Modeling Concept and Context to Improve Performance in eDiscovery

By: H. S. Hyman, ABD, University of South Florida
Warren Fridy III, MS, Fridy Enterprises

Abstract

One condition of eDiscovery making it unique from other, more routine forms of IR is that all documents retrieved are settled by human inspection. Automated IR tools are used to reduce the size of a corpus search space to produce smaller sets of documents to be reviewed. However, a limitation associated with automated tools is they mainly employ statistical use of search terms that can result in poor performance when measured by recall and precision. One reason for this limitation is that relevance -- the quality of matching a document to user criteria -- is dynamic and fluid, whereas a query -- representing the translation of a user's IR goal – is fixed. This paper reports on a design approach to eDiscovey that combines concept and context modeling to enhance search term performance. We apply this approach to the TREC 2011 Legal Track Problem Set #401. Our goal is to improve performance in eDiscovery IR results.

Keywords: Information Retrieval, Text Mining, eDiscovery, Concept-Based Modeling, Context-Based Modeling.

Introduction

Electronic discovery (eDiscovery) is the process of retrieving ESI documents for the purpose of review for anticipated or actual litigation (Oard et al. 2010). Four characteristics describe eDiscovery as a form of IR. First, eDiscovery is defined by a user who has domain knowledge about the nature of context, content, and concepts associated with the documents being sought (Grossman and Cormack, 2011). The Second characteristic of eDiscovery is the premium placed on recall over precision. The reason for this is that discovery rules take a much more harsh view of failure to disclose versus disclosing too much. The principal legal case on this issue is *Zubalake v. UBS*. The third characteristic of eDiscovery lies in the sheer volume of information to be reviewed by a human inspector prior to release (Baron, 2005). The fourth characteristic is common to all forms of high volume IR: *uncertainty* (Bates 1986). In the domain of eDiscovery, documents being sought have particular meaning and context (Oard et al, 2010; Grossman and Cormack, 2011). Performance from the use of IR tools can vary widely for a given set of search terms and weighting methods.

The problem we address in this paper is: How can we retrieve what we are looking for (recall) and leave the rest behind (precision)? In answering this question we propose an approach to model three constructs: (1) *Concepts* underlying the fixed search terms and queries, (2) *Context* of the domain and the corpus, and (3) *Elimination terms* used as counter-measures for reduction of non-relevant documents. We translate these constructs as algorithm operational calls: Document Format, Content Characteristic, and Elimination Terms. Our goal is to improve IR performance in eDiscovery by modeling (1) *internal* concepts underlying the structure of the documents sought and (2) *external* context of the nature of the corpus.

We use the components of document format and content characteristic to improve recall by extending the user's concepts beyond query search terms. We use the component of elimination terms to improve precision by reducing the number of non-relevant documents in the retrieval set through specific filtering conditions. We apply this model to eDiscovery tasks to address limitations associated with techniques such as search terms, weighting, and conditional probability matching.

Motivation

"In an era where vast amounts of electronic information is available for review, discovery in certain cases has become increasingly complex and expensive" (Judge Shira Scheindlin, 2010). The Tobacco cases and Enron litigation are specific examples of this. The Philip Morris litigation had as many

as 32 million emails in the overall corpus. TREC's version of the Enron corpus is based on the EDRM version 2. It contains approximately 650,000 to 680,000 email objects, depending on how one accounts for attachments. The Illinois Institute of Technology's version of the Tobacco corpus contains between 1 and 2 million objects. Jason Baron points out that in the case of *U.S. v. Philip Morris*, 25 individuals spent 6 months reviewing 200,000 emails, one at a time (Baron, 2005). The limitations defining this problem space are: time, volume and heuristics.

