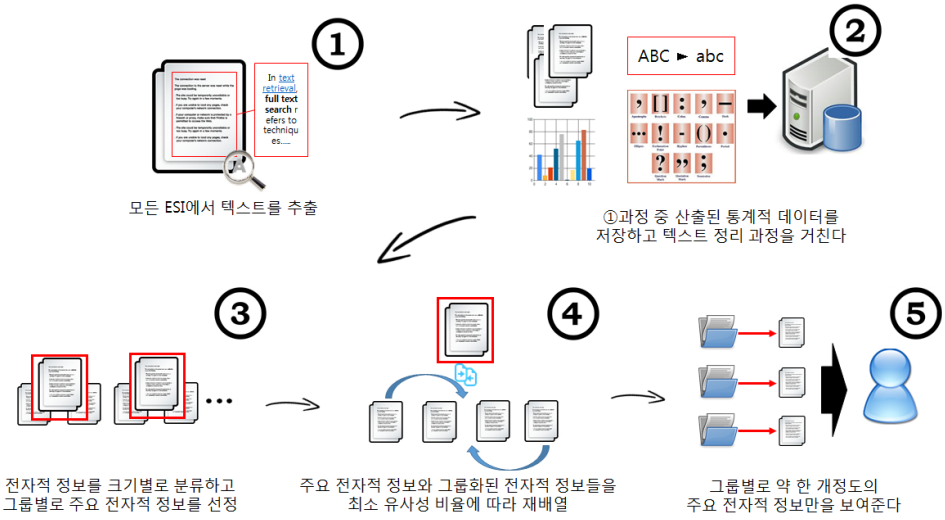


〈그림 6〉 클러스터링 과정

다섯째, 중복 식별 역시 클러스터링과 마찬가지로 대규모 데이터를 처리하는 방식이다. 보통 기업의 업무 과정 중에 생성되는 전자적 정보들은 유사한 성격의 전자적 정보가 재생산 되거나 여러 경로에 분산되어 저장되는 경우가 종종 발생한다. 예를 들어, 고용 분쟁에 관한 기안서가 초안부터 최종안까지 버전별로 존재할 경우 모든 버전의 기안서들은 유사한 내용을 가지고 있

을 것이다. 이는 근처 중복(Near Duplication)이라고 불린다. 검토 과정에서 근처 중복되는 전자적 정보를 모두 검토하는 것은 불필요한 작업이기 때문에 수동적 검토에 앞서 중복되는 전자적 정보를 걸러낸다. 대표적인 중복 식별 방식으로는 텍스트 중복 식별(Textual near duplicate identification) 방식이 있다.

텍스트 중복 식별은 소송과 관련된 문서를 검토하는데 필요한 시간을 줄이고 유사한 문서 간 차이점을 찾는데 강점이 있다. 텍스트 중복 식별은 〈그림 7〉과 같이 진행된다. ① 모든 전자적 정보에서 텍스트를 추출한다. ② 추출과정에서 텍스트 출현 빈도수와 같은 모든 통계적인 데이터를 같이 저장한다. 이 과정 중 모든 텍스트는 소문자로 변환된다. 이는 단어와 문장의 경계를 식별하기 위해 제외되고 띄어쓰기와 구두점과 같은 모든 문장부호도 무시하게 된다. ③ 전자적 정보를 데이터의 크기별로 분류하고 그룹별로 주요 전자적 정보(Principal ESI)를 선정한다. 주요 전자적 정보는 그룹에서 가장 크기가 크며 다른 전자적 정보와 중복 여부를 결정할 때 비교하는 기준이 된다. ④ 주요 전자적 정보를 기준으로 그룹화된 전자적 정보들이 최소 유사성 비율(Minimum Similarity Percentage)에 의해 해당 그룹에 재배열 된다. 최소 유사성 비율이란 그룹에 주요 전자적 정보와 비교하는 전자적 정보가 얼마나 유사한지에 대한 척도이다. 비율이 100%에 달한다면 완벽한 텍스트 중복을 의미한다. 최소 유사성 비율을 높게 설정한다면 주요 전자적 정보와 유사하다고 판단되는 전자적 정보들의 수가 감소한다. 따라서 적절한 최소 유사성 비율을 설정하는 것이 매우 중요하다. 이때 그룹의 주요 전자적 정보와 일치하지 않는 전자적 정보가 검출되는 경우, 해당문서는 새로운 그룹의 주요 전자적 정보가 된다. ⑤ 보유하고 있는 모든 전자적 정보의 ③, ④과정이 끝난다면 그룹별로 약 한 개 정도의 주요 전자적 정보만을 보여주게 된다.

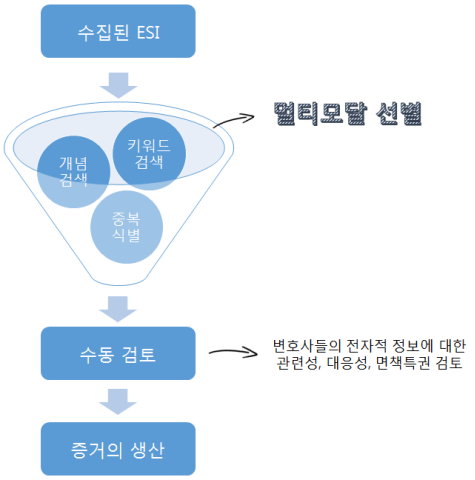


〈그림 7〉 텍스트 중복 식별 과정

3.2 검토를 위한 기계학습 기반 검색도구의 필요성

앞서 살펴본 검토에 사용되는 검색도구를 정리해 본다면 세 가지 유형으로 구분할 수 있다. 첫째, 직접적으로 해당하는 전자적 정보들을 검색할 수 있는 검색도구인 키워드 검색과 개념 검색. 둘째, 소송과 관련 있는 전자적 정보를 정리하여 변호사의 수동적인 검토를 돕는 데이터 뷰어 도구인 디스커션 스레딩. 셋째, 대량으로 수집된 전자적 정보들을 분석, 처리하여 소송과 관련이 없거나 중복되는 전자적 정보를 정리하는 대규모 데이터 처리 도구인 클러스터, 중복 식별로 정리된다.

검토과정에서는 단순히 하나의 검색도구를 사용하는 것이 아니라 두 가지 이상의 검색 도구를 사용하여 전자적 정보를 걸러내는데 이를 하이브리드 멀티모달 선별(Hybrid Multimodal Culling)이라고 부른다. 〈그림 8〉과 같이 전자적 정보를 수집하고 수집된 전자정보를 여러 도구를 사용하여 소송과 관련성이 있는 전자적 정보들을 선별한다. 다음, 선별된 전자적 정보들의 관련성, 대응성, 면책특권을 검토하기 위해 변호사들에 의해 수동적인 검토가 이루어진다. 마지막으로, 최종적 변호사들의 검토를 받은 전자적 정보들은 증거의 생산 단계로 넘어 가게 된다. 이 과정에서 사용되는 검색 도구들을 잘 조합하여 소송과 관련된 전자적 문서들을 선별해내는 것이 핵심 요소이다. 하지만 대량의 전자적 정보들을 놓고 한꺼번에 처리하는 방법으로는 기계적으로 선별된 전자적 정보들이 소송에 증거자료로 제출되기에 얼마나 적합한지 관련성을 검토하는 것은 불가능하다. 이처럼 전자적 정보의 관련성, 대응성, 면책특권을 평가하는 것은 아직까지도 변호사들의 몫이다.



〈그림 8〉 하이브리드 멀티모달 선별 모델

기존의 검색도구들은 대량의 전자적 정보들을 추려내는데 쉽고 효과적이지만, 단순히 기계적인 선별에 불과하다는 한계점이 있다. 여러 논의에서 주장되고 있는 기존 검색도구의 한계를 정리해보자면 다음과 같다.

첫째, 관련 없는 자료가 결과에 포함될 가능성이 높다. 단순 검색도구로는 소송과 전혀 관련성이 없는 자료라 하더라도 특정 단어가 포함되면 모두 검색결과로 나타나게 된다. 따라서 이를 다시 관련성이 인정되는지 여부를 판단해야하는 어려움이 있다. 즉 필요한 대상을 선별하여 검토해야 하므로 이를 위한 추가적인 시간과 비용이 요구된다.

둘째, 면책특권이 인정되는 자료가 노출될 가능성이 높다. 증거 자료를 제출하는 측이 증거자료를 수집함에 있어 추출된 전자적 정보가 방대하여 이를 모두 면밀히 검토하지 않고 제출

할 경우 이와 같은 문제가 발생한다. 예를 들어 Victor Stanley, Inc. v. Creative Pipe, Inc. 사안에서 Creative Pipe사는 약 4.9기가바이트의 검색 가능한 텍스트 자료와 33.7기가바이트의 검색 불가능한 텍스트 자료 중에서 면책 특권이 인정되는 165건의 자료를 부주의로 제출하였고, 이를 면책특권을 포기한 것으로 볼 것인지 여부에 관한 다툼이 있었다(E-Discovery Team, 2008).

