e-Discovery 솔루션의 성능 평가 방법*

이헌민†, 김헌†, 이태림†, 신상욱‡
† 부경대학교 대학원 정보보호학(협)
‡ 부경대학교 IT융합응용공학과
e-mail : galois8609@pknu.ac.kr

Methods of performance evaluation on e-Discovery

Heon-Min Lee†, Hun Kim†, Tae-Rim Lee†, Sang-Uk Shin‡
† Dept of Computer Engineering, Pukyoung Natoional University
‡ Dept of IT Convergence and Application Eng, Pukyoung Natoional University

요 약

e-Discovery 솔루션들의 성능 평가는 일반적으로 증거 확보의 측면에서 강조 되고 있으나, 소송 진행을 위해 추가적으로 제공 되어야 할 기능적인 요소들이 많다는 의견이 제기되며, 평가 기준의 확장 및 측정 할 수 있는 척도의 다양성이 요구 되고 있는 상황이다. 따라서 본 논문에서는 기존 e-Discovery 솔루션의 핵심이라 할 수 있는 증거 검색 기능 및 다양한 업무 절차에서 필수적으로 요구되는 외부적 요인을 결합하여, 객관적이며 신뢰성 있는 e-Discovery 솔루션의 성능 평가를 위한 절차를 제안한다.

1. 서론

최근 비즈니스 환경이 글로벌하게 변화되면서, 다양한 컴플라이언스 및 법 제도에 대한 기업들의 준수 여부가 필수 조건화 되고 있으며, 특히 미국의 경우 미국연방법원에서 사용되는 민사소송에 관한 규칙이 2006년 12월 1일 FRCP(Federal Rules of Civil Procedure) 개정으로, 소송 당사자가 공판 전, 공판을 준비하기 위해 법정 외에 법이 정한 방법에 의해 소송의 쟁점을 명확히 하는 정보 및 증거를 공개·수집하는 Discovery제도의 증거 범위가 디지털 형태의 전자 문서의 범위로 확장된 e-Discovery가 의무화 되고 있다.[1] 그러므로 미국을 대상으로 비즈니스를 수행하는 기업들은 e-Discovery 요구 사항에 맞는 업무 절차와 시스템을 의무적으로 도입해야 할 것이다. 또한 국내 민사 소송 역시 2000년 이후부터 계속 증가하고 있는 추세이며, 증거로서의 ESI (Electronic Stored Information) 중요성 부각 등으로 인해, 국내에도 e-Discovery 제도의 도입이 빠른 시일 내에 이루어질 전망이다. 하지만 현재 국내 대기업을 제외한 대다수의 기업들은 여전히 이에 대한 인식이 부족하여 관련 솔루션 도입이나 제도 정비 및 준비의 필요성을 크게 느끼지 못하고 있으며, 특히 소송에 대한 진행을 대부분 법률 전문 회사에 일임하는 경우가 많아, 소송을 위한 막대한 비용이 추가적으로 발생하는 실정이다.

이와 같은 문제의 해결책으로 e-Discovery 솔루션을 이용한 소송 관련 업무 수행은 시간과 비용을 절감해 주는 역할을 한다. FRCP[2]에 따르면 소송 당사자는 소송이 제기되는 날로부터 120일 내에 모든 증거 자료를 공개해야 하는데, 이는 당사자 간 소송 쟁점 및 증거 제출 양식에 대한 협의와 소송 일정을 확인하는 모든 절차가 포함 되어 있으므로, 실제 EDRM과 같은 절차를 적용하여 증거를 확보하는 작업에 소요 할 수 있는 시간이 길지 않다. 또한 가트너(Gartner) 보고에 따르면, 법률 전문가를 통한 1GB의 ESI 검토에 소모되는 비용이 $18,750 정도이며[3], 검토 대상 데이터의 규모가 나날이 증가하고 있는 상황을 고려해보았을 때, e-Discovery 솔루션의 증거로서 적합한 문서를 검색하는 기능은 시간 및 비용 소모를 절감시키기에 그 중요성이 증대되고 있다.

