TREC 2008 at the University at Buffalo: Legal and Blog Track

Jianqiang Wang and Ying Sun
Department of Library and Information Studies
The State University of New York at Buffalo
Buffalo, NY 14260, U.S.A.
(jw254, sun3)@buffalo.edu

Omar Mukhtar and Rohini Srihari
Center of Excellence for Document Analysis and Recognition
The State University of New York at Buffalo
Buffalo, NY 14260, U.S.A.
omukhtar@buffalo.edu, rohini@cedar.buffalo.edu

Abstract

In the TREC 2008, the team from the State University of New York at Buffalo participated in the Legal track and the Blog track. For the Legal track, we worked on the interactive search task using the Web-based Legacy Tobacco Document Library Boolean search system. Our experiment achieved reasonable precision but suffered significantly from low recall. These results, together with the appealing and adjudication results, suggest that the concept of document relevance in legal e-discovery deserve further investigation. For the Blog distillation task, our official runs were based on a reduced document model in which only text from several most content-bearing fields were indexed. This approach indeed yielded encouraging retrieval effectiveness while significantly decreasing the index size. We also studied query independence/dependence and link-based features for finding relevant feeds. For the Blog opinion and polarity tasks, we mainly investigated the usefulness of opinionated words contained in the SentiGI lexicon. Our experiment results showed that the effectiveness of the technique is quite limited, indicating other more sophisticated techniques are needed.

1 Interactive Legal Task

For the 2008 TREC Legal track, the University at Buffalo (UB) team participated in the interactive search task. The high level goal of the TREC Legal track is to advance computer technology on the effective search of electronic business records in the legal domain, a problem generally known as "legal e-discovery." The documents used in the track, the IIT Complex Document Information Processing (CDIP) Test Collection, contains nearly seven million business documents including correspondences, memos, publications, and regulations of tobacco products that are released under the "Master Settlement Agreement" (MSA). Two features of the collection are of particular interest: rich metadata fields that were produced by human catalogers and noisy text of full documents that were generated from some Optical Character Recognition (OCR) software. The search topics are several hypothetical "complaints" that are typically filed in court. Each complaint lays out the background of a legal case, including factual assertions and causes of action, as well as some specific requests for production of responsive documents. In practice, the main responsibility of a searcher, usually an e-discovery firm, is for each request for production to find as many responsive documents as possible while minimizing the number of false hits. For the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
TREC 2008 at the University at Buffalo: Legal and Blog Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The State University of New York at Buffalo,Department of Library and Information Studies,Buffalo,NY,14260**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**12**	

Legal track search tasks, each participating team is free to use any of the document fields and any information contained in each complaint, thus leading to a variety of potential retrieval techniques and strategies. More information of the test collection can be found in the 2006 track overview paper [1].

To focus research effort on different important aspects of the legal e-discovery task, the Legal track has three separate tasks: a routine ad hoc task, a feedback task, and an interactive task [13]. The UB team chose to work on the interactive search task for this year. In this paper, we describe our research interests, our search system and strategies, search outcomes, and the official evaluation including appealing and adjudication results.

1.1 Areas of Interest

The challenges of legal e-discovery have been well-noticed in both the research and practice communities. However, the greatest success of search technology, namely the ranked retrieval algorithms and techniques, has not been widely applied to legal e-discovery, for which the typical approach is the Boolean search. The goal of our focus on the interactive Legal e-discovery task is two-fold: to gain a better understanding of the Boolean e-discovery practice and to develop better ranked retrieval techniques based on this understanding. In addition, we would also like to learn more about the document collection, something usually not easy to achieve with batch retrieval in which the human searcher is not directly involved. An interesting feature of this year's interactive Legal task is the inclusion of Topic Authorities (TAs). Specifically, each participating team had chances to clarify its understanding of document relevance for a topic, and TAs adjudicated the initial relevance judgments of documents that a participating team disagreed. Therefore, this year's interactive task is one step closer to model the real life practice of legal e-discovery, thus more interesting.

1.2 Search System and Topic

The small size of our team and the time constraint prevented us from building a full-blown interactive search system of our own for the task. Therefore we studied several existing systems and eventually chose a Web-based Boolean search engine maintained by the Legacy Tobacco Document Library team at the University of California at San Francisco (UCSF) [1]. The system contains a superset of the documents used in the Legal track. Some of the most important features of the system include:

- *Three levels of search* Users can select from basic search, advanced search, or expert search mode. The basic search mode is for search with one term on the entire document or the searchable fields (Title, Author, Bates Number, Document Type, Entire Record, Metadata, and Text body). The advanced search mode supports Boolean query formulation with up to six search terms. With the Expert Search interface users can create more complex Boolean queries using mixed operators and wildcards while searching on any field in the collection. We used the expert search interface to generate all the queries and the corresponding official runs reported in this paper.

