Are Web User Comments Useful for Search?

Wai Gen Yee, Andrew Yates, Shizhu Liu, and Ophir Frieder
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616 USA
+1-312-567-5205
waigen@ir.iit.edu, ayates@iit.edu, sliu28@iit.edu, ophir@ir.iit.edu

ABSTRACT
We consider the potential impact of comments on search accuracy in social Web sites. We characterize YouTube comments, showing that they have the potential to distinguish videos. Furthermore, we show how they could be incorporated into the index, yielding up to a 15% increase in search accuracy.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval – *search process*.

General Terms
Measurement, Performance, Experimentation.

Keywords
search, comments, YouTube.

1. INTRODUCTION
The popularity of modern Web information sharing sites (e.g., online newspapers, shopping or video sharing sites), where users can post comments about the subject matter has increased the need for effective content search functionality. Work has been done on keyword (or "tag")-based search (e.g., [2][3]), but little work has been done on using user comments to improve search accuracy. We consider the impact of comments on search accuracy in the context of YouTube video search.

Search in YouTube currently involves comparing a query to a video's title, description and keywords. Comments are not factored into search ostensibly because they are unreliable indicators of content and do not exist when a video is first posted.

Note that external, non-YouTube search engines, such as Google, also do not index video comments (but do index video title, description and keywords). We confirmed this via informal experiments where we issued video comments chosen for their apparent selectivity as queries. The results of these queries did not include the corresponding videos. On the other hand, queries consisting of any combination of title, description or keywords of a video returned the corresponding video.

LSDS-IR Workshop. July 2009. Boston, USA.

If content is poorly described by the title/description/keywords, however, comment information may supplement or replace traditional forms of search. The title/description/keywords of a Westminster Kennel Show video, for example, may fail to mention "dog" (not to mention particular dog breeds), and thus not turn up in the results for "dog show." Searching through comment information will almost certainly solve this problem.

In this paper, we explore the "nature" of user comments and how they may aid in search. Specifically, we analyze the term distributions of user comments and attempt to apply this information to improve search accuracy.

The hazard associated with the use of comments to improve search accuracy is that they may contain noisy terms that hurt performance as well as significantly increase the size of the index. Our experimental results, however, suggest that while some queries are negatively affected by comments, overall, they can improve query accuracy by nearly 15%. Furthermore, we apply techniques that can reduce the cost of using comments by up to 70%.

2. ANALYSIS OF THE YOUTUBE DATA
We crawled YouTube during February, 2009 and collected the text associated with the 500 most popular and 3,500 random videos. Popular videos were identified using the YouTube API. Random videos were retrieved by randomly selecting results of queries consisting of terms selected randomly from the SCOWL English word list [12]. For each video, we retrieved several information "fields," including:

- Title – A title assigned to the video by the user who posted it.
- Description – A video description by the user who posted it.
- Keywords – Video "tags" by the user who posted it.
- Comments – Comments by viewers of the video.

In total, for the 4,000 videos, we retrieved over 1 million comments made by over 600,000 users. We refer to the random 3,500 videos and the popular 500 videos together as the "small" data set.

We also similarly created a "large" data set, also consisting of a random and a popular part, crawled in May, 2009. This data set consists of 10,000 randomly crawled videos and 1,500 popular videos. The four data sets are thus:

- rand3500: This data set contains data on 3,500 videos, randomly crawled from YouTube in February, 2009. This

data was found on YouTube by issuing random queries from the SCOWL word list [12].

- pop500: This data set contains data on the 500 most popular videos according to YouTube as of February, 2009.
- rand10K: This data set contains data on 10,000 videos randomly crawled from YouTube (is the same way that rand3500 was collected) in May, 2009.
- pop1500: This data set contains data on the 1,500 most popular videos according to YouTube as of May, 2009.

In our experiments, we pre-processed the data using the Porter stemming algorithm [10]. We also tried a more conservative stemming algorithm [11] in anticipation of problems with overstemming from the unique language usage found in video comments. However, the different stemmer had little effect on the final results. We also remove stop words using the Lucene stop word list.