The solution to the problem of volume is automation. After all, if the user could review every document then there would not be a problem to solve. Automation requires the use of heuristics to produce outcomes. The imperfect nature of heuristics is the subject of significant IR research. The reason for this is that an automated process is built on creating abstractions to substitute for the human reviewer. In the case of IR, we have users create queries of search terms to represent the concepts they believe distinguish a relevant document from a non-relevant document. A relevant document is one that matches the query. The difficulty of matching the query to a document has to do with uncertainty (Bates 1979, 1986). Uncertainty exists primarily due to the imperfect nature of attempting to reduce fluid concepts defining relevance, to a fixed collection of search terms. Some documents are *not* relevant but *do* contain search terms; some documents *are* relevant but do *not* contain search terms (Garron and Kontostathis, 2010). The imperfect nature of matching search terms to documents is often explained by polysemy – multiple meanings within the same word, and synonymy – multiple words with the same meaning (Deerwester, et al., 1990). Therefore, a relevant document may or may not match the query.

In this paper, we focus on the *goal* of relevancy. Vanrijsbergen, 1979, sees the goal of IR as providing users with relevant documents – such that the documents match the user's query. We adapt Vanrijsbergen's goal of IR to our goal of relevancy. In this study we focus on matching the user's mental model of a desired group of documents to the collection of documents retrieved -- relevancy. The gap between the user's mental model and the fixed query of search terms is what we address in this paper through the use of concept and context modeling, by focusing on concepts within documents and context within the corpus.

Modeling Relevance and Dealing with Uncertainty

Relevance is central to information retrieval (Park, 1993). The query is a collection of terms that abstracts the user's meaning. Document representation has been identified as a key component in IR (Vanrijsbergen 1979). There is a need to represent the content of a document in terms of its meaning.

Uncertainty describes the problem of matching the meaning of the user and the meaning of the document (Giger, 1988). The difficulty is that, *predicted* relevance by a tool is based on conditions and probabilities, whereas *actual* relevancy is defined by the human reviewer. The question becomes how to model relevancy such that the intent of the user is translated to the automated tool.

There are various methods available to model relevancy of a document. Luhn is among the first researchers to have suggested indexing documents by term vectors (Luhn 1957; Salton and Buckley, 1988). Lenk and Floyd experimented with using Bayesian probabilistic indexing in 1988. Latent Semantic Indexing has been used as a method to distinguish between the language variances of polysemy and synonymy (Deerwester, et al., 1990). April Kontostathis and William Pottenger have experimented with using Singular Value Decomposition (SVD) in Latent Semantic Indexing (LSI). Their 2005 paper reported on the first study of LSI using term by dimension vectors.

In 1996, Jiang and Conrath proposed using concept based retrieval as a method of "semantic equivalence between a user's query and the retrieved item." Grossman and Cormack identify the nature of context, content and concepts as constructs associated with document retrieval. We model these three constructs in this study and test our model using TREC 2011 Legal Track Problem #401.

Modeling Context and Concept

This paper reports on our model designed to represent context, concept and elimination terms, in combination with traditional search terms and weights. The results reported in this paper are based on the initial F1 results released by the TREC Legal Track 2011.

Users assume dependencies between concepts and expected document structures, whereas tools use process statistical and probabilistic measures of terms within documents (Giger, 1988). Concept based IR has been used to describe the problem of representing the meaning of a document's content (Vanrijsbergen, 1979). Clustering techniques have been used to model concepts within documents addressing the limitation of using search terms alone (Runkler and Bezdek, 1999, 2003). Clusters of documents can be used to describe the *concepts* within a group of documents based on characteristics within the cluster. A specific example of a concept based approach is the use of fuzzy logic (Ousallah et al., 2008) to model meaning – a fluid concept often elusive to define.

Context and content-based modeling has been identified in circumstances under which user impact and context are highly correlated to text-based searches (Chi-Ren et al., 2007). Chi-Ren et al.,

2007, applied a content approach to support geospatial IR systems. Brisboa, et al., 2009, based their indexing approach on text references and ontology for geographic IR systems. Trembley et al., 2009 and Jarmon, 2011, have used ontology methods for medical IR systems. These approaches to the above domains share two common factors associated with eDiscovery: (1) Domains highly impacted by users who are often domain experts, and (2) Term based queries alone typically result in imprecise translation of the information need. The idea of filtering methods using rules to reduce a retrieval collection has been applied to email boxes and other situations where the user is on the receiving end of push data (Singhal, 2001). There is presently little research focused on modeling concept, context and filtering, in combination with traditional search terms and weighting methods.