- *Search on individual fields or the whole document* Such flexibility turned out to be very useful. For example, searchers can easily construct queries that will return or exclude documents that contain or do not contain certain terms in their Title field. Also, this makes it easy to search based on dates.

- *Search with Wildcards* This feature makes it easy to specify variants of query terms and proximity search.

- *Full document image* Full documents in PDF format and TIFF format are provided by the system, which facilitates the review of retrieved documents.

[1] http://legacy.library.ucsf.edu/

- *Search history and bookbag* The system automatically keeps search history (queries and the number of retrieved documents for each query) for a login session, and the search history can be easily downloaded to a local computer. The "bookbag" keeps documents that a searcher has selected. The searcher can also write notes about any document in the bookbag. The bookbag needs to be downloaded before a search session ends.

We used all these features of the system quite frequently. Initially we encountered one problem with the "bookbag" feature. Since the default Web interface of the UCSF search system allows to display up to only 50 documents per page, marking relevant documents (so that their document IDs can be saved in the bookbag) can be very time-consuming if the number of retrieved documents is large. Upon our request, the system staff provided an API command that allows to change the number of documents per page up to 10,000. Given the number of documents we retrieved, it only requires a few pages to display them.

Three topics, together with the complaint that these topics are based on, were provided for the interactive search task. The complaint describes a hypothetical legal case in which a company is accused of providing false information to its security customers. Specifically, the company is suspected to have been involved in illegal domestic retail marketing practice of cigarettes and payment to foreign officials. Consequently, the three requests for production seek documents that talk about (1) marketing or advertising restrictions, (2) retail marketing campaigns for cigarettes, and (3) payments to foreign government officials. Due to constraints on the task schedule and the availability of our searchers, we were able to work only on Topic 103 for our official submission, which is the topic of retail marketing campaigns for cigarettes.

1.3 Query Formulation

The request for production of Topic 103 states that relevant documents should "describe, refer to, report on, or mention any "in-store," "on-counter," "point of sale," or other retail marketing campaigns for cigarettes." At the beginning we thought this is an easy search request but as we worked on it we realized it is indeed difficult. One of the reasons is that a relevant document may not use any of the terms in the request, except "cigarette." Our understanding is that in-store, on-counter, and point-of-sale retail marketing of cigarettes are just examples of relevant retail marketing activities - there are other forms of retail marketing campaigns that can be relevant, such as during concerts, sports, and even in streets. For example, a document that describes sales representatives of a tobacco company distribute free cigarette samples on a car racing event is deemed to be relevant, while it never uses any of the other words in the request for production. Our clarification with the TA confirmed our speculation. Therefore, terms like "in-store," "on-counter," "point of sale," and "retail marketing" should be ORed with "cigarette" if they are to appear in a Boolean query.

On the other hand, documents mentioning those words in the topic statement may not be relevant. For example, a document would be non-relevant if it discusses only the qualification of some retail stores for becoming a member of the Retail Partners Program of a tobacco company without mentioning the consumer. The TA explained that "I think the communication would need to refer or relate or imply some distribution to customers and a particular program or campaign to be responsive."

Many documents that we initially found describe federal, state, local, and company rules, codes, regulations, etc. that apply to retail marketing of cigarettes. However, it is somewhat difficult to judge the relevance of these documents. The TA clarified that the relevance of such documents should be decided based on whether they are general to all retail marketing activities or specifically about some individual events. She explained that "I think as a general matter, if the document was generated by the government and discusses rules or regulations relating to free samples of cigarettes, I would not consider this alone to be responsive to the Topic 103 request for documents relating to retail marketing campaigns. If the document was generated by a cigarette manufacturer and said that in connection with providing free samples of Merit or Pall Mall, we must follow the attached rules, I think it would be responsive. If the document was sent internally at Philip Morris

to the entire Marketing Department and said "Attached are the new government rules regarding free cigarette samples," I would probably err on the side of deeming that responsive." From these conversations, it seems that we should excluded documents of general rules, codes, regulations, and their like, but the fine line is hard to maintain.

