A Query Recommending Scheme for an Efficient Evidence Search in e-Discovery

Heon-min Lee*, Su-bin Han*, Taerim Lee*, Sang Uk Shin**

*Dept. of Information Security the Graduate School, Pukyong National University, Republic of Korea

**Dept. of IT Convergence and Application Eng., Pukyong National University, Republic of Korea

galois8609@pknu.ac.kr, subin4853@naver.com, taeri@pknu.ac.kr, shinsu@pknu.ac.kr

*Abstract***— In recent years, the importance of e-Discovery is being strongly emphasized according to the rapid increase of litigation between the business corporations. The success of e-Discovery depends on how well the litigant and lawyer search relevant evidence, and it is closely associated with making fine queries based on their analysis of complaint and data set. Therefore, this paper proposes a Query Recommending Scheme called QRS for an efficient evidence search in e-Discovery procedure. This scheme is composed with four different phases and various techniques are applied such as document parsing, machine learning and scoring. We describe how QRS works using the flow chart and introduce further researches for the improvement of QRS.**

*Keywords***— Query Recommending, Machine Learning, Electronic Discovery, e-Discovery, Evidence Search**

I. Introduction

In the spring of 2011, Apple began litigating against Samsung in patent infringement suits, while Apple and Motorola Mobility were already engaged in a patent war on several fronts. This fight between the two big companies has continued on without ceasing until now, so it has sparked global attention and this becomes a popular example for explaining the current trends of international disputes or litigation. In addition, Apple's multinational litigation over technology patents became known as part of the mobile device patent wars: extensive litigation in fierce competition in the global market for consumer mobile communications[1]. Apart from such cases related patent, the number of lawsuits is rapidly increasing among business corporations due to the conflict of their interest. As a result, e-Discovery will continue to play important roles in solving this kind of situation.

Electronic discovery (or e-Discovery, eDiscovery) refers to discovery in civil litigation which deals with the exchange of information in electronic format called ESI (Electronically Stored Information). This legal system was the subject of amendments to the Federal Rules of Civil Procedure (FRCP), effective December 1, 2006, as amended to December 1, 2010[2]. In general, most litigants depend heavily on lawyers for dealing with a series of work required in the litigation process, but it costs a large amount of money. Moreover, e-Discovery makes every litigant have a responsibility to produce their own evidence for themselves, so the use of digital forensic or e-Discovery tools becomes a necessary[3]. In the end, the most important thing is whether or not the litigant can find relevant evidence to prove his legitimacy for trial.

The keywords for evidence search are primarily suggested by a lawyer from the analysis of complaint. However, those keywords are usually ambiguous and complicated because the lawyer quoted the contents of complaint without changes. Although there are many reasons that impel the lawyer to do so, the most serious problem is that he cannot know the detail information of litigant's data from the beginning. Misuse of keyword makes poor results of evidence search, so it will lower the efficiency of entire e-Discovery work. Of course, this problem can be solved by the steady counselling or the expert's advice, but additional cost and time will be required.

This paper, therefore, proposes a Query Recommending Scheme shortly called QRS for an efficient evidence search in e-Discovery procedure. QRS creates initial queries by extracting the primary keywords from a complaint like lawyer did and makes some samples for machine learning. Using the samples, it classifies the litigant's data into relevant and non-relevant group. After that, it collects meaningful information from the relevant group and generates extended queries for recommending. To do this, this paper describes the analysis about document format of complaint and the workflow of QRS. Utilization for QRS and future work will be introduced in the final.

II. Related Work

The work for e-Discovery is generally performed by jurists and IT experts who are collaborating with each other[3]. Recently, a great deal of research is being carried out to improve the work efficiency, and most of them are about the standardization or a development of search techniques. Here are representative researches closely associated with this paper.

A. Electronic Discovery Reference Model

Figure 1 shows the Electronic Discovery Reference Model[4] usually called EDRM, and it was designed for providing essential requirements of e-Discovery work to the people concerned. There are various stages in EDRM, but QRS focus on the Identification because information gathering for the next evidence search is performed in this stage. The basic object of QRS is the preparing proper queries for search, and it is the key to the success of e-Discovery.