2.1 Basic Statistics

As shown in Table 1a, popular videos have more than 3 times the number of viewers than do random videos and more than 6 times the number of comments. Comment length for both types of videos is about 12 to 15 terms. On average, there are 2,280 terms describing a video from the rand3500 data set and 12,132 terms describing a video in the pop500 data set.. In the large data set, there is an even greater disparity between the random and popular videos, with more viewers and more comments.

The length statistics of the title, description and keyword fields, shown in Table 2, indicate that on average only 34 to 58 terms are used to describe a (random) video (assuming that comments are not used to describe videos). Including the comment field in the search returns a potential richer database of information because the average number of comment terms is at least 1,485.

Table 1. Average values for various comment statistics.

	#Views/Video	#Comments/Video	Comment Len
Popular	247,226	1,011	12
Random	71,654	152	15
Average	93,601	259	13

a. Small data set.

	#Views/Video	#Comments/Video	Comment Len
Popular	874,805	2,425	10
Random	62,807	135	11
Average	168,720	434	11

b. Large data set.

Table 2. Average lengths for non-comment video fields.

	Title	Description	Keywords
Popular	5	33	13
Random	5	44	9
Average	5	43	10

a. Small data set.

	Title	Description	Keywords
Popular	5	42	14
Random	5	24	10
Average	5	26	11

b. Large data set.

3. MEASURING INFORMATION CONTENT

As demonstrated in opinion-mining applications, many comments often describe something's "quality," rather than its "content" (e.g., how good a product is rather than what the product is) [1]. If we assume that quality-based comments come largely from a restricted vocabulary (i.e., adjectives, such as "good" or "bad"), then comments will have only a limited ability to distinguish one video from another apart from the subjective impression it left on the viewer. Specifically, comments from different videos in this case will have similar term distributions and therefore have poor discriminating power from the perspective of a search system. Furthermore, because queries generally contain content-based terms, they do not "match" the quality-based terms in the comments. In other words, comments contain little information useful to search.

To measure the discriminating power of each field, we compute each field's language model and then compute the average *KL*-divergence [13] of the individual field values to its corresponding language model. This metric is one way of identifying the potential of the field to distinguish one video from others in a search system [5].

The results shown in Table 3 confirm that the comment field is generally the least discriminating based on *KL*-divergence. For the most part, the title and the keyword fields are the most discriminating.

Table 3. *KL*-divergences for each video field.

Data Set	Title	Desc	Keywds	Comments	All
rand3500	6.77	6.14	6.82	4.98	5.19
pop500	5.46	5.14	5.35	5.68	2.59
rand10K	7.26	6.29	7.23	5.26	5.06
pop1500	5.89	5.38	5.72	5.38	2.33

4. DISTILLING INFORMATION CONTENT FROM COMMENTS

A consideration of the relative length of the average comment field explains its low *KL*-divergence. Intuitively, as a document (i.e., the comment field) gets longer, its divergence from the "background" language model decreases. (In separate experiments – not shown – we verified this phenomenon on the comment field and on the WT10G Web corpus.) In other words, the comment field becomes the language model if its size relative to the other fields is great enough.

We contend that, as a document gets longer, however, it will contain more discriminating information – as well as less discriminating information. To verify this, we identify the terms "most associated" with the comment field and see if these terms are unique to the field. We do this by pruning all but the "top terms" of each video's comment field and compare these terms to the background language model. We identify top terms with a variation of TF-IDF score (where TF measures the number of times a term appears in the video's comment field and IDF measures the number of videos' comment fields in which the term appears, as analogous to the typical definition of TF-IDF). We consider the top 68 unique terms to make the number comparable to that which is typically available in the title, description and

keyword fields, combined. (Recall our discussion on the results shown in Table 2.)

As shown in Figure 1, the *KL*-divergence of the top 68 comment terms increases quickly with the number of comment terms. The *KL*-divergence stabilizes at approximately 7.6 when the number of comment terms reaches 250 (when most of the terms are unique to the comment). This *KL*-divergence exceeds that of all the other fields (Table 3), indicating its potential in discriminating videos.