For the submitted run, two faculty members worked on the search task in two weeks, of course with other teaching and service responsibilities going on at the time. We did not accurately keep track of the time spent on the task. Our plan was that each of us focused on different types of potentially relevant documents and then merged the search results. Here are the five queries that led to our official runs:

1. *UBQ01:* (cigarette* AND ("off package"∼2 OR "off carton"∼2 OR "off pack"∼2) NOT (ebay))
2. *UBQ02:* (cigarette* AND ("discount campaign"∼10 OR "discount promotion"∼10) NOT (dt:(regu* OR "*study") OR ti:(study OR banning OR ban OR regulat* OR health OR legislat*)))
3. *UBQ03:* (cigarette* AND (giveaway OR "give away") NOT (dt:(regu* OR "*study") OR ti:(study OR banning OR ban OR regulat* OR health OR legislat*) OR ot:(violat* OR illegal OR ban OR "quitting smoker" OR "quited smoker" OR "quit smoking" OR child OR children OR minor OR "additive*" OR cancer OR "health department"∼3 OR "American health foundation")))
4. *UBQ04:* (cigarette* AND ("gift purchase"∼10 OR "store sale") NOT (dt:"regu*" OR ti:"study"))
5. *UBQ05:* ((distribut* AND (sample* OR coupon*))∼5 AND cigarette* NOT (dt:(law OR laws OR code OR codes OR regulat* OR legislat* OR publication*) OR ti:(bill OR bills OR law OR laws OR rule OR rules OR code OR codes OR regulat* OR legislat* OR restrict* OR guideline* OR manual* OR profile* OR motion* OR standard* OR requirement* OR practice*)))

The first four queries are generated by one team member with the expectation to retrieve potentially relevant documents mentioning some specific types of retail campaign events. The types of events focused in the queries are: store sale, giveaways, and gifts with purchase. Each query has three clauses: (1) cigarette and its spelling variants; (2) expressions carrying the meaning of one or two types of retail campaign events; and (3) a not clause to exclude some false positive documents. For each type of event, the searcher came up with possible expressions and only the expressions that have proved to be present in the collection are kept in the final queries. For the purpose of accessing query qualities, the queries are not defined neatly as one query per type of event, but quite ad-hoc with a general rule that the searcher feels comfortable to manipulate the number of documents retrieved. The NOT clause is added after looking at a small sample of documents retrieved by a query containing only the first two clauses. The TA was consulted to verify some sample documents.

We found a large amount of scanned coupons in the collection with very few texts in them. However the texts follow the general pattern as "certain amount off per/every/a package/carton/pack." The TA gave a positive answer to such documents. Query UBQ01 is generated to catch such documents. Query UBQ03 is about free gift or sample giveaway events. A lot of efforts were made to exclude non-responsive documents with this query. Query UBQ04 uses the general concept terms for store sale and gift with purchase events. The query is used to target documents talking about the events in general, such as a marketing report or general store sale instructions. Query UBQ02 is also about the store sale events where "discount" is mentioned. We limited the appearance of the term "discount" somewhere near "campaign" or "promotion" based on our initial understanding that the responsive documents should mentioned some specific events. However, a last-minute contact with the TA proved that the understanding is incorrect, which may be one of the reasons leading to our low recall results.

Query ID	Rtrd	Sampled	Rel(I)	Non_Rel(I)	A_A	A_A_Rel
UBQ01	38032	290	226	64	23	23
UBQ02	4827	26	21	5	0	0
UBQ03	4198	24	18	6	1	1
UBQ04	4484	23	16	7	0	0
UBQ05	38415	271	129	142	22	17

Table 1: **Run statistics: Initial and Post-Adjudication.** *Rtrd*: the number of documents retrieved by each query; Sampled: the number of retrieved documents that were sampled for the official relevance judgment; *Rel(I)*: the number of our sampled documents that were initially judged as relevant; *Non_Rel(I)*: the number of our sampled documents that were initially judged as non-relevant; *A_A*: the number of *Non_Rel(I)* that we appealed for adjudication; *A_A_Rel*: the number of *A_A* that were adjudicated as relevant.

Query UBQ05 was generated by another team member. It requires the presence of the words "distribute," "cigarette," and either of the words "sample" and "coupon" (or their variants). In addition, it requires relevant documents not to be regulations and their likes, as restricted through the "document type" (dt) field, or document titles not to contain these words. The proximity operator $\sim$ specifies the window size for words to appear together in a document. We tried several different window sizes and felt the window size of five can bring in a reasonable number of relevant documents. Of course, this heuristics is only based on some sample documents. Given the low recall of our results, we suspect that we should have probably set the window size larger.