Our model is designed to accomplish two goals. The first is to express the concepts within the eDiscovery request (Problem set #401) and match those concepts to the context defined within the corpus (Enron data set) to improve recall of documents that might otherwise failed to meet a term based query. The second goal is to improve precision by using a filtering method we call Elimination terms to reduce the number of non-relevant documents that otherwise would meet a term based query. The main objective is to improve performance in the retrieval. In last year's paper we focused on recall. In this paper we focus on precision. Figure-1 is a description of our algorithm and process. Figure-2 is a description of our job run for Problem set #401. Figure-3 is our design for combining concept and context based modeling with a traditional search term and weighting scheme.

Our Experiment

We began by making certain assumptions about what we perceived as the "concepts" contained within Problem set #401. An example of this is one of our hypotheses that; proper nouns and descriptive nouns referring to '*enrononline'* represent a specific concept that should produce a relevant document to a high degree of certainty. This hypothesis was supported. The feedback from our document run on this concept indicated we achieved a 70% precision rate; quite an improvement from our 20% precision in last year's run.

Once we identified concepts within the request, we used thesaurus and dictionary applications to amplify the terms we chose as models of the concepts. We categorized the concepts in three types: Document format, Content characteristic and Elimination component. The first two categories are an attempt to improve recall of relevant documents in situations where the document may not contain the

relevant search terms. The third category, Elimination component is designed to exclude non-relevant documents that otherwise would have been included because they contain relevant search terms.

We selected a random pull of 1000 documents to train our model. Within that set of documents we randomly pulled sets of 100 documents. We ran alternative concept models for each 100 document set. We combined the results and re-ran them on the 1000 document set and then on our entire collection of 680,000 objects.

We developed a scoring method to report our results to TREC. We based the document scores on two dimensions. First, we looked at the per word occurrence level. Second, we looked at the per word impact level. We found that certain words contained within a document had a higher significance of correlation with relevance than other words. We called this the *impact* of the word.

Discussion

The feedback we received from our submission is that our model produced an F1-measure of 49.8%. This is a tremendous improvement over last year's performance of 19%. We find this improvement encouraging and plan further experiments with our model using active participants.

We found difficulties with a significant percentage of objects within the corpus. For instance, we had trouble attempting to extract the text from email attachments containing power points. We also had some trouble with using OCR to covert some of the attachments to readable text. There were also a significant number of locked files that we were never able to extract. Other difficulties were: incompatible formats with some attachments, emails with no content having attachments only, attachments that were pasted versus native text, additional processing to prepare the data for extraction, embedded attachments to links that no longer exist, and password protected files.

In terms of the documents themselves, we found a positive correlation between the number of occurrences of a word, and the combinations of words, with relevance. We also found consistent distributions of words and occurrences throughout the collection. For example, we found that the term 'EOL' occurred within approximately 2.5 to 3 percent of the documents in the collection. It turned out that 'EOL' was a high impact word highly correlated with relevance. We also found that some terms having high co-occurrences were in fact overlapping terms. Those documents needed to be scored lower to prevent them from floating to the top falsely. Examples of such terms are '*financial*' and '*commodity*.' These two terms turned out to have a high degree of overlap, and in fact, were often overtaken by

higher impact terms such as 'EOL' and '*swaps*.' Other terms that one would expect to have a high correlation with relevance such as '*derivative'* turned out to have a low significance.

As a result of the above observations, we set a baseline for document scores using impact and occurrences as metrics. A higher impact word received a base level higher than the maximum occurrence of the next highest impact word. This created an order to the documents based on their characteristics and not simply based on the occurrence of words alone. This also allowed low impact word documents to leap over higher impact word documents if the word occurrences justified the re-ordering. Accounting for co-occurrences of words allowed documents with high occurrences between words to float to the top of the list, without preventing high impact word documents from rising to the top, if they outscored occurrence word documents.