이와 같은 추세에 편승하여 포렌식 전문 업체, 데이터 마이닝 업체, ERP 업체 등 다양한 벤더들이 E-Discovery 시장에 뛰어 들어 많은 솔루션이 출시하고 있다. 하지만 각 솔루션들의 기능에 대한 정확성과 그 성능이 어느 정도인지를 가늠할 수 있는 측정 방법이나 도구는 전무한 상태이며, 다양한 e-Discovery 솔루션들 가운데 어떠한 제품이 자사의 업무 환경에 가장 적합하며, 우수한 성능을 보유하고 있는지 판별할 수 있는 기준 또한 전무하다. 이는 기업 경영인들로 하여금 제품 선택 결정을 모호하게 하는 가장 큰 원인이 되며, e-Discovery 솔루션을 개발하는 벤더들 역시 자사의 제품을 홍보하고자 할 때 객관적

* 본 논문은 지식경제부 산업융합원천기술개발사업[10035157, 실시간 분석을 위한 디지털포렌식 기술 개발]과 정부(교육과학기술부)의 재원으로 한국연구재단-차세대정보컴퓨팅기술개발 사업[No. 2011-0029927]의 지원을 받아 수행된 연구임.

인 지표가 존재하지 않기 때문에 신뢰성 있는 자료 마련에 어려움을 겪고 있다.

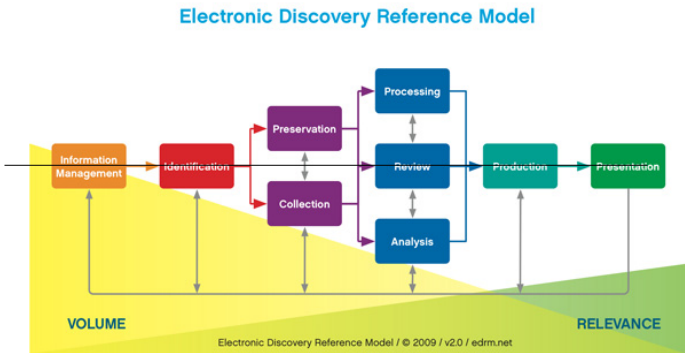
한편 정보 검색 도구의 성능 평가 기법을 연구 해 오던 TREC(Text REtrieval Conference)는 Legal Track을 통해 e-Discovery 솔루션의 성능 평가 기준 및 방안을 마련하기 위한 컨퍼런스를 개최하고 있다. 하지만 TREC의 평가 관점은 검색 기법에 국한 되어 있으며, 이외에도 소송 진행 및 e-Discovery 업무 절차 지원을 위해 제공 되어야 할 기능적인 요소들이 많다는 견해에 따라, 평가 기준의 확장에 대한 필요성이 대두 되고 있고 이를 객관적으로 측정할 수 있는 척도 역시 다양하게 요구되고 있는 상황이다.

따라서 본 논문에서는 EDRM의 Standard & Metircs project의 분석을 통해 e-Discovery 업무 절차를 지원하는 기능들만을 성능 평가 항목으로 취합하고, 일반적인 e-Discovery 솔루션의 핵심기능인 정보검색의 분석을 위해 TREC Legal Track 분석을 통해 보편화된 평가 기준을 마련하여, 일반적인 솔루션 성능에 대한 세부 절차 및 평가 방법을 제안한다.

2. 관련 연구

2.1 EDRM의 Standard & Metrics

[그림 1]의 EDRM(e-Discovery Reference Model)[4]은 현재 e-Discovery 솔루션 개발을 위하여 표준처럼 여겨지고 있는 모델이며, e-Discovery 절차 전반에 걸쳐 필수적인 업무 사항들을 기술하고 있다.