1.4 Official Evaluation Results

Table 1 shows the statistics of our submitted run, which merges documents retrieved by the five queries. We can see that among all the five queries, UBQ01 and UBQ05 retrieved about the same number of documents, which is much more than the other three queries. It's worth mentioning that there is very little overlap between these two sets of documents, suggesting the queries pulled out documents from different categories or types. Our merged (official) run contains 67,334 documents. Only less than 7% of these documents were sampled for official review. The effectiveness of our run is 4.7% for recall, 63.4% for precision, and 8.7% for F measure (before appealing and adjudication). Both the balanced F measure and recall are very low for our run, indicating we missed lots of relevant documents.

1.5 Results after Appealing and Adjudication

An interesting and important feature of this year's interactive search task is appealing and adjudication after the initial relevance assessment results were released. Each participating team was given an opportunity to question the relevance judgment of any of the sampled documents that the team disagreed. The TA would further examine the appealed documents and made a final assessment of their relevance. Appealing and adjudication was done after the TREC conference. Our team looked particularly at those documents that were initially judged as non-relevant and selected 40 of them that we felt representing different types of documents that we have retrieved. In other words, there are more documents that we wished to be adjudicated, but we believed that once we had the adjudicated results of these 40 documents we would have a better understanding of the true relevance of other documents that are similar to the adjudicated documents. Of course, another reason for us to appeal only a relatively small set of documents for adjudication is we did not want the TA to spend too much time on similar documents.

Table 1 also shows, for each of the five queries, the number of documents that were initially judged as non-relevant and we appealed for adjudication. There were six appealed documents that were retrieved by both UBQ01 and UBQ05, hence 40 unique documents for adjudication. As we can see, among these 40 documents (again, initially judged as non-relevant), 35 were adjudicated

by the TA as relevant, showing there was some noticeable discrepancy of relevance judgments between the assessor and the topic authority. Adjusting the initial official evaluation results by taking into consideration the appealing and adjudication results, the estimated recall, precision, and F measure of our submitted run are boosted to 6.1%, 71.6%, and 11.3%, respectively.

We believe that the track overview paper will provide a more detailed analysis of the process and results of appealing and adjudication. From our perspective, however, the results, together with our experience of interacting with the topic authority, make us believe that relevance assessment of documents for legal e-discovery is a very challenging task. The challenge could be due to several factors. One possible factor is that the concept of document "relevance" in the legal e-discovery domain does not seem to be quite the same as in other domains (e.g., with newswires stories).

1.6 Conclusions for the Interactive Legal Task

At the first glance, our interactive legal search results do not seem to be very encouraging. However, we have learned a lot than what these recall, precision, and F measures show. Through our participation in the task, we have become more familiar with the actual legal e-discovery process. The involvement of the topic authority in this experimental evaluation significantly helped us to understand that process as well as the concept of document relevance in the legal e-discovery domain. We also learned the usefulness of utilizing certain document features such as document types and specific fields, and query formulation techniques such as proximity search, wildcards, and relevance feedback to improve the search performance. Our immediate next step would be to study how these techniques can be automated so that they become part of some text analysis, indexing, and searching algorithms. We feel that the concept of document relevance in legal e-discovery deserves further investigation and understanding. One possible direction is to look at the different *types* of relevance [8]. Therefore, we look forward to continuing our effort in the TREC 2009.

2 Blog Distillation Task

Blog Distillation or "feed search" is the task of finding blog feeds that have a principal recurring interest in a topic X, where X is an information need expressed as a short query (title), a short sentence(description) and a short paragraph(narrative). The goal is to return a ranked list of feeds that are most relevant to the expressed information need.

The document collection contains three sets of documents: 88.8 GB permalink documents, 28.2GB homepages, and 38.6GB RSS and ATOM feeds. We used the permalink collection which is organized as 3.2 million documents. In the previous two years' Blog track, groups have looked at the so-called large document model in which all documents from a feed are used to make a single document that represents that feed [11]. An alternative model is to execute the query against the individual pages and then use a ranking method that scores feeds based on the ranking of the individual pages, hence, the small document model [4]. We choose the large document model because it has shown to work better than the small document model. We also looked at the reduced document model in which text between certain tags is indexed (described below). Another approach we tried was to integrate the PageRank of a feed with its query likelihood score. For all our experiments we used the Indri retrieval engine as well as the Lemur toolkit.