Conclusion

The goal of our experiment was to model the concepts contained within the request (Problem set #401), and the context of the corpus (Enron collection). We used concepts and context in combination with traditional search terms and weights to produce higher recall and precision. Our initial results support our hypothesis that combined concept and context modeling improves IR over search terms and weights alone. We find the initial results encouraging and plan to continue our experiments in this area by studying live user interaction with our algorithm using interface screens to capture user activity.

FIGURES

Figure 1

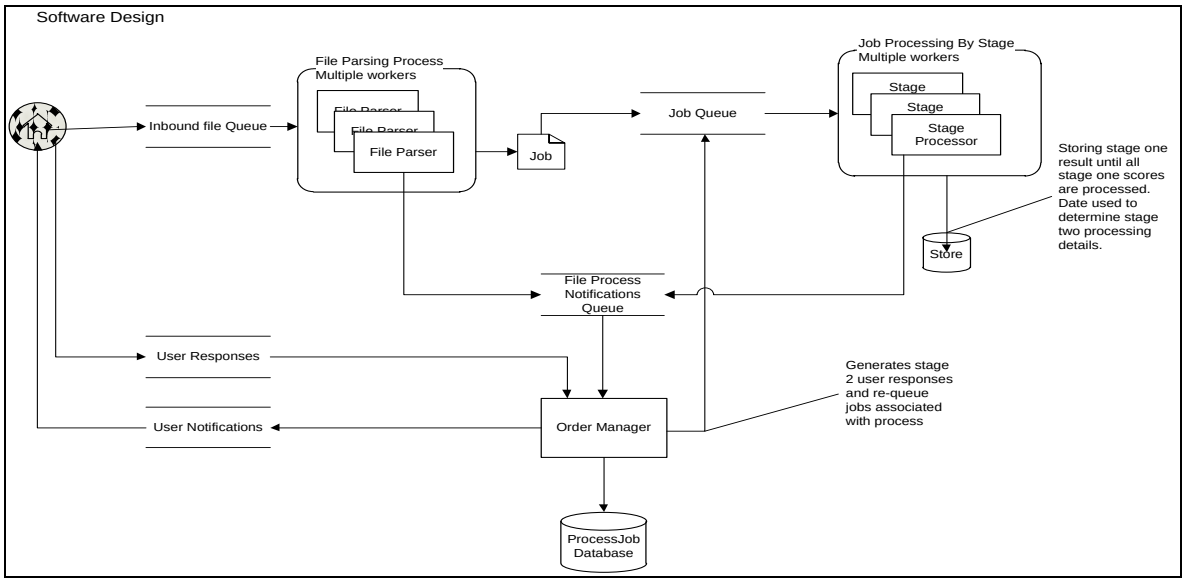


Figure 2

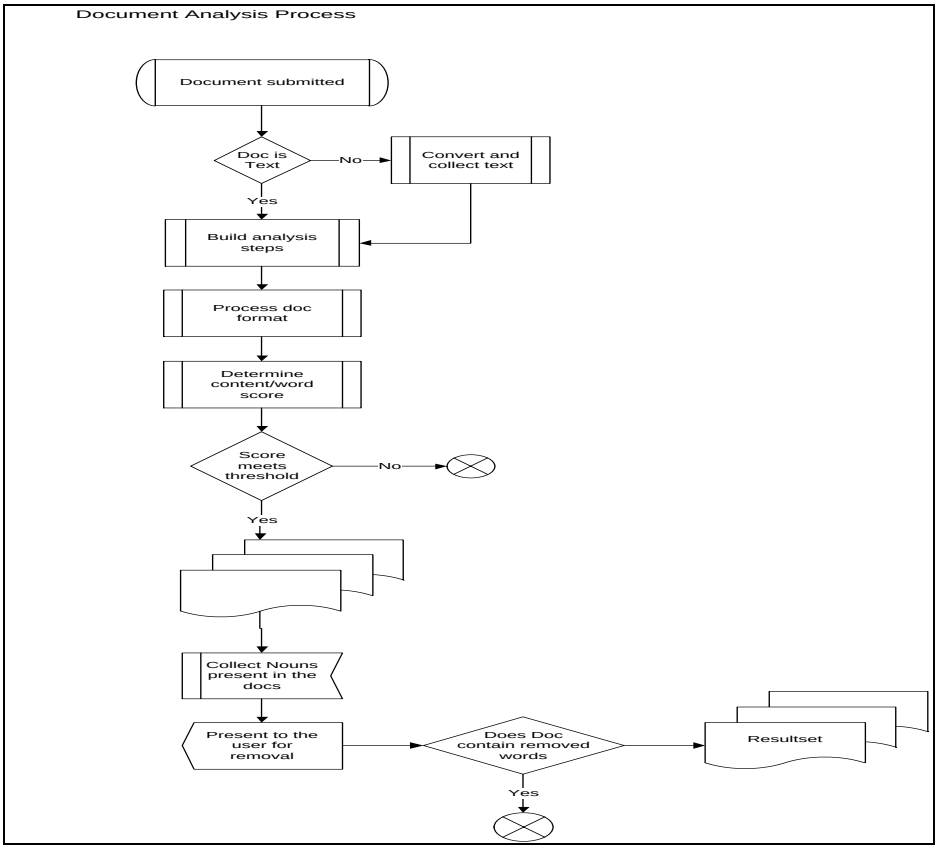
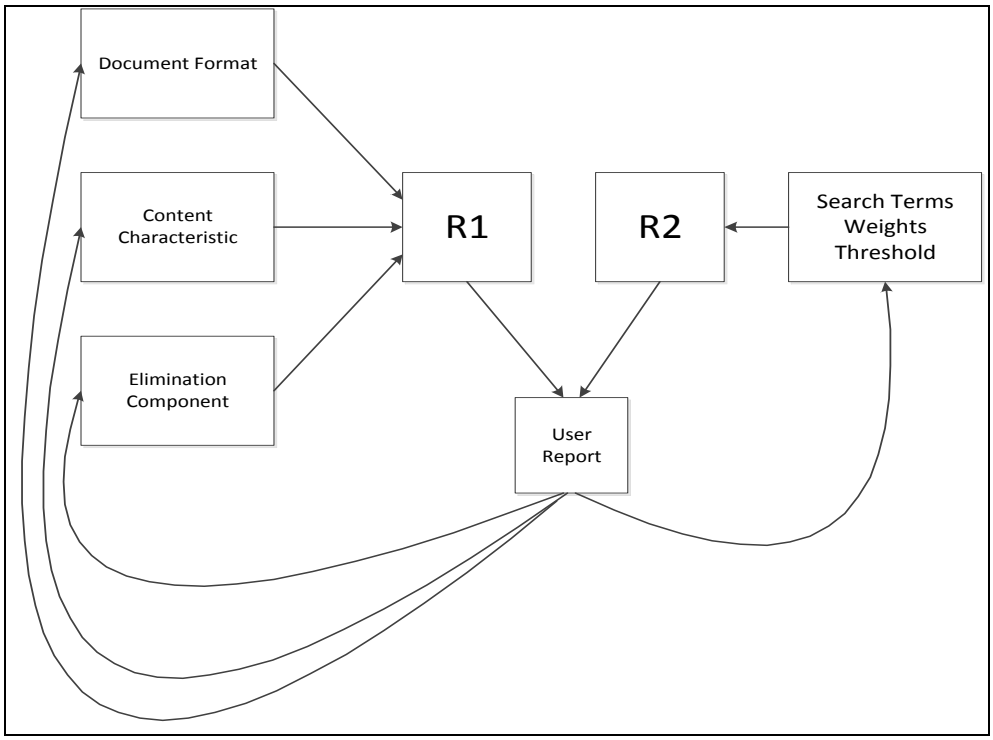


Figure 3


Document Format
Content Characteristic
Elimination Component
R1
R2
Search Terms Weights Threshold
User Report

REFERENCES

Baron, J., "Toward a Federal Benchmarking Standard For Evaluating Information Retrieval Products Used In E-Discovery," *The Sedona Conference Journal*, Vol. 6 No 1, 2005.

Baron, J., R., "The TREC Legal Track: Origins & Reflections on the First Year," Sedona Conference Journal, Volume 8 (2007).

