Evaluation of Query Formulations in the Negotiated Query Refinement Process of Legal e-Discovery: UMKC at TREC 2007 Legal Track

Feng C. Zhao, Yugyung Lee, Deep Medhi
School of Computing and Engineering
University of Missouri-Kansas City
{fczrzf,leeyu,dmedhi}@umkc.edu

Abstract—**UMKC participated in the 2007 legal track. Our experiments focused mainly on evaluating the different query formulations in the negotiated query refinement process of legal e-discovery. For our study, we considered three sets of paired runs in vector space model and language model respectively. Our experiments indicated that although the Boolean query negotiating process was successful for the standard Boolean retrieval model, it did not make statistically significant query improvements in our ranked systems. This result provided us an insight into the query negotiation process and a new direction to refine queries.**

I. Introduction

Electronically stored information has gained substantial standing during trial-preparation and litigation in recent years. However, the development of the corresponding legal e-discovery methodology and underlying engineering has not gained the same momentum. The efficacy of the legal e-discovery process fundamentally started from the initial query formulation and negotiated query refinement known as the development of a search protocol, even well before applying any information retrieval strategies and techniques.

There are four main stages in the negotiated query refinement process of legal e-discovery. First, the plaintiff states the objectives of the request for the production of documents as the legal evidences in the requested text (RequestText). Second, the defendant devises the initial query (ProposalByDefendant) from the requested text. Third, the plaintiff presents a counterproposal (RejoinderByPlaintiff) with usually more complex queries. Finally, both parties negotiate an agreement on the final query string (FinalQuery). In this escalating process, broad query terms (such as synonyms) and Boolean constrains (such as proximity) are added to the query.

The essential purpose of query refinement through negotiations in legal e-discovery is to improve the recall, which is a measure of the ability of a system to present all relevant items. Therefore, the final negotiated query is supposed to find more relevant documents than what was initially proposed by the defendant. Unfortunately, the introduced ambiguity or over-broad scope in queries may lead to the contrary. This has motivated our investigation of the dynamics of the search protocols development by comparing the performance among different query formulations.

II. Evaluation

A. Measures

We are interested in the relative performance improvement among different query formulations during the negotiation process. Traditionally, the mean average precision (MAP) and many other robust measures are common performance measures for information retrieval systems. For recall-oriented measures, the 2007 legal track further extended the concept of the inferred average precision (infAP) [4] by incorporating the deep pooling in order to obtain estimated recalls (est_R) and estimated precisions (est_P). In our experiments, we focused on the MAP and the estimated recall at B (est_RB) and estimated precision at B (est_PB), where B is the number of documents matching the final negotiated boolean query.

B. Methods

The most resources demanding component of evaluation is to obtain the relevance judgment. Pooling is the status quo in TREC to obtain the relevance judgment in which only a small percentage of documents in the pool is judged. But the traditional pooling method has its biases and limits [1] and is insufficient for the challenge of large corpus in the 2006 legal track [2] [3]. Hence, the 2007 legal track uses a deep pooling method to obtain estimated recalls and estimated precisions.

However, the above evaluation methods are impractical to evaluate query formulation in a production environment because the true relevance judgment can not be a prerequisite at the query formulation stage. Since our investigation is only interested in the relative superiority of query formulations and the absolute performance measurements are a secondary issue, this emphasis shift made it possible to set up a contingent run as the pseudo relevance judgment. The relative standing of different query formulations can be estimated if this relevance proxy is close enough to the true relevance judgement.

Our experiments compare the relative performance of different query formulations before the actual relevance assessment by using the reference Boolean run in the 2007 legal track as the pseudo relevance judgment. We hypothesize a well performed Boolean run can serve the similar role of the human relevance judgment to a certain extent based on the following observations. First, the result of the 2006 legal track showed that the reference Boolean run found 57% of the known relevant documents [2], and it is one of the top performing runs. Second, although the Boolean run is well known to contain many irrelevant documents, it can still be expected to produce a reasonable est_RB and est_PB as these measures are designed to cope with incomplete and imperfect relevance judgment. Finally, we will partially verify this hypothesis by comparing the query performances obtained from the pseudo relevance judgment to the human relevance judgment.