셋째, 자료가 포함하고 있는 내용의 유형에 따라 검색이 어려운 경우가 있다. 텍스트 검색을 기반으로 하고 있는 기존 검색도구는 검색의 대상이 되는 자료가 이미지, 음성, 영상 자료인 경우 검색 자체가 어렵다. 해당 자료의 파일명이 내용과 일치하는 경우라면 그나마 검색이 가능하겠지만 그렇지 않은 경우라면 효율적인 검색이 이루어지기 어렵다.

넷째, 자료를 포함하고 있는 파일의 형식에 따라 검색이 어려운 경우가 있다. 검색도구는 대상 자료가 컴퓨터를 사용한 접근이 가능한 것을 전제로 하고 있다. 하지만 당사자가 가진 자료가 반드시 컴퓨터 하드디스크에만 존재하는 것은 아니므로 검색 자체가 얼마든지 존재할 수 있다.

과거의 전자증거개시 서비스 및 솔루션의 주요 목적은 사람에 의한, 즉 변호사에 의한 소송 쟁점과의 관련성 검토 작업 준비 단계로써, 단순히 전자적 정보를 식별하고 수집하는 역할만을 수행했다. 기업이 처리하는 전자적 정보의 양이 기하급수적으로 급증하고 있음을 고려할 때, 변호사들의 수동적인 검토를 하는데 있어 큰 부담을 안겨주고 있다. 또, 실제로 증거개시 절차에 소모되는 비용의 대부분이 변호사들의 인건비에 해당하므로 단순한 검색도구만으로는 비용 절감이 불가능에 가깝다.

이에 따라 새로운 검색도구 기술의 필요성이 대두되었고, 기존 검색도구 의 대안으로 기계 학습을 바탕으로 한 Predictive Coding 기술이 떠오르기 시작했다. 기계 학습은 인공 지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다. 전자증거개시 업계에서는 2010년부터 기계 학습을 적용한 검색 도구의 활용을 시작하였고, 2011년 TREC의 Legal Track을 비롯한 전자증거개시 관련 컨퍼런스에서는 자동화 된 기계 학습을 활용한 텍스트 마이닝(text mining)기반의 관련성 분류 기법이 주목받기 시작했다. 전자증거개시 과정에서는 기계 학습을 통하여 〈표 2〉와 같

〈표 2〉 기계 학습을 통한 검토 단계의 기대 효과

적용 대상 기능	기대 효과
- 기밀 정보 분류 - 특허 분류 - 그룹핑 - 관련성 재검토	기밀 또는 특허 관련 ESI를 분류하기 위한 범주를 정의하고, 정보관리 단계에서 미리 분류함으로써, 해당 범주의 ESI들이 검색되었을 때 별도의 검토 작업 없이 제외시킬 수 있게 됨
	소송 이슈와의 얼마나 관련성이 있는지에 대한 척도인 유사도에 따라 ESI를 그룹핑 하고, 증거 생산 시 비용적 한계로 인해 우선 생산 되어야 하는 ESI를 선별하고자 할 때, 유사도를 연관성 정보로 활용 가능함
	해당 효과들로 인해 검토 및 분석 작업에 소요되는 시간과 비용을 절감 할 수 있음

이 검토 단계에서의 효과를 기대할 수 있다.

이후 Predictive Coding은 2011년 전자증거개시 절차에서 가장 많은 관심을 받은 주제로 지목되었고, 2012년 2월에 Da silva Moore v. Publicis Groupe SA 사안에서 처음으로 연방법원에서 인정되었다. 이를 기점으로 Predictive Coding에 관한 다양한 포럼이 미국 전역에서 개최되고 있고, 일부 법조인들은 이를 전자증거개시의 미래로 받아들이고 있다고 한다(김도훈, 2014).

한편 미국 기록관리 학계에서도 전자증거개시에 대해 지속적으로 관심을 가져왔다. ARMA (Association of Records Managers and Administrators)에서는 출판물인 Information management에서 2006년부터 전자증거개시 제도에 대해 언급[3]하기 시작했고, Predictive Coding에 대해서는 2012년부터 언급하기 시작했다.

2012년 11/12월호[4]에는 Predictive Coding을 기밀문서들을 예측(predict)하는 데 도움을 주는 기계 학습 기반 기술의 한 종류라고 Sementic사의 전자증거개시 변호사인 Matthew Nelson의 말을 빌려 소개하고 있다. 이외에도 전자증거개시 관련한 연구들과 기사들을 발췌하여 Predictive Coding이 전자증거개시 과정을 진행하는데 있어 많은 비용감소를 가져 올 수 있다고 언급하고 있다. 마지막으로 당시에 Predictive Coding을 이용한 Da silva Moore v. Publicis Groupe SA사안에 대해 소개하고 있다.

2013년 1/2월호[5]에는 해마다 기하급수적으로 늘어나는 데이터를 처리하기 위해 적합한 도구가 Predictive Coding이라고 언급하며 전자증거개시 이외에도 보존(retention), 처분(disposition), 기록관(archives), 증거보존 통지(legal holds), 보호 명령(protective orders), 데이터 개선(data remediation), 파일 전송(file transfer), 지적 재산권(intellectual property), 지식 관리(knowledge management), 비즈니스 인텔리전스(business intelligence), 민감한 정보(sensitive information), 데이터 보안(data security), 가짜 데이터(data fraud), 파일 공유(file shares) 등 여러 분야에서 쓰일 수 있다고 소개했다.

2013년 3/4월호[6]에는 Predictive Coding이 전자증거개시 과정에서 윤리적인 의무가 되고 있다고 언급했고 이에 대한 근거로 EDBP에 Predicative Coding의 과정이 추가됨을 들었다.

2014년 5/6월호[7]에는 Predictive Coding이 기록 보존 전문가(records retention expert)의 도움을 받아 학습된다면, Predictive Coding이 훨씬 더 좋은 솔루션이 될 수 있다고 언급했다.

---

3) Swartz, N. (2006). New Rules for E-Discovery. Information Management, 22-26.
4) An ARMA International Publication (2012, November/December). Making 'Predictive Coding' Pay Needs Cooperation. Information Management, 8.
5) Issacs, L. (2013). Rolling the Dice with Predictive Coding Leveraging Analytics Technology for Information Governance. The Information Management Journal, 47(1), 23-26.
6) An ARMA International Publication (2013, March/April). Predictive Coding to Become an Ethical Obligation. Information Management, 12.
7) An ARMA International Publication (2014, May/June). Predictive Coding: Not Just for E-Discovery. Information Management, 17.

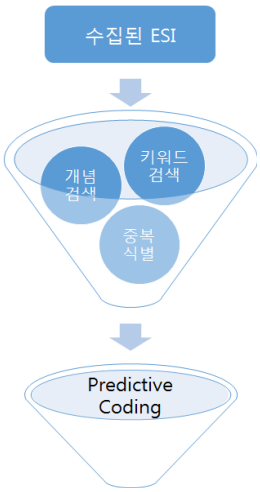
4. Predictive Coding 도구 적용 방안

4.1 기계학습기반의 Predictive Coding 원리

Predictive Coding은 전자적 정보의 검색과 수집 그리고 이를 통해 얻어진 결과에 대한 검토에 이르는 전 과정에서 기술 지원을 받는 것으로 컴퓨터가 전자적 정보의 관련성, 대응성, 면책특권 등에 관한 분석을 수행하여 이를 수치화 하여 제시해주는 검색 도구이다.

Predictive Coding은 선별 단계에서부터 소프트웨어 테스트 단계까지 총 다섯 단계로 구성되며 그 자세한 내용은 다음과 같다.

첫째, 선별 단계에서는 앞서 소개했던 검색 도구들을 이용하여 일차적으로 전자적 정보들의 전자적 정보들의 컬렉션들 중 소송과 관계없는 불필요한(junk) 정보를 걸러낸다.