2.1 Corpus Processing

A FEEDNO tag is included in the DOCHDR section of each permalink document, indicating which feed it belongs to. With that information, we were able to aggregate all documents with the same FEEDNO in the permalink collection into one document. There are about 100K individual feeds so we created an index with about 100K documents. We initially built two indices, one with stop-words removed and the other without. Our preliminary experiments showed that removing

stop words seemed to help improve retrieval effectiveness. Therefore, our official runs were based on document indices with stop words removed.

For the reduced documents model, we built three separate indices: (1) Title (T) only index, by indexing only the text in the TITLE field, (2) Title and Heading (TH) index, by taking only the text in the TITLE and the heading fields(H1, H2, and H3), and (3) Title, Heading and Paragraph (THP) index, by taking the text in the TITLE, H1, H2, H3, and P fields. We did not remove any stop words in the reduced document model but did use Krovetz stemming.

We did not use any technique to remove non-English or spam blogs. According to statistics of the 2007 Blogosphere [2], only 36% of the blogs are English. Our motivation for not removing non-English blogs is that we want to see how robust our approaches are in dealing with English Blogs mixed with non-English ones. We note, however, that excluding non-English and spam blogs could improve performance on Blog Distillation [10].

2.2 Query Construction

A retrieval engine will perform only as well as the quality of the words that are used to express an information need. Query expansion techniques such as relevance feedback (including pseudo relevance feedback), WordNet association, and Wikipedia term association are all ways to enhance the quality and clarity of the information need. We performed experiments on different methods of combining words in title, description and narrative. We used last year's relevance judgments and Mean Average Precision(MAP) as the metric to test the performance of our techniques.

For query construction, our baseline run was produced with queries containing only the words from the title field. We looked at the MAP obtained from running the full title, description and narrative and their combinations. In our experiments removing all stop words from the title, description and narrative improved performance. More complicated techniques such as, automatically extracting the most important words by finding the capitalized words, phrases such as "University at Buffalo" where "at" is a connective between two capitalized words, showed further improvement. The addition of lower case words did not improve MAP or R-Precision but we did get the highest P@10 when we take the title, capital words and lower case words. (see Table 2. The pre-processing we did is similar to query expansion in that we were trying to clarify the information need of the user. The similarity between our approach and query expansion is that our aim is to add words to the short original query to clarify the information need.

2.3 Documenting Ranking Techniques

We tried three document ranking techniques in our experiments, namely, query likelihood and PageRank, support vector machine, and reduced document model.

2.3.1 Query Likelihood and PageRank

We first looked at whether combining the query likelihood and a popularity metric improves performance on the blog distillation task. The popularity metric we used is PageRank [9]. Specifically, we used the link structure within the document collection to calculate the PageRank. To calculate the PageRank values of the feeds we used Lemur's PageRank utility.

We first calculated the PageRank of the individual feeds. Then we used a linear combination of the PageRank and the query likelihood score to rank the feeds. There were two parameters that we needed to learn, the relative weights to be given to the query likelihood and the PageRank of the feeds. Another interesting thing we found was that only 27K blogs out of 100K blogs had a PageRank i.e. 72K feeds had a PageRank of zero. The distribution of the PageRank's of the feeds are given in table 3. We tried different combinations of these scores but found that using the PageRank degraded performance unless the weight of the PageRank was set so low (less than 0.01) so that it would be almost ignored. In conclusion, we found that using the PageRank did not improve performance.

[2] http://www.sifry.com/alerts/archives/000493.html

Query Fields	Index used	Processing	MAP	R-Precision	P@10
T	T	none	0.2012	0.2705	0.4541
T	T & H	none	0.1837	0.2626	0.4054
T	T, H & P	none	0.2075	0.2887	0.3895
T	Full	none	0.2430	0.3167	0.4103
T & C wd	T only	T, C from D & N	0.2432	0.3225	0.4711
T & C wd	T & Hd	T, C from D & N	0.2249	0.3027	0.4156
T & C wd	T, Hd & P	T, C from D & N	0.2416	0.3229	0.4244
T & C wd	Full	T, C from D & N	**0.2915**	**0.3700**	0.4400
T ,C & L	Full	T, C from D & N, S	0.2890	0.3594	0.4636
T ,C & L	T	T, C from D & N, all wd x S	0.2483	0.3333	**0.4955**
T ,C & L	T & H	T, C from D & N, all wd x S	0.2171	0.2856	0.4455
T ,C & L	T, H & P	T, C from D & N, all wd x S	0.2477	0.3239	0.4705

Table 2: The effect of query construction on different indices. Evaluation was on the Blog Distillation Relevance Judgments from last year. T is Title, C is capital words, H is heading, P is paragraph, S is stop words, wd is words , D is description, N is narrative , L is lower case words, x is except.