This result shows that longer comment fields contain more discriminating information. However, it is also likely that the rate of discriminating terms in comment fields decreases with comment length. Therefore, while we claim that longer comment fields contain more discriminating information, the rate at which we yield this information should decrease as the comment field gets longer. In any case, the long comment fields are more discriminating than the other fields.

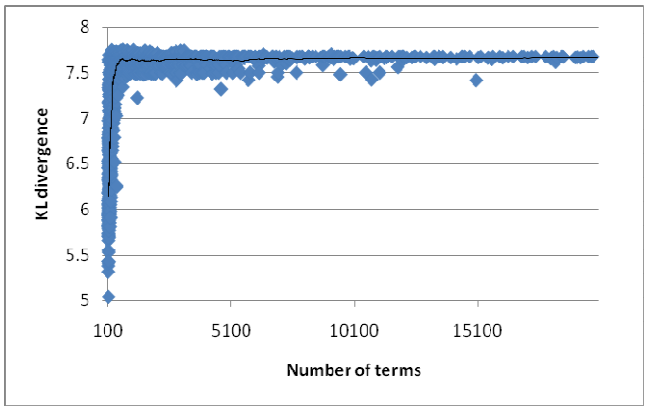

Figure 1. *KL*-divergences of the top 68 terms in each comment field as a function of number of terms in the comment field with the rand3500 data set (the trendline indicates the 50-point moving average).

Note that we only consider comment fields with at least 100 terms. With fewer terms, the comments often lacked 68 *unique* terms, making their *KL*-divergences as a function of length unstable, obscuring the results. Also, experiments with different numbers of top terms yielded similar, predictable results.

Table 4. Overlap percentage of top 30 terms and various fields with the rand3500 data set.

N	Title	Description	Keywords	Comments
10	12.58%	22.44%	31.93%	52.05%
20	10.05%	18.71%	30.24%	52.24%
30	8.14%	15.49%	27.90%	52.41%

4.1 Potential Impact of Comments on Query Accuracy

To estimate the potential of using comment terms to improve search accuracy, we use a technique described in [4] that effectively identifies the terms that are most likely to occur in a query that retrieves a given document. For each video, we extract the top *N* of these terms and calculate their overlap with the various video information fields. Note that the overlap is not necessarily disjoint, so the overlap percentages may exceed 100%.

The results in Table 4 show that most of these terms come from the comments. Of course, the comment field contains many more terms than the other fields, so the overlap will be greater. (For example, the title field's overlap is limited because titles generally contain fewer than 30 terms.) But the point is that it is exactly the size of the comment field that is the source of its potential. Although it contains many meaningless terms, it also contains a lion's share of the top terms. This suggests including comment terms in queries can improve search accuracy.

4.2 Waiting for Comments

One of the problems with using comments in search is that they take time for users to generate. In the results discussed in Section 4, we need about 250 comment terms before the *KL*-divergence stabilizes. If we assume each comment is 13 terms long, then we would need about 20 comments to yield 250 terms.

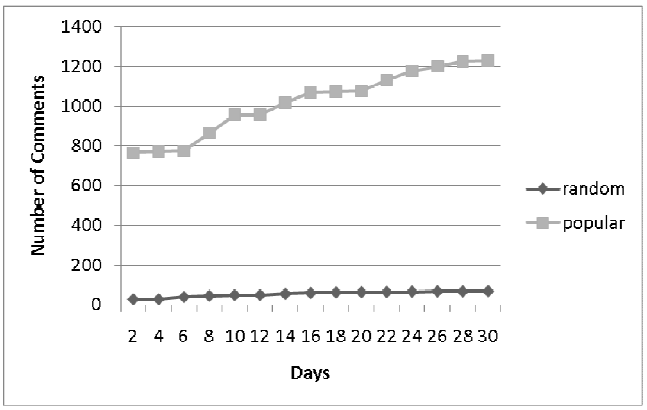

Figure 2. Number of comments as a function of time for the small data set.