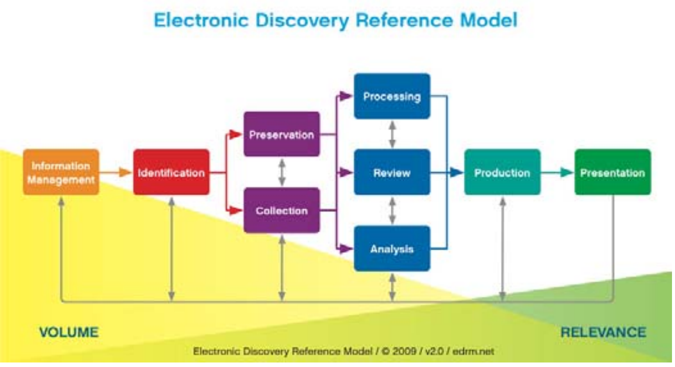

Figure 1. The workflow of Query Recommeding Method

B. TREC Legal Track

The Text REtrieval Conference (TREC) is an on-going series of workshops focusing on a list of different information retrieval (IR) research areas, or tracks, and the learning task has been conducted in legal track from 2010. The goal of this task is to determine which documents (email messages or attachments, treated separately) should be produced in response to a production request for which a set of "training" relevance judgments are available[5]. This is a kind of experiment to find the best way to use a machine learning algorithms for evidence search.

QRS should find potentially meaningful information from the data set given by the litigant, so it requires a special data mining skills for correlation analysis with complaint. Machine Learning can be used as a proper method if the considerations for e-Discovery could be applied well.

C. Machine Learning

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. The core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory[6].

QRS should be designed by considering all requirements to use this technique and to prove its validity of generalization method. Also, additional studies should be followed to confirm what type of algorithm is more suitable for QRS.

III. The Main Considerations of Complaint

A plaintiff starts a civil action by filing a pleading called a complaint. A complaint must state all of the plaintiff's claims against the defendant, and must also specify what remedy the plaintiff wants[7]. This means that main issues of e-Discovery are almost included in this document, so a series of keywords for evidence search are created by the lawyer's review. This is very common way to take action against the lawsuit and to prepare the ensuing e-Discovery work, but it has a high-cost and low-efficiency problem because of the complete dependence on a specific person like a lawyer. For example, let's suppose that the lawyer gives a sentence as a query: "Find all documents about legal violation of tobacco advertising by the company A" which was extracted from the allegation part of original complaint. Using some keywords in this query as it is, the search results will be poor because hands-on workers could be used their own name for advertising project or specific name of product like Marlboro instead of tobacco on site. This is not a problem which can be solved easily by just using an approximate search. As a result, the special method should be required to get this kind of useful information for search from the litigant's data set.

First of all, the most pressing matter is to decide on how to deal with complaint to extract initial query without lawyer. This is not a simple because there are many different kinds of complaint format due to the litigation types or jurisdictions. Therefore, based on the analysis of complaint examples provided by Legal Information Institute (LII) in Cornell University Law School[7] and TREC Legal Track[5], the common components of complaint were organized as follows.

TABLE 1. The Common Components of Complaint

Title of Paragraph		Notes
LII	TREC	
N/A (Caption or Heading)		Outline (Jurisdictions, Plaintiff, Defendant, Type of action)
Preliminary Statement	Nature of the action	Summary of the causes for demand trial
Jurisdiction and Venue		Reason why the case should be heard in the selected court rather than some other court
General Allegations	Plaintiff's Allegations	List of facts that brought the case to the court
	Substantive Allegations	
Count	Cause of Action	A numbered list of legal allegations, with specific details about application of the governing law to the each court
Demand for relief		The relief that plaintiff is seeking as a result of the lawsuit

Among the items of Table 1, the useful information to make queries for e-Discovery is the part associated with the allegations. In some cases, the other parts could be more important, but we considered only this part to focus on finding evidence for the proof of facts.