PageRank	10	9	8	7	6	5	4	3	2	1	0
Frequency	1	3	9	25	69	189	517	1,415	3,866	17,303	77,252

Table 3: **PageRank of Feeds**

2.3.2 Support Vector Machine

Another system we built was based on Support Vector Machines(SVM). For this purpose we used the LibSVM toolkit [3]. We took the qrels from last year, extract features for each feed, and then use SVM to classify the feeds. The features that we used to train our SVM include: PageRank of a feed within the permalink collection, the number of documents in the feed, the length of the feed document, the average document length in the feed, the number of times, each word in query title is found, and the number of times, the title of the query is found (Phrase Matches). As we can see, some of the features rely on the relevance judgments from last year's track.

The performance metric we used was the percentage of correctly identified feeds in the testing set. The distribution of relevant/irrelevant feeds in the relevance judgment file is skewed in favor of irrelevant feeds. To overcome this problem, we penalized our SVM whenever it misclassified a relevant feed. Our best performance using SVM and the features itemized above was 62%.

2.3.3 Reduced Document Model

Fields Indexed	Size	Avg. Doc. Length	MAP	R-Precision	P@10
Full	24.3GB	16,000	**0.2430**	**0.3167**	0.4103
T	236MB	235	0.2012	0.2795	**0.4541**
TH	922MB	1,356	0.1837	0.2626	0.4054
THP	6.58GB	7,634	0.2075	0.2887	0.3895

Table 4: Index Sizes. T is Title-only index created from words inside TITLE field of a webpage, H is heading index created from words inside H1, H2, or H3 fields, and P is paragraph index created from words inside P field.

If one looks at documents belonging to a feed, there is a certain structure to them. The Blog (Feed) title is repeated on each page, there are links to similar blogs (blog roll), there is an archive of previous posts, there is the main "post" and finally the "comments." The archive, blog roll, and feed title do not change much and are almost static, while the "post" and the "comment" changes in each document. So there are static (or almost static elements) and dynamic content on each documents belonging to a feed. For the distillation task, our goal is to find out whether posts or comments to a post from a feed have a "principal recurring interest in X (information need)."

We used the reduced document model to filter out the non-relevant elements of a feed. Instead of labeling different HTML tags from a feed as relevant/irrelevant, we used a heuristic method in which texts within certain HTML tags is indexed while texts within other tags is simply ignored. The tags we chose to index were the TIELE, H1, H2, H3, and P tags. Our approach is much simpler than the approach taken by [6] and [7]. The former group's approach is based on VIPS (Vision based Page Segmentation) algorithm [2] for page segmentation. They used layout features and language features to identify important content blocks. The latter group used content extraction on the Opinion track but their approach is also based on page segmentation and uniqueness of text within tags. [14] also preprocessed the permalink collection for the opinion retrieval task to remove the script and style information in documents, which led to a reduction of 80% in page size (from 88GB to about 18GB).

The motivation for the reduced document model is to see the effect of index reduction on performance and to simplify the filtering of document elements. Both techniques (VIPS and Uniqueness of text) are computation-intensive, and might not be suitable for real time filtering. Another motivation is to mimic what people do when they subscribe to blogs. People seldom read every single document from a feed. Instead, they perform a search based on their information need and look at the titles and snippets to decide if the feed might be relevant and worth subscribing to. Table 4 shows the reduction in index as well as the performance penalty of index size reduction. We used the queries and relevance judgments from last year to measure performance. We achieved a reduction in index size to 0.01% with a performance penalty of -18%. However, we received a 10% boost in P@10. For the heading and paragraph indices we observed similar trends but no such improvement in P@10.

2.4 Experiments and Results

Since we got disappointing results from the combination of query likelihood and PageRank feature, we did not submit any run based on this technique. With SVM, we used classification errors as our metric. Therefore, all our submissions are based on query likelihood on the feed index as well as the reduced document model. For our baseline, we used the title only query on the feed index. For other submissions, we built the query from the title and the capital words as explained in the query construction section above. All evaluations are on the 50 new queries proposed this year. 5 of those were proposed by our group.