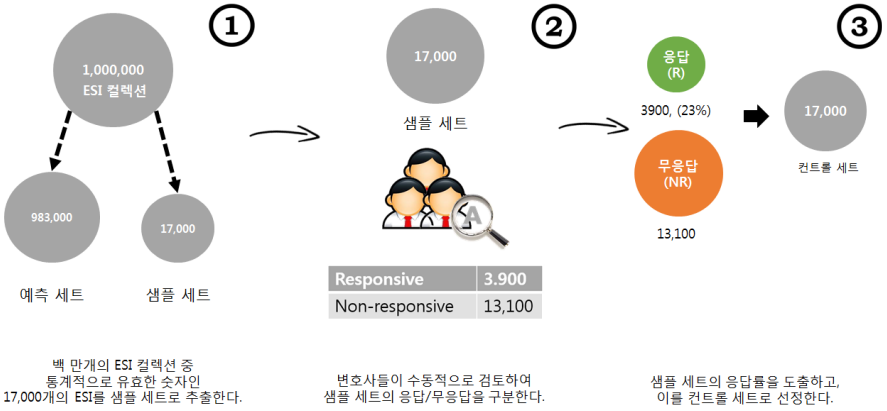
〈그림 9〉 선별 단계

Predictive Coding을 진행하기 전 거대한 규모의 전자적 정보 컬렉션에서 소송과 관련성이 있는 전자적 정보들을 일차적으로 선별하는 것은 프로세스 진행 과정에서 시간과 비용을 절약할 수 있게 된다. 따라서 Predictive Coding 도구는 Predictive Coding 기술 이외에도 기존의 검색 도구들과 연계시키는 것이 중요하다.

둘째, 수율(yield) 추정 단계는 앞에서 일차적으로 정제된 전자적 정보의 컬렉션들이 소송과 관련성이 있는 컬렉션 인지 평가하게 된다. 수율의 사전적 의미는 투입 수에 대한 완성된 양품의 비율을 말하는 것으로, 흔히 양품률 이라고도 하며, 불량률과는 반대되는 의미를 갖는다. 이 단계에서 수율이란 전체 전자적 정보 컬렉션에서 소송과 반응성(responsive)이 있는 전자적 정보의 비율을 말한다. 수율 추정 방법은 정제된 전자적 정보의 컬렉션 중 임의로 뽑아낸 통계적으로 유효한 수의 전자적 정보들을 수동으로 검토하고 반응성이 있는지 측정한다. 이때의 반응성 비율이 수율이 된다. 일반적으로 천만 개의 전자적 정보들을 대상으로 수율 추정을 하였을 때 20%에 해당하는 것으로 나타났다.

셋째, 컨트롤 세트 선정 단계는 앞선 두 단계를 거친 전자적 정보들 중 일부를 샘플 세트(sample set)로 추출하고 이를 토대로 변호사들이 소송과 관련성, 대응성, 면책특권에 대해 검토하여 응답률을 도출한다. 도출한 응답률이 소송 당사자나 법원에 의해 합의된 기준을 만족한다면 이를 컨트롤 세트로 삼는다.

컨트롤 세트 선정 단계의 자세한 내용은 〈그림 10〉과 같다. ① 전자적 정보 컬렉션에서 통계적으로 유효한 수인 10,000개에서 20,000개 정도의 전자적 정보들을 임의적으로 샘플 세트로 뽑아낸다. 이때 전자적 정보 컬렉션은 샘플



〈그림 10〉 컨트롤 세트 선정 예

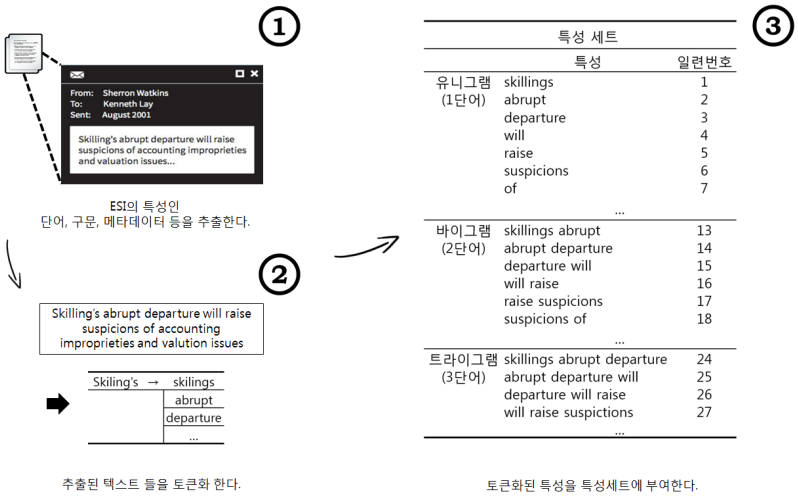
세트와 나머지 예측 세트(prediction set)로 나누어지게 된다. ② 샘플 세트를 변호사들이 수동적으로 검토를 하고 이를 응답(responsive, R)과 무응답(non-responsive, NR)로 구분한다. ③ 샘플 세트의 응답률이 오차가 일정 기준을 충족할 경우 이를 컨트롤 세트로 선정한다. 〈그림 10〉에서는 17,000개의 전자적 정보들을 샘플 세트로 삼았는데, 일반적으로 17,000개의 통계적으로 유효한 무작위 표본은 99%신뢰도 수준에서 ±1%의 오차를 보인다. 따라서 샘플 세트에서 23%의 응답률이 나올 경우 컨트롤 세트는 22~24%의 응답률을 보일 것이라고 확신할 수 있다.

넷째, 소프트웨어 학습단계는 전 단계에서 도출한 컨트롤 세트를 중심으로 소프트웨어를 학습시키는 단계로써 크게 특성 세트(feature set) 생성 과정과 가중치 정제(weight refine) 과정으로 나뉜다.

특성 세트 생성 과정에서는 Predictive Coding 소프트웨어가 컨트롤 세트들의 전자적 정보들을 검토하고 해당 전자적 정보의 특성(feature)에 가중치(weight)를 할당한다. 본 과정에서 말하는 특성이란 하나의 단어나 혹은 연속된 단어를 말한다. 각 특성에는 가중치가 붙게 되는데, 가중치는 처음 0에서부터 시작한다. 해당 전자적 정보에 포함하고 있는 특성들이 응답(R)로 구분된다면 가중치가 증가하고, 특성들이 무응답(NR)로 구분된다면 가중치가 감소하게 된다.

특성 세트 생성 과정의 상세한 내용은 〈그림 11〉과 같다. ① 컨트롤 세트에 속하는 전자적 정보들의 특성인 단어, 구문, 메타데이터 등을 추출한다. ② 추출된 텍스트들을 '토큰화' 한다. 토큰화란 유럽과 미국의 언어의 경우 단어와 단어 사이가 분리의 구분이 되는데, 토큰화 된 개개의 단어 단위를 토큰이라 부른다. 먼저 텍스트들을 단어로 쪼갠다. 이후 공백과 문장 부호를 제거하고 대문자들을 소문자들로 변경하여 토큰을 생성하게 된다. ③ 토큰화 된 특성을 특성 세트에 부여한다. 〈그림 11〉에서는 최대 3개의 연속된 단어까지 토큰으로 구성하였다.

가중치 정제 과정 에서는 Predictive Coding 소프트웨어가 특성 세트에 있는 특성별로 컨트롤 세트에 포함된 모든 전자적 정보와 비교하며 가중치 점수를 판단하게 된다. 가중치가 판



〈그림 11〉 특성 세트 생성 과정 예

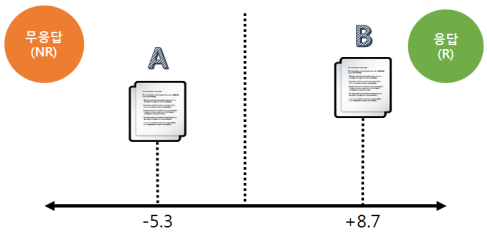
단된 특성 세트를 〈표 3〉과 같이 가중치 테이블(weight table)이라 부른다.

〈표 3〉 가중치 테이블 예시

특성(Feature)	가중치(Weight)
will raise	-0.6
suspicions	0.7
of	-
improprieties and valuation	1.8

이 과정에서는 특성별로 가중치를 산출하기 위해 많은 시행착오를 거치게 된다. 소프트웨어가 많은 전자적 정보를 검토하여 가중치를 정제한다면, 가중치 테이블은 변호사들의 판단을 잘 반영하게 될 것이다. 가중치 테이블로 소프트웨어가 변호사와 비슷한 판단을 내릴 때 까지 반복하는 가중치 정제 과정을 기계 학습이라고 부른다.