Based on our data, popular and random videos receive approximately 20 and 1.4 comments per day, respectively. Therefore, popular videos collect enough comments in one day and random videos require about 2 weeks to yield enough terms to be useful for search. In Figure 2, we show the number of comments for the data set as a function of time. Popular videos are commented at a higher rate as expected, but both types of videos have a consistent increase in the number of comments.

Table 5. Analysis of DTC results on the rand3500 data set with length 3 top-IDF queries.

MRR Improvement	# of Videos	Avg(Len(DT))	Avg(Len(C))	|C∩K| / |K|	|DT∩K| / |K|
-1.00 - -0.75	40	19.225	719.475	0.2734	0.5314
-0.75 - -0.50	67	22.8209	345.6268	0.2904	0.4992
-0.50 - -0.25	188	27.1436	386.6968	0.2372	0.5081
-0.25 – 0.00	165	48.3273	118.8242	0.1764	0.5383
0	2576	33.5093	277.5613	0.2557	0.5617
0.00 - 0.25	152	58.4145	304.3158	0.3600	0.4049
0.25 - 0.50	151	45.4172	533.2848	0.4766	0.4402
0.50 - 0.75	38	66.2105	492.9474	0.3900	0.5529
0.75 - 1.00	116	37.2931	895.0776	0.6260	0.3291

5. EXPERIMENTAL RESULTS

5.1 Data Set and Metrics

We use data sets mentioned in Section 1 for our experiments. We simulate user queries by removing the keyword field from the video data set and using them to generate known-item queries. From the keyword set, we generate queries in two ways:

- top-IDF – Top-IDF queries are generated by the top *K* terms in the keyword field, where IDF is computed based on keyword fields.
- random – Random queries are generated by randomly picking *K* terms from the keyword field.

In the alternatives above, we use *K* values 2, 3, and 4 as these are the most common query lengths in Web and P2P applications [7][8].

We generate queries in this way because keywords are meant to help users index content. Top-IDF queries are meant to simulate users who generate very specific queries and their generation is similar to the query generation techniques described in [4][9]. Random queries are appropriate if we assume that all keywords are appropriate for queries.

Note that the choice of using the keyword field to create queries is somewhat arbitrary. Recent work shows that the terms used as keywords do not necessarily match those used in user queries [20]. For example, people tagging music would use the terms associated with genre, such as "pop," whereas people generally do not search for music via genre – title and artist are more likely in cases of known-item search. In future work, we will consider queries generated by other techniques described in [4][9].

Because we use keywords to generate queries, we also strip the keywords from the data set that we index. If we did not do this, then the baseline query accuracy would be so high – given our experimental setup – that we would not be able to reasonably run experiments that would show any meaningful positive change. One might worry that stripping keywords from the data set will result in an artificially low baseline for performance because keywords are expected to match queries very precisely. However, referring again to the results from [20], keywords do not necessarily match query behavior. Furthermore, the title and description fields were shown in Table 3 to be at least as discriminating of the video as the keywords, so these fields could have been chose as well as sources for queries.

In any case, our goal is to show whether the addition of comments can improve query performance over not using them. We could have therefore generated queries from any field provided that we remove that field from the data that is indexed. A positive result, therefore, would suggest that indexing comments in addition to all of the other fields is beneficial to query accuracy.

Because we are assuming known-item search, we use *MRR* as our main performance metric, defined as the average reciprocal rank of the desired result over all queries:

$$MRR = \frac{1}{N_Q}\sum_{i=1}^{N_Q}\frac{1}{r_i}$$

In the expression above, N_Q is the number of queries issued and r_i is the rank of the known item in the result set of query *i*. *MRR* is a metric that ranges from 0 to 1, where *MRR* = 1 indicates ideal ranking accuracy.

The data are indexed in the Terrier search engine [6]. Each of the videos is searched for via their respective queries.

Table 6. Query performance with and without comments with various query lengths on the rand3500 data set.