C. Systems

The information retrieval system that we use in this experiment is a modified Lucene search engine [5] on a Cray XD1 system in the Arctic Region Supercomputing Center. Both the vector space model (VSM) and the language model (LM) were implemented, which represent algebraic models and probabilistic models respectively [9]. The primary advantage of utilizing two representative retrieval models is to mitigate the potential bias in any one particular model during the evaluation. The key relevance measurement in the VSM of query q for document d correlates to the cosine-distance or the dot-product between document and query vectors; its formula is explained in the Lucene book [5]. In language modeling approaches for information retrieval [6], we estimate a language model for each document and then rank documents according to their likelihood of generating the query. For a collection C, document d, query q, term t, term frequency of t in d ($\mathrm{tf}_{t,d}$) and document frequency of t in C ($\mathrm{df}_{t,C}$), the language modeling formula in this system is given by the equation 1. The Jelinek-Mercer smoothing parameter λ holds the responsibility for a linear interpolation of the maximum likelihood estimation in document language model and the collection language model.

$$P(d|q) = P(d) \prod_{t \in q} (\lambda P(t|d) + (1-\lambda) P(t|C)) \quad (1)$$

where
$P(d) = \frac{|d|}{\sum_{d' \in C} |d'|}$
$P(t|d) = \mathrm{tf}_{t,d}$
$P(t|C) = \frac{\mathrm{df}_{t,C}}{\sum_{t' \in C} \mathrm{df}_{t',C}}$.

The most significant technique we utilized is the query expansion model based on the conceptual relevance framework [7]. Conceptual relevant concepts are expanded into a query based on its query centroid. The query is expanded before the initial search, so there is no relevance feedback required. But since this query expansion process does not observe the full Boolean syntax and simply concatenates every query term with an OR operator, information of phrase and proximity is lost from original Boolean queries.

D. Indexes

We indexed only the OCR text portion and its document number from the IIT CDIP test collection. Porter stemmer was invoked, but its potential was hindered by the numerous OCR errors. A customized stop words list of 1,236 items was used to reduce the index size and to clean the OCR error. We then crafted the most of the two-letter permutations into this stop words list, and they are counted as almost half of the list. All of the above efforts are mounted at a common goal to create a centralized index with a manageable size.

E. Runs

There are six runs designed as stated in Table II-E; essentially, three paired runs utilize the various query fields in two different retrieval models. They are labeled UMKC$_1$, UMKC$_2$, and so on. Although the RejoinderByPlaintiff is different with the FinalQuery, they do not significantly diverge. Therefore, we do not show the RejoinderByPlaintiff as a separate query genre in this table.

TABLE I
SIX SUBMITTED RUNS WITH THEIR QUERY SOURCES AND RETRIEVAL MODELS.

Runs	**Query Source**	**Retrieval Model**
UMKC$_1$	ProposalByDefendant	LM
UMKC$_2$	RequestText	LM
UMKC$_3$	FinalQuery	LM
UMKC$_4$	ProposalByDefendant	VSM
UMKC$_5$	RequestText	VSM
UMKC$_6$	FinalQuery	VSM

F. Queries

The final query strings of paired runs are the same for both retrieval models if they share the same query source. The actual query string generated from the query expansion model is quite distinguishable from its query source text. As an example, for request number 56, the query source of UMKC$_2$ run is:

RequestText: Please produce any and all documents concerning soil water management as it pertains to commercial irrigation.

And its final query string contains a list of weighted relevant terms:

irrig (0.3084472), soil (0.25898176), water (0.2516427), pertain (0.20087002), commerci (0.1465618), tobacco (0.08702624), cigarett (0.03782016), plant (0.037246022), product (0.031287868), smoke (0.029167147) ...