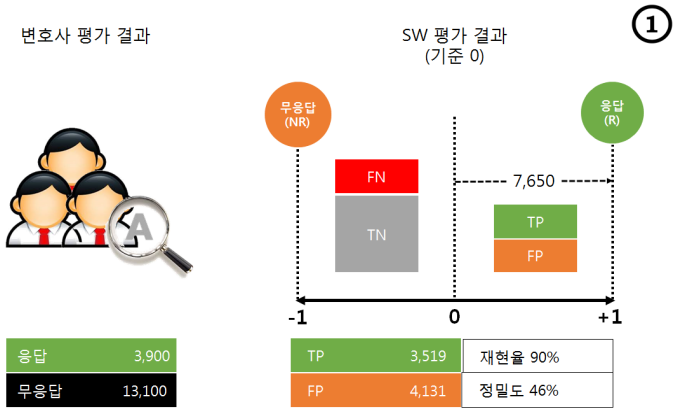
컨트롤 세트 내에서 가중치 테이블의 정제가 완료되면, 소프트웨어는 컨트롤 세트 내의 각 전자적 정보에 대한 점수를 매기게 된다. 전자적 정보에 대한 점수는 각 전자적 정보들이 포함하는 특성들의 가중치의 합이다. 이를 토대로 소프트웨어는 〈그림 12〉와 같이 전자적 정보의 점수가 양수(+)인 전자적 정보 B는 응답(R)으로, 점수가 음수(-)인 전자적 정보 A는 무응답(NR)으로 판단하게 된다.



〈그림 12〉 전자적 정보의 점수에 따른 응답/무응답 판별

다섯째, 소프트웨어 테스트 단계는 학습을 마친 소프트웨어를 이용하여 컨트롤 세트 내 전자적 정보를 선별 테스트하고, 통계적으로 유효한 결과를 도출하는지 여부를 평가하기 위한 단계이다.

〈그림 13〉과 같이 첫 번째 과정에서는 컨트



〈그림 13〉 소프트웨어 테스트 1과정

롤 세트 내의 각각의 전자적 정보들을 소프트웨어가 검토한 가중치의 합을 통계를 낸다. 이때 가중치의 합을 최대, 최소 ±1점으로 치환하고 0점을 기준으로 응답(R), 무응답(NR)을 구분한다. 이후 변호사들이 검토한 응답(N)과 무응답(NR)과 비교한다.

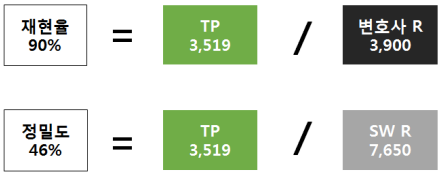
예를 들어, 변호사들은 17,000개의 전자적 정보를 응답(R)이 3,900개, 무응답(NR)이 13,100개라고 평가했다. 소프트웨어는 점수 0점을 기준으로 응답(R)이 7,650개 무응답(NR)이 9,350개라고 평가했다. 소프트웨어와 변호사들이 평가한 응답(R)과 무응답(NR)을 비교하면, TR(true positive), TN(true negative), FP(false positive), FN(false negative)을 도출해낼 수 있다. 각각의 의미는 〈표 4〉와 같다.

예시에서 응답(R)을 비교하여 3,119개의 TP와 4,131개의 FP를 구할 수 있다고 가정한다. FN은 0점이 기준일 경우 변호사들의 평가가 완벽하다고 가정할 때 FN의 값은 0에 수렴한다. 앞에서 도출한 TP, FP을 이용하여 0점을 기준으로 한 소프트웨어의 재현율(recall)과 정밀도(precision)를 도출한다. 정보 검색 분야에서, 재현율과 정밀도는 검색된 전자적 정보 및 관련 있는 전자적 정보에 의해 정의된다.

〈표 4〉 TP, TN, FP, FN의 의미

이름	의미
True Positive	수동 검토와 SW 검토가 R로 일치함 (실제 정답과 실험 결과가 일치함)
True Negative	수동 검토와 SW 검토가 NR로 일치함 (실제 정답과 실험 결과가 일치함)
False Positive	수동 검토는 NR, SW검토는 R로 불일치함 (True로 예측했지만 결과가 False)
False Negative	수동 검토는 R, SW검토는 NR로 불일치함 (False로 예측했지만 결과가 True)

예시에서 재현율은 수동으로 검토한 응답(R)된 전자적 정보 중 소프트웨어 검토가 수동 검토와 응답(R)으로 일치한 전자적 정보를 재현하였는지를 나타낸다. 정밀도는 소프트웨어가 검토한 응답(R)된 전자적 정보 중 소프트웨어 검토와 수동 검토가 응답(R)으로 일치한 전자적 정보를 정확성을 나타낸다. 재현율과 정밀도를 구하는 방법은 〈그림 14〉와 같다. 따라서 0점을 기준으로 재현율은 90%, 정밀도는 46%가 된다.

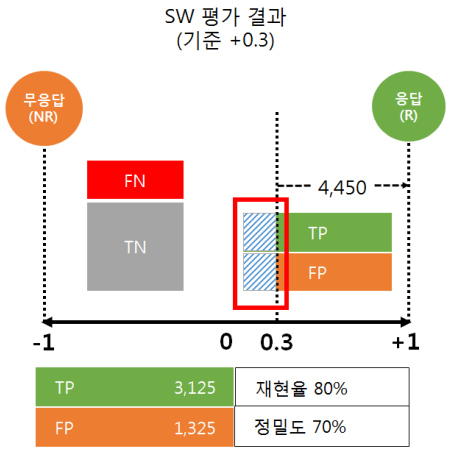


- 변호사 R : 변호사가 검토한 응답 ESI 수량
- SW R : SW가 검토한 응답 ESI 수량
- TP : SW 검토하여 응답한 ESI의 수량 중 변호사 검토와 비교하여 응답 일치한 ESI 수량

〈그림 14〉 0점 기준 재현율/정밀도

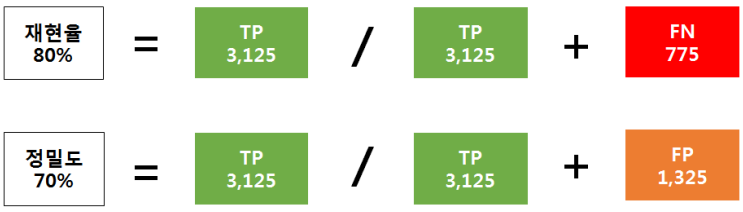
0점 기준에서는 재현율은 90%에 달하지만, 정밀도는 46%로 정밀도가 매우 떨어진다. 이는 적합한 기준이라 할 수 없다. 따라서 정밀도를 높이기 위한 과정이 이어진다.

정밀도를 높이기 위한 공정으로 기준점을 조절하며 재현율과 정밀도를 체크한다. 예를 들어, 〈그림 15〉와 같이 소프트웨어의 기준점을 +0.3으로 이동한다면 정밀도는 70%로 높아지게 된다. 그러나 그림에서 빗금 친 부분처럼 기준점이 이동하며 소프트웨어가 검토한 기준점 0에서 +0.3 사이의 응답(R)된 전자적 정보가 무응답(NR)처리 되기 때문에 재현율은 80%로 떨어지게 된다. 기준점을 조절하였을 때 재현율과 정밀도를 구하는 방법은 〈그림 16〉과 같다.



〈그림 15〉 소프트웨어 테스트 2과정

예시로 든, 기준점을 +0.3으로 조절하여 기준점 0과 비교했을 때 변호사가 검토한 응답(R)된 전자적 정보 중 소프트웨어가 무응답(NR)으로 판단하는 775개의 전자적 정보가 발생하는데 이를 FN이라 한다. 따라서 FN과 TP의 합은 항상 변호사가 검토한 응답(R)의 수인 3,900개와 같기 때문에, 재현율은 80%가 된다. 정밀도는 0.3점 기준으로 했을 때, 소프트웨어



- TP : 기준점을 0.3으로 이동하여 TP의 수량이 기준점 0보다 감소
- FP : 기준점을 0.3으로 이동하여 FP의 수량이 기준점 0보다 감소
- FN : 기준점을 0.3으로 이동하였기 때문에 775개의 FN발생
- TP + FN = 변호사 R
- TP + FP = 0.3점 기준의 SW R

〈그림 16〉 0.3점 기준 재현율/정밀도

가 검토한 응답(R)된 전자적 정보 중 TP의 비율이다. TP와 FP의 합은 소프트웨어가 검토한 응답(R)된 전자적 정보의 수와 동일하고, 정밀도는 70%가 된다.

이처럼 기준점을 조절하는 것으로 결과의 재현율과 정밀도를 조절할 수 있다. 재현율이 높더라도 정밀도가 낮다면 소프트웨어로 도출된 전자적 정보들의 신뢰성이 떨어진다. 반대로 정밀도가 높더라도 재현율이 낮다면 전체 전자적 정보 컬렉션 중 소송과 관련 있는 전자적 정보들을 놓칠 확률이 높다. 재현율과 정밀도는 반비례하기 때문에 적절한 재현율과 정밀도를 도출할 수 있는 기준점을 찾는 것이 중요하다.