IV. The Proposed Scheme for Query Recommending

This section describes the proposed scheme of query recommending for evidence search. As above mentioned, the search results only using the information in complaint could be ineffective, so queries will be expanded by using the analysis result of litigant's whole data. The denoted notations, functions and workflow for this method are as follows. This method extracts some information first to make initial queries, and gives scores to each terms of query.

TABLE 2. NOTATIONS

Notation	Description
t_i	A *i*-th term in text
$T = \{t_1, t_2, \ldots t_n\}$	A general text with *n*-terms
C	A complaint
$S = \{T_1, T_2, \ldots T_n\}$	A sample with a set of *n*-texts
$D = \{T_1, T_2, \ldots T_n\}$	A set of entire litigant's data
$R = \{T_1, T_2, \ldots T_n\}$	A set of relevant texts in D
$NR = \{T_1, T_2, \ldots T_n\}$	A set of non-relevant texts in D
wt_i	A weighted value of *i*-th term
$RWT_i = \{wt_1, wt_2, \ldots wt_n\}$	A *i*-th text with a set of *n*-weighted values of terms in R
wt_{avg}	An average of weighted values
$RT_i = \{t_1, t_2, \ldots t_n\}$	A *i*-th set of recommended terms for query expansion
$Q = \{t_1, t_2, \ldots t_n\}$	A query with a set of *n*-terms
$WQ = \{wt_1, wt_2, \ldots wt_n\}$	A query with a set of *n*-weighted terms
Q_i	A *i*-th query
$Q_{init} = \{Q_1, Q_2, \ldots Q_n\}$	A set of initial queries
$Q_{rmd} = \{Q_1, Q_2, \ldots Q_n\}$	A set of recommended queries

TABLE 3. FUNCTIONS

Function	Description
$GQ : C \rightarrow Q_{init}$	Generate initial queries based on the contents of allegation from complaint
$SCR : Q_i \rightarrow WQ_i$	Calculate weighted values of each term in *i*-th query
$SCH(Q_i) : D \rightarrow S_i$	Make samples for machine learning by using the search result of each Q
$ML(S_i) : D \rightarrow R, NR$	Classify the D into R or NR according to the likelihood with S
$SEL(RWT_i, WQ_i) : RWT_i \rightarrow RT_i$	Calculate wt_{avg} in Q_i, Remove same terms between RWT_i and WQ_i, Select terms having a bigger weighted value than wt_{avg} in RWT_i
$E(RT_i) : Q_{init} \rightarrow Q_{rmd}$	Make Q_{rmd} using the combinations of RT_i and Q_i in Q_{init}

Figure 2 shows the workflow of query recommending scheme based on the notations and functions in Table 2, 3. This workflow includes 4 phases, and each phase has own input and output according to its purpose.

A. *Pre-Processing*

The contents of allegation are extracted from the original complaint in this step. To do this, QRS should perform a series of tasks such as text parsing, structure analysis, tokenizing and filtering. So, it is very similar to the general document parser for indexing tools.

B. *Sampling*

This phase is to make samples from the search results produced by initial queries. These samples will be used as the training sets in Data Mining phase. In order to generate initial queries, QRS should be able to perform a syntax analysis about extracted contents of allegation because each allegation is a long sentence in general. For example:

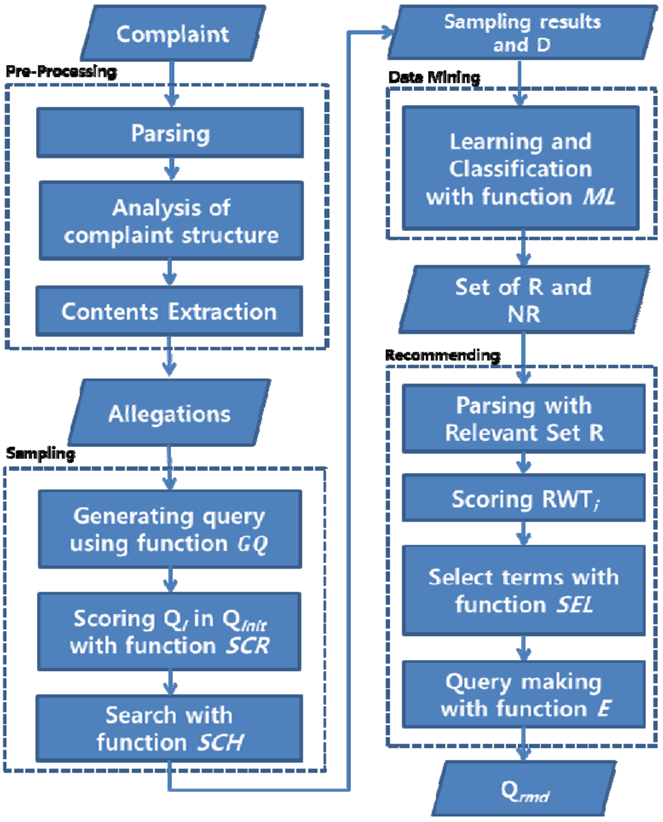
Complaint
Pre-Processing
Parsing
Analysis of complaint structure
Contents Extraction
Allegations
Sampling
Generating query using function GQ
Scoring Q_i in Q_{init} with function SCR
Search with function SCH
Sampling results and D
Data Mining
Learning and Classification with function ML
Set of R and NR
Recommending
Parsing with Relevant Set R
Scoring RWT_i
Select terms with function SEL
Query making with function E
Q_{rmd}

Figure 2. The workflow of Query Recommeding Scheme

- The case of copyright infringement: The defendant A plagiarized the plaintiff B's song, "I'm angry" to make a title song, "I'm sorry" in his debut album which was released on January 1, 2013.
- The expected terms for query: the name of A, B, debut album, "I'm angry", "I'm sorry", the release date, etc.

Also, QRS should calculate weighted values for comparing the importance of terms in Recommending phase. TF-iDF is one of the easiest functions for scoring.

C. *Data Mining*

In this phase, all documents in D (see the Table 2) are divided into two groups, relevant set of R and non-relevant set of NR. Machine learning algorithm like SVM can be used for this. After training from the collected texts in Sampling, it starts to classify each document according to its likelihood. When the classification is completed, QRS only take care of documents belong to a set of R for next phase.

D. *Recommending*

As QRS did in the previous phases, it starts parsing and scoring with document set of R to make different sets of RWT_i (see the Table 2). This *i* is the number of queries used for search. Separately, it also calculates each wt_{avg} of Q_i in Q_{init}. After that, QRS can select the potentially useful terms for query expansion through the following procedure.

1) De-Duplication: For expansion, QRS does not have to consider all terms included in Q_i because they are already used for search. So, it changes a set of RWT_i by $RWT_i – (RTW_i \cap Q_i)$.

2) Selection: Using the changed RTW_i, QRS selects the new terms satisfied that $wt_i > wt_{avg}$ of all elements in RTW_i. At this time, all selected new terms become the elements of set RT_i in Table 2.

This procedure shows the roles of function *SEL* in Table 3, and the only thing left to do is to make new queries by the combinations of RT_i with Q_i in Q_{init}. The name of this function is *E*, and it makes the final result of Q_{rmd} for query recommendation.

V. Utilization and Expected Effect of QRS

The fundamental object of QRS is reducing the cost and improving the efficiency of e-Discovery work by lessening the dependence on lawyers. To achieve this, the most important duty of QRS is changing queries to be more suitable for the characteristics of litigant's data set. It leads to avoid the bad outcomes of evidence search when the litigant uses some queries generated by the lawyer as they are because the lawyer cannot know the details of litigant's data set or situation from the first. Above this, here are other examples for the utilization of QRS.