In the above final query string, those functional terms in the requested text are automatically eliminated due to their marginal weight and other relevant terms have been expanded into the query. The connotation of a term being relevant to the query is local to this particular IIT CDIP test collection and may not strictly correspond to our common sense.

G. Results

The six runs were evaluated through the l07_eval program where both the traditional measures and the 2007 legal track specific measures were produced. As every document in the reference Boolean run is assumed to be selected and judged as relevant, the probability of including document d in the judging sample is one. Hence, $p(d) = 1.0$ was added to the reference Boolean run to accommodate the l07_eval program.

In the TREC environment, it has been suggested that the t-Test significance coupled with at least a 10% relative difference in MAP between two runs is significant [8]. Therefore, the absolute performance measures, including MAP, est_RB, and est_PB, are shown in Table II. Both the relative difference on MAP and the two-tail P-values from the paired t-Test of MAP, est_RB, and est_PB are shown in Table III, where the notion of (m, n) indicates a comparison between UMKC$_m$ and UMKC$_n$. A particular query formulation is compared with not only the other query formulations using the same retrieval model but also its corresponding run in the other retrieval model.

There are two observations from the above results:

- *Observation-1*: there is no statistical significant improvement at the 0.05 level among different runs regardless of the query sources.
- *Observation-2*: the language model generally outperforms the vector space model.

The corresponding performance evaluation tables after the human relevance judgment being obtained are shown in Tables IV and V. In the case of using the human relevance judgment, observation-1 still holds except for the P-value of est_RB between UMKC$_1$ and UMKC$_2$. But the observation-2 is reversed in that the vector space model actually outperforms the language model, especially when queries are derived from the FinalQuery field.

TABLE II
PERFORMANCE MEASURES WITH PSEUDO RELEVANCE JUDGMENT

Measures	**UMKC$_1$**	**UMKC$_2$**	**UMKC$_3$**	**UMKC$_4$**	**UMKC$_5$**	**UMKC$_6$**
MAP	0.1731	0.1486	0.1412	0.1386	0.1253	0.1253
est_RB	0.2333	0.2166	0.2091	0.2129	0.1966	0.2010
est_PB	0.9302	0.9767	0.9767	0.9535	0.9767	0.9767

TABLE III
PERFORMANCE COMPARISONS WITH PSEUDO RELEVANCE JUDGMENT

Measures	**(1,2)**	**(2,3)**	**(1,3)**	**(4,5)**	**(5,6)**	**(4,6)**	**(1,4)**	**(2,5)**	**(3,6)**
MAP Diff %	0.1649	0.0523	0.2259	0.1060	0.0004	0.1065	0.2482	0.1851	0.1266
MAP P-value	0.0576	0.7374	0.1866	0.2110	0.9975	0.4782	***0.0002***	***0.0039***	***0.0202***
est_RB P-value	0.1507	0.7299	0.3171	0.1769	0.8065	0.5403	***0.0148***	***0.0073***	0.2364
est_PB P-value	0.1597	n/a	0.1597	0.3230	n/a	0.3230	0.3230	n/a	n/a

TABLE IV
PERFORMANCE MEASURES WITH HUMAN RELEVANCE JUDGMENT

Measures	**UMKC$_1$**	**UMKC$_2$**	**UMKC$_3$**	**UMKC$_4$**	**UMKC$_5$**	**UMKC$_6$**
MAP	0.0940	0.0906	0.0842	0.1029	0.0987	0.1050
est_RB	0.1351	0.1003	0.1065	0.1571	0.1259	0.1371
est_PB	0.2410	0.2262	0.2426	0.2528	0.2597	0.2580

TABLE V
PERFORMANCE COMPARISONS WITH HUMAN RELEVANCE JUDGMENT

Measures	**(1,2)**	**(2,3)**	**(1,3)**	**(4,5)**	**(5,6)**	**(4,6)**	**(1,4)**	**(2,5)**	**(3,6)**
MAP Diff %	0.0367	0.0701	0.1042	0.0405	-0.0638	-0.0206	-0.0861	-0.0824	-0.1980
MAP P-value	0.6805	0.4628	0.3652	0.6700	0.6046	0.8658	0.3895	0.3805	***0.0151***
est_RB P-value	***0.0159***	0.7284	0.0812	0.0559	0.5634	0.3072	0.2016	***0.0297***	***0.0329***
est_PB P-value	0.6043	0.5390	0.9444	0.8191	0.9551	0.8508	0.3075	***0.0238***	0.2093