재현율과 정확도 간의 적절한 선을 맞추고 기준점을 설정하면 소프트웨어 검토의 범위를 컨트롤 세트에서 전체 전자적 정보 컬렉션으로 변경하여 최종적으로 소프트웨어를 테스트 한다. 앞서 무작위하게 도출된 통계적으로 유효한 수의 컨트롤세트는 전체 전자적 정보의 컬렉션을 대표할 수 있다. 따라서 소프트웨어의 검토 범위가 전체 전자적 정보 범위로 넓어지더라도 약간의 성능 저하는 있겠지만, 앞선 컨트롤 세트의 검토 결과와 유사하게 도출될 것이다.

이후 최종 품질검사로 변호사들이 샘플 전자적 정보를 뽑아 검사한다. 대부분의 경우에 2,000개의 샘플을 검사하였을 경우 99%에 신뢰도에서 ±3%의 오차를 보인다.

이처럼 Predictive Coding은 일차적으로 기존의 검색 도구들을 이용하여 불필요한 전자적 정보를 걸러낸 후 이차적으로 기계 학습을 활용함으로써 기존의 그 어떤 검색방식보다 효율성이 높다. Predictive Coding은 초기의 일부 기초 전자적 정보에 대한 변호사의 검토를 필요로 하지만 컴퓨터가 전자적 정보의 관련성, 대응성, 면책특권에 관한 검토를 상당부분 수행해 준다는 점에서 시간과 비용 면에서 기존의 검색 도구들에 비해 사건당사자나 변호사의 부담을 현저하게 줄여준다.

실제로 Predictive Coding 서비스를 제공하는 11개의 업체를 대상으로 한 연구결과에서 기존의 방식에 비해 4개 업체는 45%의 비용을 절감하였고, 7개의 업체는 70%까지 비용을 절감한 것으로 나타났다. 다른 연구결과에서는 변호사가 전자증거개시과정을 수행하는데 소용되는 시간을 80% 가량 줄여줄 수 있는 것으로 나타났다(Acosta, 2012).

〈표 5〉 Da Silva Moore v. Publicis Groupe SA 사안 Predictive Coding 방법론

- 2,399개의 임의의 문서 샘플을 검토한다. (신뢰수준: 95%, 오차율: ±2%)
- Predictive Coding 도구의 결과를 개선하기 위해 반복적으로 기계학습을 진행한다. (7라운드)
- 적절한 재현율과 정밀도를 확인하고, 검토하여 소송과 관련성 있는 문서를 4만 건 도출한다.
- 훈련되고 테스트되는 전자적 정보들은 투명성을 유지하기 위해 원고 측의 전자적 정보들을 테스트 한다.
- 피고는 원고에게 컨트롤 세트를 만들 모든 면책특권이 없는 문서를 넘긴다.
- 피고는 원고에게 7라운드의 반복적인 검토 동안 모든 면책특권이 없는 문서를 넘긴다.
- 피고는 원고에게 최종 검토시 모든 면책특권이 없는 문서를 넘긴다.

출처: Schoenecker, E. Jr. (2015). Nine Cases on Predictive Coding From Modus 재구성. Retrieved from https://www.linkedin.com/pulse/nine-cases-predictive-coding-from-modus-edward-schoenecker

4.2 Predictive Coding을 이용한 소송 사례

4.2.1 Da Silva Moore v. Publicis Groupe SA 사안[8)]

Da Silva Moore v. Publicis Groupe SA 사안은 Predictive Coding의 사용이 법적으로 승인된 최초의 사례로써, 법원으로부터 Predictive Coding을 이용하여 제출한 전자적 정보가 증거로 인정되었다.

해당 사안은 소송이 개시되고 소송 당사자 간 협의를 통하여 검색 도구를 선정할 때 발생한 핵심 쟁점이 있다. 피고 측은 수집한 3백만 개의 전자적 정보를 평가하는데 어려움을 겪었다. 피고 측은 원고 측 요구에 따라 소송과 관련된 증거인 전자적 정보를 제출하는데 있어 20만 달러 이상은 쓸 수 없다고 밝히고, Predictive Coding을 활용할 경우 한 건의 전자적 정보 당 5달러의 비용으로 약 4만 건의 전자적 정보를 증거로 제출할 수 있을 것으로 판단했다(김도훈, 2014). 이에 전자증거개시 서비스 제공업체와 협의하여 Predictive Coding 도구와 방법론을 사용할 것을 제안하였다. 방법론의 일부 내용을 살펴보면 〈표 5〉와 같다.

피고 측은 방법론을 토대로 원고 측과 해당 사안에서 증거개시를 위한 Predictive Coding 프로토콜을 법원에 제시하였고, 법원은 원고 측의 동의가 없음에도 불구하고 사용을 허가했다. 원고 측은 미연방민사소송규칙 제26조와 미연방 증거규칙 제702조를 근거로 이의를 제기했다. 법원에서는 전자증거개시상 증거자료의 제출은 변호사로 하여금 100% 완전할 것을 요구하는 것이 아니며, Predictive Coding이 기존의 키워드 검색과 비교하여 성능이 떨어지지 않는다고 평가하여 원고 측의 이의제기를 받아들이지 않았다.

해당 사안은 최초로 Predictive Coding이 사용되어 가공된 전자적 정보가 증거자료로 채택된 사안이라는 점과 기존의 키워드검색과 Predictive Coding을 비교하여 더 나은 평가를 내렸다는 점, Predictive Coding이 소송 당사자 간의 합의 없이 법원에 명령에 의해 사용되었다는 점에서 의의가 있다.

4.2.2 In re Actos: (Pioglitazone) Prod. Liability Litigation 사안[9)]

In re Actos: (Pioglitazone) Prod. Liability Litig 사안은 당뇨병 치료제인 'Actos'를 복용하는 당뇨병 환자들의 방광암 유발율을 증가시킨다는 피고의 주장아래 제기된 소송이다. 해당 사안은 피고 측과 원고 측이 Predictive Coding 프로토콜에 합의에 도달한 첫 번째 소송으로 소송 당사자 양측 모두 Predictive Coding의 결과물 도출을 위해 협력했다는데 의의가 있다.

해당 사안의 Predictive Coding 프로토콜을 살펴보면, 첫째, Predictive Coding을 '개념 증명(proof of concept)'과정을 거쳐 검토 도구로 삼았다는 점이다. 개념 증명이란 제품, 기술, 정보 시스템 등이 조직 특수 문제 해결을 실현할 수 있다는 증명과정으로써, 소송 당시 검증되지 않은 Predictive Coding 도구를 개념 증명하기 위해 소송 당사자들 간의 전자적 정보들을 제공

8) Moore v. Publicis Groupe SA, 2012 WL 1446534(S.D.N.Y. Apr. 26, 2012).
9) In re Actos (Pioglitazone) Product Liability Litigation, No. 611-md-2299 (W.D. La. July 27, 2012).

하기로 합의했다.

둘째, 원고 측에 대한 소송 당사자 간의 협력과 협업과정의 투명성을 강조했다는 점이다. 앞서 소개한 Predictive Coding의 각각의 과정에서 피고 측과 원고 측이 모두 참여하여 Predictive Coding의 전체 과정의 투명성을 확보했다. 또, 일련의 프로세스가 종료되고 결과를 검토하여 피고 측과 원고 측 모두 만족하는 결과 얻었을 때, 상호 동의하에 다음 단계가 진행되도록 했다.

셋째, 보다 완벽한 최종 결과물을 도출하기 위해 노력했다는 점이다. 이를 위해 컨트롤 세트를 기준으로 한 검토 결과가 ±5%의 오차 범위 내 95% 신뢰도 수준에 이르기까지 Predictive Coding 소프트웨어의 기계학습과정을 반복했다. 또, 누락되는 증거자료를 최소화하기 위해 Predictive Coding 소프트웨어의 기준점을 소송 당사자 간 상호 합의하에 지정했다. 마지막으로 상호 합의하에 지정된 기준점을 이용하여 도출해낸 최종 결과물을 가지고 최종적으로 수동 검토를 진행하였다.

4.2.3 EORHB, Inc. v. HOA Holdings LLC 사안[10]

반면, EORHB, Inc. v. HOA Holdings LLC 사안에서는 피고 측과 원고 측 양쪽에서 검색 도구들의 선정에 관해 아무런 주장을 하지 않았음에도 불구하고 법원은 해당 사안의 경우 대량의 전자적 정보가 대상이 되므로 Predictive Coding을 사용하여 이를 진행할 것을 명했다. 추가적으로 법원에서는 소송 당사자 양측 모두 Predictive Coding의 사용에 동의하여야 하며, 동의하지 않는 경우 판사에게 이유를 제시해야 한다고 요구하였다. 마지막으로 양측에서 Predictive Coding 도구 제공업체를 제시하면 이중 법원에서 선정하는 제공업체의 소프트웨어를 이용하여 증거를 제출할 것을 요구하였다.