[그림 1] e-Discovery Reference Model

이외에도 EDRM에서 진행 중인 Standards and Metrics 프로젝트는 업무 평가에 활용 가능한 세부 항목과 평가 방법들을 기술하고 있다. 평가 대상 절차는 Identification, Collection, Processing, Review, Production의 5 단계로 축약되었으며, 이는 각 단계의 업무 성격 및 평가 대상의 특징에 따라 정량화 할 수 있는 수치 적용 평가가 가능한 Collection, Processing, Review는 Metric 프로젝트로, 그렇지 못한 나머지 2개의 절차는 평가를 위한 비교 대상이나 체크 항목과 같은 성향이 강한 Standards 프로젝트[5]로 구분되어 있다.

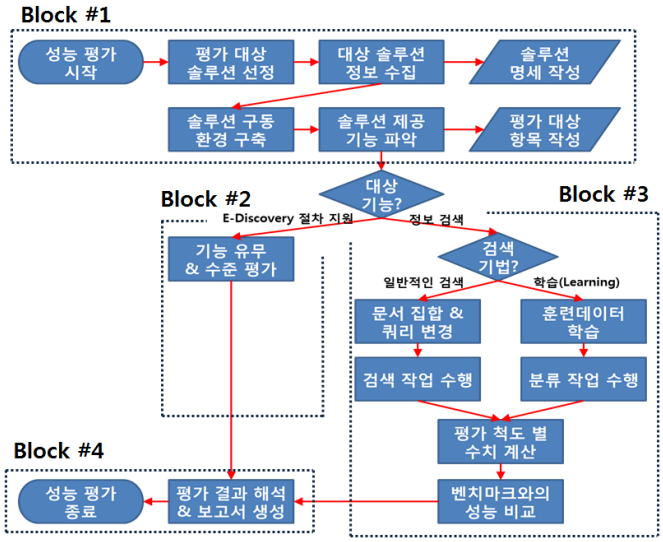
2.2 TREC Legal Track

TREC Legal Track[6]은 e-Discovery 솔루션의 성능 평가 기준 및 연구를 진행하고 있는 TREC의 한 프로젝트이다. 이것은 앞 서 언급한 EDRM의 평가 기준과는 대조적으로 오직 e-Discovery 솔루션의 핵심 기능이라 할 수 있는 정보 검색 기법에 대한 성능 평가 방법들을 제시하고 있다.

초기에는 소장에 대한 분석을 바탕으로 증거 데이터 확보를 위한 효과적인 부울 검색(Boolean search) 기법 개발을 목적으로 연구를 진행하였으며, 최근에는 관련 분야의 최신 기술로서 기계 학습(Machine Learning) 기반의 자동화된 문서 검색 기능에 대한 성능 평가 기법을 연구하고 있다. 2006년부터 연도 별 TREC Legal Track의 Task들이 평가 대상으로 삼은 검색 기법들은 e-Discovery의 실무에 가장 밀접한 연관성을 지니고 있는 부분으로 받아드릴 수 있으며, 매년 컨퍼런스 개최를 통해 벤치마킹을 위한 Task 참가 팀들의 실험 결과들을 공개 하고 있기 때문에 동일한 성능 평가의 목적을 가진 실험자들이 자유롭게 활용 가능하다.

3. e-Discovery 솔루션의 성능 평가 연구

본 논문에서 e-Discovery 솔루션 성능 평가를 위한 일반화된 절차를 제안한다. [그림2]에서 제시된 절차는 Block #1에서 #4까지 크게 4단계로 세분화되어 구성된다.



[그림 2] e-Discovery 솔루션 성능 평가를 위한 절차

▪ Block #1 : 솔루션 성능 평가를 위한 준비 단계

1. 평가 대상 솔루션 정보 수집 / 테스트 환경 구축

솔루션의 업무 성능 평가를 위해 평가 대상을 선정하고, 대상 이름, 제조사 등을 비롯한 솔루션의 정보들을 바탕으로 솔루션 명세를 작성한다. 이후 Test 환경 구축을 위한 플랫폼을 구성 및 구동을 위해 필요한 라이브러리 및 실행 파일등을 설치한다.