Bates, M., J., "Information Search Tactics," *Journal of the American Society for Information Science*, July, (1979).

Bates, M., J., "Subject Access in Online Catalogs: A Design Model," *Journal of the American Society for Information Science*, November (1986).

Blair, D., C., and Maron, M., E., "An Evaluation of Retrieval Effectiveness for A Full-Text Document-Retrieval System, *Communications of the ACM*, Volume 28, Number 3 (1985).

Brisboa, N.R., Luances, M.R., Places, A., S., Seco, D., "Exploiting Geographic References of Documents in a Geographical Information Retrieval System Using an Ontology-Based Index," Geoinformatica, 14:307-331, (2010).

Chi-Ren, S., Klaric, M., Scott, G., J., Barb, A., S., Davis, C., H., Palaniappan, K., "GeoIRIS: Geospatical Information Retrieval and Indexing System – Content Mining, Semantics Modeling, and Complex Queries," *IEEE Transactions on Geoscience and Remote Sensing*, Volume 45, Number 4, April (2007).

Deerwester, S., Dumais, S. T., Furnas, G.W., Landauer, T. K., Harshman, R., "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*. (Sep. 1990).

Garon, A., Kontostathis, A., "Applying Latent Semantic Indexing on The TREC 2010 Legal Dataset," NIST Special Publication*, Proceedings: Text Retrieval Conference (TREC) 2010*.

Giger, H.,P., **"**Concept Based Retrieval in Classical IR Systems," *SIGIR '88 Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York (1988).

Grossman, M., R., Cormack, G., V., "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," *Richmond Journal of Law and Technology*, Volume 27, Issue 3 (2011).

Jarmon, J., "Combining natural Language Processing and Statistical Text Mining: A Study of Specialized Versus Common Languages," Working Paper (2011).

Jiang, J.,J., Conrath, D.W., "A Concept-Based Approach to Retrieval from an Electronic Industrial Directory," *International Journal of Electronic Commerce*, Volume 1, Number1 Fall (1996).

Kontostathis, A., Pottenger, W., "A Framework for Understanding Latent Semantic Indexing (LSI) Performance," *Information Processing and Management*, Volume 42, pp. 56 – 73, (2006).

Lenk, P.,J., Floyd, B.,D., "Dynamically Updating Relevance Judgements In Probabilistic Information Systems Via Users' Feedback," *Management Science*, Volume 34, Number 12, December (1988).

Luhn, H., P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development*, (1957).

Oard, D. W., Baron, J. R., Hedin, B., Lewis, D. D., Tomlinson, S., "Evaluation of Information Retrieval for E-discovery," *Artificial Intelligence and Law*, 18:347 (2010).

Oussalaleh, M., Khan, S., Nefti, S., "Personalized Information retrieval System in the Framework of Fuzzy Logic," *Expert Systems with Applications*, Volume 35, Page 423 (2008).

Park, T.,K., "The Nature of Relevance in Information Retrieval: An Empirical Study," *Library Quarterly*, Volume 63, Number 3, p. 318 (1993).

Runkler, T., A., Bezedek, J., C., "Alternating Cluster Estimation: A New Tool for Clustering and Function Approximation," *IEEE Transactions on Fuzzy Systems*, Volume 7, Page 377 (1999).

Salton, G., Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, MA, (1989).

Salton, G., Buckley, C., "Term-Weighting Approaches In Automatic Text Retrieval," *Information Processing and Management*, Volume 24, Number 5, (1988).

Salton, G., McGill, M., J., Introduction to Modern Information Retrieval. McGraw Hill Book Company, New York (1983).

Singhal, A., Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, (2001).

Van Rijsbergen, C. J, Information Retrieval, Butterworths, London, Boston. 1979.

Van Rijsbergen, C. J., Harper, D. J., Porter, M.F., "The Selection of Good Search Terms," *Information Processing and Management*. Vol. 17, Pg. 77-91 (1981).

Vorhees, E., M., "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness," *Information Processing and Management*, Volume 36, Page 697 (2000).