해당사안은 소송 당사자들이 검색 도구들을 선정하지 않았음에도 불구하고 법원에서 직접 Predictive Coding을 사용하게 명령했다는 점에서 그 의의가 있다.

4.2.4 In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation 사안[11]

In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation 사안은 Biomet사에서 제조한 "metal-on-metal" 관련 제조물 책임 소송이다. 본 소송은 미국 전역에서 별도로 수십 건의 소송이 진행되었지만, 법원의 재판 통합 명령으로 수십 건의 소송이 통합되어 진행되었다. 이에 따라 증거 자료를 제출하기 위한 광범위한 전자적 정보의 검색이 진행 되었으며, 두 건의 Predictive Coding 도구 사용요청이 제안되었다.

피고 측은 초기에 잠재적으로 소송과 관련이 있는 1,950만개의 전자적 정보를 수집했다. 그 중 키워드 검색을 이용하여 390만개의 전자적 정보로 추려냈고, 중복 제거를 이용하여 250만개의 전자적 정보로 최종적으로 추려냈다. 이

10) EORHB, Inc., et al. v. HOA Holdings, LLC, C.A. No. 7409-VCL (Del. Ch. Oct. 15, 2012).
11) In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation, No. 3:12-MD-2391 (N.D. Ind. Aug. 21, 2013).

어 250만개의 전자적 정보를 검토하기 위해 Predictive Coding 도구를 적용했다.

원고 측은 Predictive Coding 도구를 이용한 검토 과정을 진행하는 것이 반대했다. 정확하게는 Predictive Coding 도구의 사용을 반대한 것이 아니라 피고 측이 초기에 수집한 1,950만개의 전자적 정보 중 키워드 검색을 적용하여 390만개로 추려낸 점을 지적했다. 키워드 검색을 한 전자적 정보들이 Predictive Coding 도구를 적용하기 전에 오염되었다는 것이다. 이어 원고 측은 피고 측이 제시한 Predictive Coding 도구의 우수성을 보여주는 연구를 인용하여, 초기의 정제되지 않은 전자적 정보에 Predictive Coding 도구를 적용해야 한다고 주장했다.

법원은 두 가지 이유로 원고 측의 요청을 거부했다. 첫째, 피고 측의 키워드 검색을 이용한 초기 검토 과정은 법적인 규칙과 원칙을 모두 준수하였기 때문에 피고 측의 의무를 다하였다. 둘째, 키워드 검색을 이용한 초기의 검토 과정 중 누락된 소송과 관련된 전자적 정보들이 있더라도, Predictive Coding을 적용하여 소모되는 비용과 비교했을 때 비용적인 면이 더 중요하다는 점을 지적했다.

이외에도 Dynamo Holdings v. Comm'r 사안[12]이나 Rio Tinto PLC v. Vale, S.A 사안[13]과 같이 Predictive Coding과 연관되어 있는 소송 사례들을 살펴봤을 때 Predictive Coding의 신뢰도는 법학적 측면에서 검증된 것으로 판단되며, 비용적인 면에서도 강점을 보여 소송 당사자들이 제안하거나 법원에 명령에 의해 Predictive Coding 도구를 사용하기에 이르렀다. 다만 도구의 사용에 있어서 소송 당사자 간의 협의를 통한 양측의 동의와 참여로 투명한 Predictive Coding 과정을 진행하여야 하며, 진행하기 전 법원의 사전 승인 허가를 받아야 함이 강조되고 있다.

4.3 Predictive Coding과 기업기록관리를 바탕으로 한 전자증거개시 대응 전략

4.3.1 소송 절차에 따른 전자증거개시 TF 구성요소

일반적으로 기업들의 전자증거개시 대응전략을 살펴보면 소송이 발생하는 시점에서 전자증거개시를 위한 TF팀이 조직된다. 기업이 소송을 준비하는데 필요한 TF팀의 핵심요소로는 〈표 6〉과 같이 법무담당자와, 전산담당자, 기록관리자, 보안담당자로 제시할 수 있다.

법무담당자는 전자증거개시과정에서 핵심역할을 하게 된다. 소송이 개시되기 전 공소장의 작성부터 진행과정 중 사건당사자간의 협의와 최종 심리까지 모두 법무담당자를 중심으로 일련의 과정이 진행된다. 국내 기업의 경우 대기업을 중심으로 전담 법무팀이 구성되어 있는 경우도 있다. 하지만 기업 내부의 법무팀 만으로는 전자증거개시제도를 진행하는데 부족함이 있다고 파악된다. 따라서 기업 내부사항을 잘 알고 있는 내부 법무담당자와 연방민사소송규칙을 확실히 파악하고 전자증거개시 대응 서비스를 제공하는 외부 법무법인과 연계하여 법무

12) Dynamo Holdings Limited Partnership, et al., v. Commissioner of Internal Revenue, 143 T.C. No. 9 (Sept. 17, 2014).
13) Rio Tinto PLC v. Vale, S.A., et al., No. 1:14-cv-03042-RMB-AJP (S.D.N.Y. Mar. 2, 2015).

〈표 6〉 전자증거개시 대응 TF팀 핵심요소

담당 업무	역할
법무담당자	- 소송 준비와 수행의 핵심 역할 - 법률 대리인(법무 법인)을 선정하기 위한 분석과 판단 - 소송에 필요한 솔루션 제공자(검색 도구 기술 제공자)를 선정하기 위한 분석과 판단 - 관련 전자적 정보 보존 명령(Litigation Hold) 작성
전산담당자	- 기업의 네트워크 시스템과 저장장치에 대한 확인과 설명 제공 - 소송과 잠재적으로 관련된 전자적 정보가 어떠한 저장장치에 저장되어 있는지 확인 - 저장된 전자적 정보를 보존・수집 - 기업에서 사용되는 파일의 포맷에 대한 설명 제공
기록관리자	- 기록관리 정책 운영 현황에 대한 증거 기록 제공 - 어떤 내용의 전자적 정보가 어디에 저장되는지 설명 제공 (전산 담당자가 물리적인 의미에서의 정보 위치 식별을 담당한다면 정보 관리자는 내용적인 위치 식별을 담당) - 정보의 등급이 구분되어 있을 경우 해당 구분 정책에 대한 설명을 제공
보안담당자	- 소송과 잠재적 관련이 있는 전자적 정보의 삭제, 백업정책 정보 제공 - 기업 전자적 정보의 보호 활동

담당자를 구성할 필요가 있다.

전산담당자는 전자증거개시과정 진행시 전자적 정보와 관련된 기술적 지원을 맡게 된다. 소송이 발생하면 잠재적으로 관련된 전자적 정보의 위치를 파악하고 수집・보존하게 된다. 이어 법무담당자를 도와 소송과 관련된 전자적 정보와 연결된 업무담당자의 리스트를 작성하기도 하며, 기업에 구축된 네트워크 시스템과 데이터베이스에 대한 확인과 설명을 제공하고 기업 내부에서 사용되는 파일 포맷 등을 관리하게 된다.

기록관리자는 기업 내 기록관리 정책과 규정을 세우는데 핵심적인 역할을 하고 기업에서 업무활동 중 생산되는 전자적 정보가 어떠한 시스템에서 어떻게 관리되는지 설명할 수 있어야 한다. 소송이 발생하게 되면 기록관리자는 법무담당자를 도와 잠재적으로 소송과 관련된 전자적 정보들의 수집・보존을 위하여 내용 정보를 제공하게 된다.

전산담당자와 기록관리자가 제공하는 정보는 비슷하게 느껴질 수 있으나 서로 다르다. 전산담당자의 위치 정보는 특정 서버에 저장된 전자적 정보가 어떤 형태로 저장되고, 그 저장장치가 어떤 네트워크로 구성되어 있고, 어떻게 운용하는지 등에 대한 정보를 제공한다. 반면 기록관리자의 내용정보는 해당 전자적 정보는 어디서 생산되고 어디서 관리하고 있는지, 취급은 어떻게 하고 있는지, 어떤 부서에서 접근하여 활용하는지 등에 대한 정보를 제공한다.

보안담당자는 기업 내 정보 보안정책을 수립하고, 시스템에 대한 접근을 통제・관리하여 기업의 전자적 정보를 보호하는 역할을 한다. 소송이 발생하면 소송과 관련 있는 전자적 정보의 보존에 지장을 줄 수 있는 삭제 정책과 각종 위험으로부터의 백업정책을 통제하고 해당 정보를 법무담당자에게 제공한다.

기업의 입장에서 소송을 진행할 때 소송에 적합한 법무담당자와 그 법무담당자를 도와 기업

내부를 잘 파악하고 소송과 관련된 정보를 제공하는 전산담당자, 기록관리자, 보안담당자의 선정은 매우 중요하다. 소송이 발생하는 시점에서 전자증거개시 대응을 위한 TF팀을 구성하고 나면 민사소송절차를 따라 전자증거개시를 진행 하게 된다.