2. 평가 대상 항목 작성

솔루션 성능 평가를 위한 준비를 완료한 후, 사용 가이

드 및 매뉴얼, 화이트페이퍼, 직접적인 사용 등을 통해 솔루션이 제공하는 기능들을 파악하고, 파악된 결과를 토대로 평가 대상 항목을 작성한다.

▪ Block #2 : e-Discovery 절차 지원 기능 평가

솔루션의 기능이 e-Discovery 절차를 지원한다면, Block #1의 과정을 통해 파악한 솔루션의 기능들이 EDRM의 Identification에서 Production 절차에서 필요한 기능들의 존재 여부에 대한 성능 평가 대상으로 항목화 한다. 각 절차에 따른 도구평가 기준으로서 다음과 같은 항목들을 예상 할 수 있다.

[표 1] 평가를 위한 EDRM 절차 상 예상 가능한 항목들

절차	예상 가능한 항목들
Identification	(1) e-Discovery 업무 진행과 관련되는 사항의 기록 및 열람의 여부 (2) ECA[1]를 위한 기능의 제공 여부 (3) EDA[2]를 위한 절차의 제공 여부
Collection	(1) Collection 업무 수행과 관련된 사항의 기록 및 저장 여부 (2) 수집이 불가능한 작업들에 대한 오류를 보고 및 기록 여부 (3) 사전 수집 예측 양(검색 결과) 대비 실 수집 양(Collection 수행결과)에 대한 비교를 수행할 수 있는 수치 정보를 제공 여부
Processing	(1) Processing 업무 수행과 관련된 사항의 기록 및 저장 여부 (2) 처리가 불가능한 작업들에 대한 오류를 보고하고 기록 여부 (3) Collection 단계에서 전달된 데이터 셋 관련 정보를 기반으로 실 처리율에 대한 수치 정보를 제공 여부
Review	(1) Review 업무 수행과 관련된 사항의 기록 및 저장 여부 (2) 검토자들의 작업 관련 로그에 대한 통계적인 기록 여부 (3) 검토가 불가능한 작업 관련 오류를 보고 및 기록 여부 (4) 검토 작업의 효율성 평가를 위한 수치화 된 정보 제공 여부
Production	(1) 협의된 증거 제출 양식의 데이터 변환 기능 제공 여부 (2) 데이터 변환 후, 증거 검색 과정의 재현을 위한 솔루션 기능 적용 가능 여부 (3) 원본에 대한 무결성의 증명 제공 여부

위의 [표 1]에 대한 체크리스트를 작성하여, 제시된 e-Discovery의 각 단계 별 평가 절차에 대한 정보를 각 단계에 따라 모두 취합한다. 취합된 정보들을 통해 선정한 솔루션이 e-Discovery 업무 절차에 부합하는 정도에 따라 상/중/하의 단계로 구분하여 평가한다.

▪ Block #3 : 정보검색 지원 기능 평가

솔루션 기능이 정보검색을 지원한다면, 제공하는 기법에 따라 부울 검색 기법 혹은 기계 학습 검색 기법을 제공하기 때문에 Block #1에서 파악된 솔루션 기능에 따라서 평가를 진행 한다.

솔루션이 부울 검색 기법에 대한 평가를 TREC Legal Track의 2006년부터 2009년까지의 각 Task들에 대한 적용을 통한 벤치마킹 할 수 있다. 2006년부터 2009년까지의 각 년도 별 Task들은 소송내용을 기반으로 작성된 부울 쿼리들을 활용하는 검색 기법의 맥락에서 진행이 되었기 때문에, 이를 솔루션의 부울 검색 기법 평가에 대한 벤치마킹 자료로서 활용 가능하다. 모든 Task는 TREC에서 제공하는 문서 집합인 IIT CDIP v1.0 test Collection[7]을 사용하고, 소장의 Topic 별 주요 키워드에 대한 쿼리문을 사용해 부울 검색을 진행하기 때문에, 솔루션 평가를 위해 동일한 문서 집합을 사용하고, 주어진 쿼리문을 솔루션이 제공하는 검색 방법으로 적절하게 변환하여 평가를 진행하여 적합 문서를 찾는다.