The MAP, R-Precision and P@10 for our submitted runs are given in Table 5. Our baseline run is UBDist1 which is a title only run on the full index. Our second run is UBDist2, for this run, we built the query from title, the index we used is the title, heading and paragraph index. We see a drop in MAP, R-Precision and P@10 although the performance was not as severe as observed on the queries from last year (Table 4). This could be because of the change in nature or the queries themselves. Our third run is UBDist3, which is produced with queries built from the title words, capital and lower case words and index with the title, heading and paragraph. For our final run UBDist4, we built the query from title and capital words and again used the title, heading and paragraph index. We achieved the best performance with this run, which is an improvement of 9% in MAP, 8% in R-precision and 2% in P@10 over our baseline run.

Runs	MAP	R-Prec	P@10
UBDist1	0.2410	0.2916	0.3720
UBDist2	0.2348	0.2949	0.3640
UBDist3	0.2304	0.2951	0.3460
UBDist4	0.**2633**	**0.3160**	**0.3820**

Table 5: **Performance of official runs on Feed Distillation task**

UB Runs	MAP	R-Prec	P@10
UB (baseline)	0.1700	0.2267	0.3793
UBop1	0.1570	0.1925	0.3093

Table 6: **Performance on Opinion Mining task**

3 Opinion and Polarity Task

The Opinion retrieval task involves locating blog posts that express an opinion about a given target. It can be summarized as, "What do people think about "target" where the target is an information need." Thus, the Opinion task is a subjective task. Each search topic in this task is about a person, a location, an organization, a product, or an event. The topic of a post does not necessarily have to be the target, but an opinion about the target must be present in the post or one of the comments to the post in order for the blog to be considered relevant.

The goal of the Polarity task is, for each topic, all positive opinionated documents and negative opinionated documents should be ranked and retrieved. This task is very similar to the Opinion task with a small difference: in the Polarity task no documents that have both negative and positive opinion about a target should be returned.

Since both tasks are subjective, we need some human judgments on words that are considered opinionated (positive or negative or objective). That is, a lexicon with words labeled as positive, negative or objective is needed. In our experiment, we used SentiGI lexicon [5], as described below.

3.1 Lexicon

The SentiGI lexicon is based on the General Inquirer (GI) lexicon [12]. The GI lexicon contains positive and negative words. The SentiGI lexicon adds a new category of objective words. The SentiGI positive and negative word lists are made from 1,612/1,982 terms obtained by two original GI sets of 1,915/2,291 terms after removing 17 terms appearing in both categories (e.g. deal). For the Objective category, the 7,582 GI terms that are not labeled as either Positive or Negative, are labeled as Objective. Then all the multiple entries of a single term caused by multiple senses are reduced to a single entry. After this reduction we have 5,009 objective terms. We used this list of positive, negative and objective words in our experiment. We trained our neural network based on the relative word distribution on last years judgments.

3.2 Opinion Task

To train our neural network we used the following features: (1) document length, (2) number of positive words, (3) number of negative words, (4) number of objective words, (5) ratio of positive words to total number of words in document (after stop word removal, same below), (6) ratio of negative words to total number of words in document, and (7) ratio of objective words to total number of words in document.

After training, we first ran the title-only query and got an initial set of documents. Since for this task, 1000 documents had to be returned and some documents may not have any opinion or might be rejected, our initial set was made of the top 2000 documents returned for a query. We then used the neural network to obtain a class label (Opinionated or Unopinionated) for the document. For the Opinion task the class labels were Positive, Negative, Mixed and Unopinionated. If a

UB Runs	MAP	R-Prec	P@10
UBpol1 (positive)	0.0666	0.0785	0.0906
UBneg1 (negative)	0.0529	0.0686	0.0725

Table 7: **Performance on Polarity task**

Method	QRel	MAP	R-Prec	P@10
A	OQ	0.3476	0.3803	0.6250
A with O	AQ	0.2698	0.3471	0.6208

Table 8: Performance on Opinion Mining Task with and without opinion mining. A means Ad-hoc retrieval techniques used, O means Opinion mining technique used, OQ means 2007 Opinion qrel, AQ means 2006 Ad-hoc retrieval.

document is judged to be positive, negative or objective, we take a linear combination of its initial relevance and the probability assigned to the label (positive, negative or objective). We finally reranked the documents based on this modified score.