전자증거개시절차제도의 프로세스에 따라 소송이 개시되고 99일 이내에 소송 당사자 간 협의를 통해 공개되어야할 증거 범위와 형식, 시간 및 비용을 결정하게 되는데, 소송 당사자 간 협의에서 어떤 검색 도구를 사용하여 증거를 도출할지 결정된다. 증거의 제출 방법에 관한 선택권은 당사자에 있고, 세도나 원칙상으로도 증거를 제출하는 측에서 정하는 것이 바람직 한 것으로 제시되어 있다. 앞 절에서 살펴본 다양한 Predictive Coding 사례들에서도 협의 과정에서 소송 당사자 양측에서 사용할 검색 도구를 제안하는 경우가 일반적이나, 법원 측에서 특정한 검색 도구를 사용하라고 제안하거나 명령하는 경우도 있다. 소송 당사자 간 협의를 마치고 나면, 소송 개시 120일 이내에 증거자료를 제출해야 한다.

앞서 말했듯이 검토과정에서 시간과 비용이 가장 많이 소모되고 제출된 증거가 소송의 승패와 직결되는 만큼 전자증거개시 과정을 도와줄 검색 도구 제공 업체를 선정하는 것 역시 중요한 핵심 쟁점이라 볼 수 있다. 만약 소송 당사자 간 협의에서 증거 제출 시 사용하는 검색 도구를 Predictive Coding을 사용한다고 협의한다면, Predictive Coding 서비스를 제공하는 제공업체의 선정과 이와 협력하여 전체적인 소송을 진행하고 최종적인 검토를 하는 법무담당자의 선정은 매우 중요할 것이다.

4.3.2 기록관리를 통한 대응 전략

기존의 전자증거개시 대응전략은 소송이 예상되는 시점부터 준비되어 소송발생을 기점으로 TF팀이 구성되어 진행된다. 하루에도 대량의 전자적 정보가 생산되는 기업들의 현 상황에서는 소송이 발생되는 시점에서 일정 시간 내에 소송과 잠재적으로 관련이 있는 전자적 정보를 수집하고 분류하여 평가하는 과정은 시간과 비용이 많이 소모되는 일이며, 이 과정에서 소송의 승패가 결정된다고 해도 과언이 아니다.

이에 따라 관련 업계에서도 증거개시과정 중 첫 번째 단계인 정보의 관리 단계의 중요성이 점점 중요해지고 있다. 소송이 예상되는 시점에서부터 전자증거개시 과정의 대응이 아닌 일반적인 상황에서의 법률적인 체계적인 수집, 분류, 평가, 보존 등의 기업기록관리를 통해 전자증거개시과정의 시간과 비용의 소모를 크게 줄일 수 있을 것이다.

따라서 체계적인 정보의 관리를 위해 기업 기록관리 정책을 수립하고 기록관리 시스템을 구축하여 매일 생산되는 방대한 양의 전자적 정보를 관리해야 한다. 소송에 대응하기 위한 기록관리 정책에 포함되어야 할 내용은 〈표 7〉과 같다.

기업의 전자적 정보에 체계적인 기록관리를 적용하게 된다면 소송과 관된 정보의 충분한 분류와 기술이 되어있다는 전제하에 소송과 연관 기록을 찾기 쉽도록 도울 것이다. 기록이 조직화되어 있고 계층별 기술이 되어있다면 덩어리별로 연관기록을 찾아내기 쉬워지기 때문에 더욱더 효과적일 것이다. 이는 Predictive Coding 도구 적용 시 선별 단계에서 시간・비용적 자원의 소모를 크게 줄일 것으로 예상된다. Predictive Coding 알고리즘 상 일차적인 선별을 거치고 나

〈표 7〉 소송 대응을 위한 기록관리 정책에 포함되어야 할 내용

- 기업의 사업 운영에 필수적인 기록과 민감한 정보의 처리
- 소송에 필요한 자료 보존에 대한 의무 준수(Compliance of Legal Holds)
- 소송의 보존 대상이 될 수 있는 임직원의 문서와 기록의 송신 및 전송 취급
- 정책과 관련된 임직원
- 기업 정보의 소유권
- 사생활과 의사소통에 관련된 기업 소유 장비의 사용
- 국제적으로 적용되는 보존 필요성과 저장 장소
- 자체적인 평가와 감사 기록
- 문서나 다른 정보의 적절한 보존과 유지에 대한 정책이나 지시의 실패로 인해 발생한 법률 미준수 사례

온 자원의 덩어리의 순도가 높을수록 더욱더 정확하고 민감한 특성 세트가 구성된다. 이는 정제된 전자적 정보 덩어리의 도출로 이어져 차후 변호사들의 최종적인 검토의 부담을 대폭 줄여주게 된다.

또, 처분일정에 따른 기록관리를 적용한다면 기업에 불필요한 전자적 정보들을 처분계획에 따라 삭제할 수 있기 때문에 효과적이다. 다만 무분별한 처분을 막기 위해 사안이 발생하였을 때 관련 기록을 찾아 처분을 보류할 수 있는 작업과 사안이 종료된 후 처분 보류를 해제하는 작업이 있어야 한다.

기존에 생산되어 있던 방대한 양의 전자적 정보들을 모두 관리하려면 오랜 시간과 천문학적인 비용이 필요할 것으로 예상되며 현실적으로 불가능해 보인다. 따라서 단기적으로는 기존에 생산되었던 전자적 정보들을 적절한 평가・폐기정책을 수립하여 관리하고 만약 소송이 발생한다면 기존의 전자증거개시대응 전략을 따라 TF팀을 꾸리고 Predictive Coding 도구를 이용하여 소송에 대응한다. 장기적으로는 기존의 생산되었던 전자적 정보들을 체계적으로 관리할 수 있는 기록관리 정책을 수립하고 기록관리 시스템을 구축해야 할 것이다. 이 때 역으로 Predictive Coding을 이용하여 방대한 양의 전자적 정보를 분류, 평가하는 도구로 이용할 수 있을 것으로 판단된다.

5. 맺음말

해외에서는 이미 전자증거개시를 통한 소송이 진행되고 있고 이에 따라 국내 기업들의 전자증거개시 대응의 필요성이 높아지고 있다. 국내 법조계에서도 소송 시 증거 제출을 위한 전자증거개시제도를 국내 실정에 맞게 수정하여 국내에도 도입할 것으로 보인다. 따라서 기업들은 전자증거개시제도에 대응할 전략을 세우고 준비해야 할 것이다.

기업들이 전자증거개시를 통한 소송에 승소하기 위해서는 제한된 시간 내에 적절한 비용으로 소송과 관련성이 있는 전자적 정보를 제출하는 것이 중요하다. 따라서 기업들은 소송을 위한 TF팀을 구성하고 증거개시를 도와줄 검색도구를 선정하는 것이 중요하다.

기존의 전자증거개시 과정에서 사용되고 있는 검색도구들은 대량의 전자적 정보들을 추려내는데 쉽고 효과적이지만, 단순히 기계적인 선

별에 불과하다는 한계점이 나타났다. 기업이 처리하는 전자적 정보의 양이 점점 늘어남에 따라 기존의 검색도구로는 비용 절감의 한계가 있다.

Predictive Coding은 전자증거개시 검토에 사용할 수 있는 검색도구들의 대안으로 떠올랐다. Predictive Coding은 기계학습을 이용하여 기업들이 보유하고 있는 전자적 정보들의 검토를 도와주는 도구이다. Predictive Coding은 기존의 검색 도구들 보다 효율성이 높고 잠재적으로 소송과 관련된 전자적 정보를 추려내는데 강점이 있는 것으로 나타났다.

다만, Predictive Cording 역시 프로그래밍 된 알고리즘에 의한 기계적인 검토에 불과하기 때문에 선별된 전자적 정보의 이차적인 수동적인 검토가 필요하다. 또, 현재 Predictive Coding 기술 상 텍스트로 구성된 전자적 정보 이외의 다른 유형의 전자적 정보인 동영상, 음성, 데이터베이스를 분석하는데 제한점이 있다. 마지막으로, 현재 사용되고 있는 Predictive Coding 알고리즘은 영어로 구성된 전자적 정보들을 선별하는데 특화되어 있다. 글자, 어순, 문장구조가 영어와는 다른 한글로 구성된 텍스트를 토큰화하여 적용하는데 기술적인 어려움이 따를 것으로 예상된다.