솔루션의 정보검색 기능이 기계학습 기법을 지원한다면, Legal Track의 2010년의 Task에 대한 적용을 통한 벤치마킹을 할 수 있다. 기계 학습을 위해 법률 전문가들에 의한 사전평가를 통해 관련성이 있다고 판단된 seed set을 이용하여, 기계 학습을 하고 EnRon 문서 컬렉션(EDRM EnRon v2 Collection)[8]을 사용해 적합 문서를 찾는다.

두 검색 기법의 결과로 얻은 소장의 Topic에 대한 쿼리 결과인 적합 문서와 비 적합 문서 수를 얻을 수 있으므로, 결과에 대하여 정보 검색에서 사용되는 대표적인 정보검색의 평가방법인 정확률, 재현률, F1척도를 계산한다.[9]

- $\text{정확률}(Pricision) = \frac{\text{도구를통해검색된적합문서수}}{\text{도구를통해검색된전체문서수}}$
- $\text{재현률}(recall) = \frac{\text{도구를통해검색된적합문서수}}{\text{주어진전체적합문서수}}$
- $F1\text{척도} = \frac{2\times(\text{정확률}\times\text{재현률})}{(\text{정확률}+\text{재현률})}$

계산된 수치를 이용하여 TREC Legal Track에서 공개한 결과들과 벤치마킹하여 솔루션의 검색 기법에 대한 평가를 진행 한다. 상호비교를 위해 2006년부터 2009년까지의 결과는 각 참가 팀 당 Topic에 대한 검색 결과의 정확률, 재현률, F1척도 평균치 등과 모든 참가 팀의 각 Topic에 대한 검색 결과의 정확률, 재현률, F1척도 평균치 등이

1) ECA(Early Case Assessment) : EDA 뿐만 아니라 소송 전반에 관한 대응 전략을 수립하기 위해 다양한 정보를 획득하는 절차

2) EDA (Early Data Assessment) : Data에 대한 형태와 저장된 위치 등과 같은 정보 요구에 대한 분석이 수행되는 절차

주어진다. 또한 기계 학습 검색 기법의 경우 2010년과 2011년의 각 참가팀의 각 Topic 별 F1척도와 AUC 수치를 결과로서 제공된다. 공통되는 수치 정보인 정확률, 재현률, F1척도를 이용하여 상호비교를 할 수 있다. 다음의 [표 2]는 부울 검색 기법을 이용한 벤치마크 데이터[10], [표 3]은 기계학습을 이용한 벤치마크 데이터[11]의 예시이다.

[표 2] TREC 2006 Task 의 각 팀의 보고 결과

실험 명	Topics	Retrieved	Relevant	Rel-ret	Precision@B	MAP	R-Prec
humL06t	39	111724	4323	2168	0.0523	**0.1109**	**0.1612**
NUSCHUA1	39	180295	4323	1087	0.0134	0.0264	0.0559
SabLeg06aa1	39	194996	4323	2174	**0.0916**	0.1071	0.1575
UmdComb	39	**195000**	4323	1025	0.0661	0.0560	0.0757
UMKCB	39	145847	4323	1762	0.0422	0.0671	0.1116
york06la01	39	**195000**	4323	**2621**	0.0655	0.1031	0.1115
평균 수치	39	170477	4323	1086	0.0552	0.07843	0.1122