3.3 Polarity Task

Unlike the Opinion task, documents of mixed opinion should be excluded from the returned set. Thus we did not need to retrain our neural network and needed to make only some minor changes. The class labels were Positive, Negative, Mixed or Objective and Unopinionated. If a document was judged to have mixed opinion, we rejected it. All documents ranked positive were separated from the the documents ranked negative. We then output the final ranked set of positive and negative documents.

3.4 Experiments and Results

For testing we used the Indri query language; our query looks like the following

$$\#1(MarchofthePenguins)\#uw(MarchPenguins)$$

where "#(March of the Penguins)" was taken from the title of the query and executed as a phrase query, and the words "March" and "Penguins" are extracted from title after removing stop words from the query itself. We use Indri's built-in Language Modeling (query likelihood) ranking method to rank the documents.

We found that there was a high correlation between relevant blogs and opinionated blogs. Indeed, applying opinion mining degraded the performance (see Table 8). Such degradation after processing was also observed by a few other teams in last year Opinion Mining task (e.g., [10]). Our hypothesis for the drop in performance is that most blogs are opinionated, thus this kind of shallow opinion mining could have filtered out some relevant blogs. In that table, **UB** is our baseline submission, for which we used the title as our query and the default query likelihood model in Indri to rank documents. For **UBop1**, we used the title as our query, got an initial set of documents, and used our neural network to rerank these documents.

Table 7 shows the result for the polarity task. From the results we see that the performance of our system is better on positive opinionated documents than on negative opinionated documents. We used the title as our query, got an initial set of documents, and used our neural network trained to filter out blogs with mixed opinions. The reranked list of positive opinionated blogs was submitted as **UBpol1** and the list of negative opinionated blogs was submitted as **UBneg1**.

3.5 Conclusions for the Blog tasks

For the Feed Distillation task, we applied link-based feature PageRank in combination with query likelihood model but the combination did not work well. We also studied query independent

features on an Support Vector Machine classifier. Our official submission, however, was based on the reduced document model in which text between certain tags was indexed. We got encouraging results and improved on our baseline while achieving a reduction in index size.

For the Opinion and Polarity task, we used the SentiGI lexicon to find opinionated words in a document. We trained a Bayesian Neural Network using the relevance judgments from the last two years. The primary feature we used was the distribution of the positive, negative and objective words in opinionated documents. Our results are not very encouraging, suggesting that simple word distribution doesn't capture opinions about an information need. The same can be said about ranking documents that are positive or negative about an information need, suggesting more sophisticated techniques based on deeper lexical and semantical analysis of documents need to be considered.

References

[1] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC 2006 legal track overview. In *The Sixteenth Text REtrieval Conference TREC*, 2007. http://trec.nist.gov/.

[2] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Vips: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft Research, 2003.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[4] Jonathan Elsas, Jaime Arguello, Jamie Callan, and Jaime Carbonell. Retrieval and feedback models for blog distillation. *Proceedings of TREC 2007*, 2007.

[5] Andrea Esuli. *Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications*. PhD in Information Engineering, PhD School "Leonardo da Vinci", University of Pisa, 2008.

[6] Liu Kang, Wang Gen, Han Xianpei, and Zhao Jun. Nlpr in trec 2007 blog track. *Proceedings of TREC 2007*, 2007.

[7] Xiangwen Liao, Donglin Cao, Yu Wang, Wei Liu, Songbro Tan, Hongbo Xu, and Xueqi Cheng. Experiments in trec 2007 blog opinion task at cas-ict. *Proceedings of TREC 2007*, 2007.

[8] Douglas W. Oard, Jianqiang Wang, Gareth J. F. Jones, Ryen W. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. Overview of the CLEF-2006 cross-language speech retrieval track. In *Proceedings of CLEF'06*, 2006.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[10] Kazuhiro Seki, Yoshihiro Kino, Shohei Sato, , and Kuniaki Uehara. Trec 2007 blog track experiments at kobe university. *Proceedings of TREC 2007*, 2007.

[11] Jangwon Seo and W. Bruce Croft. Umass at trec 2007 blog distillation task. *Proceedings of TREC 2007*, 2007.

[12] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[13] Steven Tomlinson, Douglas W. Oard, Jason R. Baron, and Paul Thompson. Overview of the 2007 TREC legal track. In *The Sixteenth Text REtrieval Conference TREC*. National Institutes of Standards and Technology, 2007. http://trec.nist.gov/.

[14] Qi Zhang, Bingqing Wang, Lide Wu, and Xuanjing Huang. Fdu at trec 2007: opinion retrieval of blog track. *Proceedings of TREC 2007*, 2007.