기업에서는 하루에도 수많은 전자적 정보들이 생산된다. 전자적 정보는 결제문서뿐만 아니라 데이터베이스, 이메일 등 모든 종류의 전자기록이 그 대상이 된다. 따라서 기업들은 기존의 전통적인 소송 대응 방식이 아니라 전자증거개시에 대비한 수집, 분류, 평가, 보존, 폐기 등의 각각 절차에 맞는 기업 기록관리 시스템 구축이 필요하다. 이를 통해 대량으로 발생하는 전자적 정보를 관리하여 잠재적으로 소송과 일차적으로 관련된 전자적 정보의 양을 대략적으로 줄인다면 소송에 들어가는 시간과 비용의 절감을 가져 올 수 있을 것이다. 다만 기업들이 하루에 생산되는 전자적 정보의 양을 볼 때 기존에 생산되었던 전자적 정보들을 모두 분류하고 평가하는 것은 불가능해 보인다. 따라서 소송이 발생할 때 기존에 생산되어 있던 전자적 정보들은 Predictive Coding 도구를 이용하여 검토하고, 향후 기업에 맞는 기록관리 정책의 수립과 기록관리 시스템을 구축하여 매일 생산되는 전자적 정보들을 관리해야 할 것이다.

기업의 효율적인 검색도구의 선택과 지속적인 기록관리를 통해 검토비용의 시간적, 비용적 절감을 꾀할 수 있을 것으로 예상된다. 따라서 기업이 전자증거개시제도에 대응하기 위해서 시간과 비용적 측면 모두를 고려한 전문적인 Predictive Coding 솔루션 도입과 지속적이고 장기적인 기록관리를 통해 가장 효과적인 방법을 모색해야 한다.

참 고 문 헌

김도훈 (2014). 미국 전자증거개시절차상 증거검색 및 수집방법에 관한 연구. 강원법학, 41(1), 217-252.
김승범 (2015). 기록관리의 기회와 위협요인으로서의 전자증거개시(E-Discovery)제도 연구. 석사학위

논문, 명지대학교 기록정보과학대학원, 기록관리전공.

김영수, 홍도원 (2011a). E-Discovery 대상 ESI의 컬링 성능 향상을 위한 핵심 기술. 한국통신학회 학술대회논문집, 650-651.

김영수, 홍도원 (2011b). E-Discovery 프로젝트: EDRM과 Sedona Conference. 주간기술동향, 1509, 14-27.

김일아 (2016). 전자증거개시(E-Discovery)에 대응하는 미국 기업의 기록관리 동향 분석. 석사학위논문, 명지대학교 기록정보과학대학원, 기록관리전공.

김종호 (2015). 세도나 캐나다 원칙상 전자증거개시제도의 준비에 관한 실무상의 문제점. 법학연구, 59, 147-183.

안정혜 (2010). 국제중재에서의 전자증거개시. 중재연구, 20(2), 67-90. https://doi.org/10.16998/jas.2010.20.2.67

이태림, 신상욱 (2012). 기업의 효과적인 소송 대응을 위한 전자증거개시 절차 모델과 대체 기술. 디지털융복합연구, 10(8), 287-297.

전복만, 박지훈 (2012). 지식재산분쟁에서 중재제도 활성화를 위한 전자증거개시제도의 정비. 과학기술법연구, 18(3), 267-402.

천우성, 박대우 (2011). e-Discovery 시스템 설계와 관리를 위한 인증과 암호화. 한국컴퓨터정보학회 학술발표논문집, 19(2), 139-142.

채은선 (2008). 디지털포렌식을 통한 E-Discovery의 실용화에 관한 연구. 석사학위논문, 동국대학교 대학원, 정보보호학과.

탁희성 (2011). 전자증거개시제도(E-Discovery)에 관한 연구. 서울: 한국형사정책연구원.

Acosta, A. M. (2012). Predictive coding: the beginning of a new e-discovery era. Res Gestae, 56, 8.

An ARMA International Publication (2012, November/December). Making 'predictive coding' pay needs cooperation. Information Management, 8.

An ARMA International Publication (2013, March/April). Predictive coding to become an ethical obligation. Information Management, 12.

An ARMA International Publication (2014, May/June). Predictive coding: Not just for E-Discovery. Information Management, 17.

Debra, L. (2008). Using the electronic discovery reference model to process, review and analyze digital evidence. Gartner Research. ID:G00159094

EDPB 공식홈페이지. Retrieved from http://www.edbp.com

EDRM 공식홈페이지. Retrieved from http://www.edrm.net

Schoenecker, E. Jr. (2015, April 4). Nine cases on predictive coding from modus. Retrieved from

https://www.linkedin.com/pulse/nine-cases-predictive-coding-from-modus-edward-schoenecker

E-Discovery Team (2008). Thoughts on SEARCH and Victor Stanley, Inc. v. Creative Pipe, Inc. Retrieved from https://e-discoveryteam.com/2008/06/08/hundredth-blog-thoughts-on-search-and-victor-stanley-inc-v-creative-pipe-inc/

E-Discovery Team (2013). My basic plan for document reviews: The "Bottom Line Driven" approach. Retrieved from https://e-discoveryteam.com/2013/10/01/my-basic-plan-for-document-reviews-the-bottom-line-driven-approach/

Issacs, L. (2013). Rolling the dice with predictive coding leveraging analytics technology for information governance. The Information Management Journal, 47(1), 23-26.

Volinino, L., & Redpath, I. (2009). E-Discovery for dummies. New Jersey: Wiley.

Pace, N. M., & Zakaras, L. (2012). Where the money goes: Understanding litigant expenditures for producing electronic discovery. Santa Monica, CA: Rand.

Swartz, N. (2006). New rules for E-Discovery. Information Management, 22-26.

[사안]

Dynamo Holdings Limited Partnership, et al., v. Commissioner of Internal Revenue, 143 T.C. No. 9 (Sept. 17, 2014)

EORHB, Inc., et al. v. HOA Holdings, LLC, C.A. No. 7409-VCL (Del. Ch. Oct. 15, 2012).

In re Actos (Pioglitazone) Product Liability Litigation, No. 611-md-2299 (W.D. La. July 27, 2012)

In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation, No. 3:12-MD-2391 (N.D. Ind. Aug. 21, 2013)

Moore v. Publicis Groupe SA, 2012 WL 1446534(S.D.N.Y. Apr. 26, 2012).

No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *3(S.D.N.Y. Feb. 24, 2012).

Rio Tinto PLC v. Vale, S.A., et al., No. 1:14-cv-03042-RMB-AJP (S.D.N.Y. Mar. 2, 2015)

• **국문 참고문헌에 대한 영문 표기**
(English translation of references written in Korean)

Ahn, Jeong-Hye (2010). Electronic discovery in international arbitration - Focusing on the establishment of rules regarding electronic discovery. Journal of Arbitration Studies, 20(2),

67-90. https://doi.org/10.16998/jas.2010.20.2.67

Chae, Eun-Sun (2008). A study on the practical use of E-Discovery through digital forensics. Master's thesis, Graduate School of Dongguk University, Department of Information Protection.

Chun, Woo-Sung, & Park, Dea-Woo (2011). Design of emergency response e-Discovery systems using encryption and authentication. Proceedings of the Korean Society of Computer Information Conference, 19(2), 139-142.

Jun, Bokman, & Park, Juhoon (2012). A study on maintenance of the E-Discovery for intellectual property disputes in the arbitration. Institute for Law of Science & Technology, 18(3), 267-402.

Kim, Do Hoon (2014). A study on the search and information retrieval methods in the U.S. E-discovery - Focusing on the technology-assisted review. Kangwon Law Review, 41(1), 217-252.

Kim, Il-a (2016). The trends analysis on the American business records management responding to E-Discovery. Master's thesis, Graduate School of Records, Archives & Information Science of Myongji University, Department of Records and Archival Information Management.

Kim, Jongho (2015). Some practical issues in the preparing for E-Discovery under the Sedona Canada Principles addressing electronic discovery. Law Review, 59, 147-183.

Kim, Seungbum (2015). An analysis on E-Discovery system as opportunities and threats for record management. Master's thesis, Graduate School of Records, Archives & Information Science of Myongji University, Department of Records and Archival Information Management.

Kim, Youngsoo, & Hong, Dowon (2011a). Core technologies for advancing curling performances of ESI concerning e-Discovery. Proceedings of Symposium of the Korean Institute of Communications and Information Sciences, 650-651.

Kim, Youngsoo, & Hong, Dowon (2011b). E-Discovery project: EDRM and Sedona Conference. Weekly Technology Trends, 1509, 14-27.

Lee, Tae-Rim, & Shin, Sang-uk (2012). E-Discovery process model and alternative technologies for an effective litigation response of the company. The Journal of Digital Policy & Management, 10(8), 287-297.

Tak, Heesung (2011). A study on E-Discovery. Seoul: Korean Institute of Criminology.