III. DISCUSSIONS

There is generally no statistically significant performance difference among different query formulations, regardless of whether we latch on to the pseudo relevance judgment or the human relevance judgement. Therefore, the pseudo relevance judgement can be justified to practitioners as an economical mean to compare the query formulations. However, such justification is largely dependent on the actual performance of the reference Boolean run. In other words, we have to choose a sound reference Boolean run to evaluate query formulations. On the other side, the available performance measures between the vector space model and the language model do not yield any significant conclusion in this experiment. They simply indicate that the language model performs more closely to the reference Boolean run, whereas the vector space model performs more closely to the human judgement. Hence, we need additional and different retrieval systems in order to verify whether the pseudo relevance judgement can also be used to compare different retrieval systems.

The results indicate the deficiency in the current Boolean query negotiation process as the negotiation has not improved retrieval performance, at least in our ranked systems. The utility of using the reference Boolean run as the pseudo relevance judgement also supports the particular finding in the 2006 legal track that the effectiveness of the Boolean query system is compatible with the effectiveness of the best ranked retrieval systems [2]. But we have to realize that the Boolean query refinement process is specifically intended for Boolean retrieval systems, and that may be the chief advantage of the reference Boolean run. Table VI shows the definite statistically significant improvement when the Boolean queries move from ProposalByDefendant to either RejoinderByPlaintiff or FinalQuery. But generally, there is no further improvement from RejoinderByPlaintiff to FinalQuery and this is consistent with our previous observation that there is only little difference between these two queries.

Thus, what made the current query refinement

TABLE VI
PERFORMANCE COMPARISONS WITH HUMAN RELEVANCE JUDGMENT FOR REFERENCE BOOLEAN RUNS

Measures	**Defendant (D)**	**Plaintiff (P)**	**Final (F)**
est_RB	0.0272	0.1863	0.2158
est_PB	0.0264	0.2349	0.2921
	(D, P)	**(P, F)**	**(D,F)**
est_RB Diff%	5.8487	0.1582	6.9327
est_PB Diff%	7.9096	0.2436	10.0802
est_RB P-value	***1.29E-05***	0.3277	***1.34E-05***
est_PB P-value	***3.96E-07***	***0.0087***	***1.86E-09***

process successful to the standard Boolean system, and how may we adapt it for the benefit of the ranked systems as well? The primary techniques used in the Boolean refinement process are enriching the query with synonym-like terms and relaxing Boolean constraints. Request number 65 and 71 are typical examples of where the above two techniques are applied.

- *RequestNumber*: 65
  ProposalByDefendant: candy w/5 (packag! OR label! OR wrapper!)
  RejoinderByPlaintiff: Candy AND (pack! OR label! OR wrap! OR adverti! OR box OR ingredient! OR contain!)
- *RequestNumber*: 71
  ProposalByDefendant: bromhidrosis
  RejoinderByPlaintiff: bromhidrosis OR ((body OR human OR person) AND odor!))

The intention of performing the above two techniques is to increase the recall; as a matter of fact, this objective is well achieved in the standard Boolean query system. But in our ranked systems, we discarded the Boolean constraints and leveled query coverage through query expansion; hence, both techniques lost their thrust. From another point of view, the infertility of the Boolean query negotiation revealed the fact that the negotiation is ineffective to discover and inject semantically independent terms into queries. In other words, the negotiated final query essentially has the same semantic coverage as what was initially proposed by the defendant after we drop all the Boolean syntax.