[표 3] TREC 2010 Learning Task 의 각 팀의 보고 결과

Topic & R 실험명	200 R : 2544	201 1886	202 6312	203 3125	204 6362	205 67938	206 866	207 20929
BckBigA	F1 : 18.9 AUC : 78.0	40.7 84.8	56.6 82.7	36.4 78.8	15.2 79.7	47.3 81.2	3.4 83.7	81.8 87.4
DUTHIrgA	8.9 82.7	11.6 90.4	32.1 81.0	24.2 92.9	8.2 86.8	51.4 87.3	7.6 91.3	16.7 60.3
otL10bT	25.8 67.6	23.1 91.1	32.8 84.1	32.9 87.9	8.1 52.2	47.3 71.0	37.0 97.0	22.9 72.8
rmitindA	15.3 81.6	13.0 91.2	37.8 86.8	32.2 94.3	17.1 87.0	52.1 87.9	5.9 87.6	24.3 27.0
xrceCalA	19.1 73.8	34.5 94.1	70.6 97.1	39.4 92.3	26.6 77.6	45.9 76.8	14.3 84.8	90.3 96.6
평균	17.6 76.7	24.6 90.3	46.0 86.3	33.0 89.2	15.0 76.7	48.8 80.8	13.6 88.9	47.2 68.8

- Block #4 : 솔루션의 평가 결과 해석 및 보고서 생성

Block #1부터 #3까지의 솔루션 평가를 거치며 작성된 솔루션 명세, 평가 대상 항목, 벤치마킹을 통한 성능 비교를 종합하여 보고서를 작성하고, 작성된 보고서를 통하여 솔루션이 e-Discovery 업무 절차 및 정보 검색 능력에 대한 성능에 대한 분석을 하고 성능 평가를 종료한다.

5. 결론 및 향후 연구

본 논문에서는 기존 e-Discovery 솔루션의 핵심이라 할 수 있는 증거 검색 기능 및 다양한 업무 절차에서 필수적으로 요구되는 외부적 요인을 결합하여, 객관적이며 신뢰성 있는 e-Discovery 솔루션의 성능 평가를 위한 절차를 제안하였다. 이를 위하여 EDRM의 Standard & Metrics와 TREC Legal Track 분석을 통해, e-Discovery 솔루션의 성능을 기능별로 테스트 할 수 있는 평가 항목 및 시나리오를 마련하였으며, 여러 솔루션들 간 상호 비교가 가능한 벤치마크 데이터와 평가 결과 해석 방안을 제시 하였다.

향후 각 과정의 정밀한 세분화와 평가 척도에 대한 심화적인 통계적 기법을 통해 제시한 솔루션 성능 평가에 대한 정밀도 및 통계학적인 분석의 기능 추가와 제안한 평가 절차에 대한 구현을 통해 실제 e-Discovery 솔루션들에 대한 실험 및 분석 할 예정이며, 솔루션을 선택 할 때 기준이 될 수 있는 근거가 될 수 있도록 모든 내용들을 데이터베이스로 구축을 할 예정이다. 또한 EDRM의 Standards와 Metrics 프로젝트를 보다 더 심화하여 평가 항목을 더 엄밀화시킬 예정이다.

참고문헌

[1] 김영수, 신상욱, 홍도원, “ESI 관점에서의 e-Discovery”, 정보통신산업진흥원 주간 기술 동향, 2010

[2] Federal Rules of Civil Procedure(FRCP), http://www.law.cornell.edu/rules/frcp

[3] Linda Volonio and Ian J. Redpath, e-Discovery For Dummies, Wiley Publishing, Inc., Indianapolis, Indiana, 2010

[4] EDRM Framwork Guides, http://www.edrm.net/resources/guides/edrm-framework-guides

[5] EDRM Metrics & Standars project, http://www.edrm.net/resources/standards

[6] TREC Legal Track, http://trec-legal.umiacs.umd.edu/

[7] IIT CDIP v1.0 test Collection, https://ir.nist.gov/cdip/

[8] EnRon v2 Collection, http://trec-legal.umiacs.umd.edu/corpora/trec/legal10/

[9] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, 최신 정보검색론, 교보문고; 2010

[10] TREC 2006 Legal Track, http://trec.nist.gov/pubs/trec15/appendices/legal.results.html

[11] TREC 2010 Legal Track Learning Task Result, http://trec.nist.gov/pubs/trec19/appendices/legal-learning.html