- For the EDRM Identification: In this stage, there are two major tasks for developing overall e-Discovery strategies, Early Case Assessment (ECA) and Early Data Assessment (EDA)[4]. These tasks are usually performed by human activity for the analysis of complaint and litigant's data set. QRS can automatically deal with this kind of work in real time.
- For the meet-and-confer: Before the trial, parties must "meet-and-confer" to try to resolve the matter or at least determine the points of conflict. This has the beneficial effect of making the lawyer and clients face up to the realities of their positions[8]. All the information produced by QRS (ex, about whether securing important evidence is feasible or not) let them to foresee who win or lose in trial, so they can actively negotiate from a position of strength.
- For an appeal or a similar case in future: QRS uses a machine learning algorithm for extracting the meaningful information. If the litigant can preserve some important evidence in previous cases, it can be used again by QRS when a similar case happens.

VI. Conclusions and Future Work

This paper proposed a Query Recommending Scheme for an efficient evidence search in e-Discovery. In the field of digital investigation like e-Discovery, information retrieval is the most basic and frequently used technology. As the application fields of this technology are diversified, although a lot of latest requirements are already upon us, the accuracy improvement of retrieval still remains one of the principal challenges. There exist various recommending techniques to solve this kind of problems, but most of them are based on the user's feedback and not suitable for e-Discovery because it requires comparatively a lot of time. The proposal of QRS has a meaning in new attempts to satisfy those two requirements at the same time, but there are many things to do.

First of all, the implementation of QRS and proper experiments should be done to evaluate its availability. After that, additional researches for finding a better functions or factors such as *GQ*, *SEL*, *E* or wt_{avg} of QRS should be followed in order to improve the accuracy of search result.

Acknowledgment

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (No.2011-0029927).

References

[1] Apple Inc. V. Samsung Electronics Co., Ltd., Wikipedia, 2013

[2] L. Smith et al., Federal Rules of Civil Procedure. U.S. Government Printing Office. [Online]. Available: http://www.uscourts.gov/uscourts/rules/civil-procedure.pdf, Aug. 2013.

[3] T. Lee, H. Kim, K. H. Rhee and S. U. Shin, *Design and Implementation of e-Discovery as a Service based on Cloud Computing*, Computer Science and Information Systems. Serbia: ComSIS Consortium, 2013, vol. 10 (2). pp. 703—724.

[4] (2013) The EDRM website. [Online]. Available: http://www.edrm.net/

[5] (2013) The TREC website. [Online]. Available: http://trec.nist.gov/

[6] F. Sebastiani, Machine Learning in Automated Text Categorization, Journal of ACM Computing Surveys. 2002, vol. 34. no. 1. pp. 1—47.

[7] Complaint Overview, Legal Information Institute in Cornell University Law School, 2010.

[8] L. Volonino and I. Redpath, *e-Discovery For Dummies*, John Wiley & Sons Inc. New York, USA, 2009.

Heon-min Lee received his B.S. degree in Major of Computer and Multimedia Engineering and applied mathematics from Pukyong National University, Busan, Korea in 2013. He is currently pursuing his master's degree in Department of Information Security, Graduate School, Pukyong National University. His research interests include machine learning, data-mining and e-Discovery.

Su-bin Han received his B.S. degree in Major of IT Convergence and Application Engineering from Pukyong National University, Busan, Korea in 2013. She is currently pursuing his master's degree in Department of Information Security, Graduate School, Pukyong National University. Her research interests include digital forensics, cloud computing, and e-Discovery.

Taerim Lee received his Bachelor and Master of Engineering degrees from Pukyong National University, Busan Korea in 2008 and 2010, respectively. He is currently doing a Ph.D. program in Department of Information Security, Graduate School, Pukyong National University. His research interests include digital forensics, e-Discovery, cloud computing, and machine learning.

Sang Uk Shin received his M.S. and Ph.D. degrees from Pukyong National University, Busan, Korea in 1997 and 2000, respectively. He worked as a senior researcher in Electronics and Telecommunications Research Institute, Daejeon Korea from 2000 to 2003. He is currently an associate professor in Department of IT Convergence and Application Engineering, Pukyong National University. His research interests include digital forensics, e-Discovery, cryptographic protocol, mobile/wireless network security and multimedia content security.