Interestingly, among all 70 submitted main task runs in the 2007 legal track, the UMKC$_5$ run has the highest estimated precision at depth of 25,000 (est_P25000), and the UMKC$_2$ run has the highest estimated relevance retrieved measure (est_rel_ret). Both UMKC$_5$ and UMKC$_2$ are using the RequestText only, and they are paired runs in the vector space model and the language model respectively. If we choose to view the Boolean negotiation process as a kind of manual query expansion, then the above results indicate that the unsupervised query expansion model we deployed is more effective than the manual query refinement in terms of retrieving more relevant documents at the 25,000 level–the designated depth of the 2007 legal track.

In order for the ranked systems to take further advantage of the query refinement process, we suggest enriching the legal query with new concepts which are pertinent to the overall query intention in the RequestText but located in some other semantic dimensions. From the perspective of information retrieval, it will be more helpful to our ranked systems if the query negotiators simply identify a solid list of core concepts. From the perspective of e-discovery, it is desirable to limit the query scope to avoid unduly burdensome or expensive discovery requests [10]. Therefore, we should investigate the criteria which qualifies a term to be considered for query refinement and the effect of a term on the overall query performance. Furthermore, in the absence of insights of underlying information retrieval engines, the emphasis of the query negotiators should focus on identifying those basic core concepts, rather than expanding the query according to the needs of the standard Boolean retrieval system. Additionally, from a broader perspective of e-discovery, we need to negotiate the information retrieval system beyond the e-discovery queries.

Finally, in the review of our system design, there are still several techniques that may be added to improve its performance. For example, sentence boundary detection and phrase identification can be utilized during conceptual relevance building, and full Boolean syntax can be observed during query expansion.

IV. CONCLUSION

Our experiments indicate that although the Boolean query negotiating process was successful for the standard Boolean retrieval model, it did not make statistically significant query improvements in our ranked systems. This implies a new challenge to the legal query negotiator, who has to discover new semantically independent query terms during negotiation. We also found that the utility of the

Boolean run as the pseudo relevance judgment can serve as a potential economical mean to evaluate and direct the query refinement process without the expensive human relevance judgment. As the query negotiators usually are uncertain with the nature of the underlying information retrieval system, we further propose a direction for the query refinement in legal e-discovery, which is to shift from amplifying a particular query term to identifying the core concepts pertaining to the overall query intention.

ACKNOWLEDGEMENT

We thank Christopher Fallen of the Arctic Region Supercomputing Center for generously facilitating the set up of the modified Lucene search engine. The Boolean run results listed in Table VI are provided by Dr. Stephen Tomlinson of Open Text Corporation.

REFERENCES

[1] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, "Bias and the limits of pooling," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval Seattle, Washington, USA: ACM Press, 2006.

[2] J. R. Baron, D. D. Lewis and D. W. Oard. TREC-2006 Legal Track Overview. In Proceedings of TREC 2006.

[3] S. Tomlinson. "Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track,". In Proceedings of TREC 2006.

[4] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in Proceedings of the 15th ACM international conference on Information and knowledge management Arlington, Virginia, USA: ACM Press, 2006.

[5] O. Gospodnetic and E. Hatcher, Lucene in action. Greenwich, CT: Manning Publications, 2005.

[6] D. Hiemstra, "Using Language Models for Information Retrieval," in Center for Telematics and Information Technology. vol. Ph.D.: University of Twente, 2001.

[7] F. C. Zhao, Y. Lee and D. Medhi, "Experiments with Query Expansion at TREC 2006 Legal Track". In Proceedings of TREC 2006.

[8] M. Sanderson and J. Zobel, "Information retrieval system evaluation: effort, sensitivity, and reliability," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval Salvador, Brazil: ACM Press, 2005.

[9] D. A. Grossman and O. Frieder, Information retrieval : algorithms and heuristics, 2nd ed. Dordrecht ; Great Britain: Springer, 2004.

[10] Daniel B. Garrie and Matthew J. Armstrong, Electronic Discovery and the Challenge Posed by the Sarbanes-Oxley Act, UCLA Journal of Law and Technology. 2, 2005