Query Type	Query Length	DT	DTC	Pct Change
top-IDF	2	0.5912	0.6045	2.24%
top-IDF	3	0.6645	0.6761	1.74%
top-IDF	4	0.7064	0.7136	1.01%
random	2	0.4697	0.4761	1.36%
random	3	0.5736	0.5839	1.80%
random	4	0.6377	0.6459	1.29%

5.2 Basic Results

In our first experiment, we test the impact of indexing queries. We issue the queries twice. First, we issue the queries on an index that contains the title and description (a setup that we refer to as DT) of each video, but not the comments. Second, we issue

the queries on an index that contains the title, description, and comments (a setup that we refer to as DTC) of each video.

In Table 6, we show query performance for different query lengths when the index does not and does contain comments. The results show that there is a consistent difference in *MRR* in the range of about 1% to 2% when using the comments on the rand3500 data set.

We search for the source of the performance improvement by dividing the results of each query into buckets based on the impact that the comment field has on the query results and then search for correlations between the change in *MRR* and "features" of the video data to which the query corresponds. The goal is to find some correlation between *MRR* improvement and a video feature. We considered several features of the video data, including, the lengths of the various fields in terms of the number of unique terms and the similarities between the fields.

A subset of our correlation analysis is shown in Table 5. Each bucket corresponds to a 0.25 point difference in *MRR*. We see that approximately 450 videos have their *MRR*s improved and about the same number have their *MRR*s worsened. Most videos (2,576 or about 75%) are not affected by the addition of comments.

We see that the length of the title and description fields have little impact on *MRR*. There is no clear correlation between them and change in *MRR*.

On the other hand, both the length of the comment field and the similarity between the comment and keyword fields are correlated with *MRR* change. Note that the similarity between the comment and keyword field is measured by how much the comment field covers the keyword field:

$$\frac{|C \cap K|}{|K|}$$

The coefficient of correlation between the similarity of the comment and keyword fields and the change in *MRR* is 0.7589. The coverage of the keyword field is also related to the length of the comments. If we remove the comment length of the first row of Table 5, then the coefficient of correlation between the change in *MRR* and the length of the comment field is 0.7214. (With the first row, the coefficient of correlation is 0.2964.) Finally, the coefficient of correlation between the length of the comment field and the similarity between the comment and keyword fields is 0.9351 without the first row of data and 0.7552 with the first row of data.

There is also a negative correlation between the similarity of the title and description fields with the keyword field (|DT∩K| / |K|) and *MRR* (-0.5177) and between |DT∩K| / |K| and |C∩K| / |K| (-0.8077). These results show that in the cases where titles and descriptions do not contain enough information to match the queries, then the long comment field is able to compensate. (We observe, for example, some videos with non-English descriptions and English keywords.)

The conclusion that we draw from these results is that comments help:

- *MRR* improves when the comments contain keywords (equivalently, query terms, since we generate queries from the keywords).
- Comments are particularly important when the title and description do not contain the appropriate terms that match the query.
- Longer comment fields are more likely to contain keywords.

So, despite all of the irrelevant terms contained in the comments – particularly long comments – the existence of the relevant terms helps.

In our next experiments, we run the same test on the pop500 data set. The results of this experiment show how comments affect the search for videos that users actually want to find (i.e., popular videos).

Table 7. Query performance with and without comments with various query lengths on the pop500 data set.

Query Type	Query Length	DT	DTC	Pct Change
top-idf	2	0.5193	0.6239	20.14%
top-idf	3	0.5984	0.6709	12.12%
top-idf	4	0.6561	0.7150	8.99%
random	2	0.4895	0.5455	11.44%
random	3	0.5592	0.6010	7.48%
random	4	0.6105	0.6650	8.93%

Table 8. Analysis of DTC results on the pop500 data set with length 3 top-IDF queries.

MRR Improvement	# of Videos	Avg(Len(C))	|C∩K| / |K|
-1.00 - -0.75	18	2267.7	0.5071
-0.75 - -0.50	10	2842.0	0.6795
-0.50 - -0.25	45	2481.2	0.6927
-0.25 – 0.00	16	2328.4	0.7800
0	267	3337.4	0.6546
0.00 - 0.25	41	3668.8	0.6731
0.25 - 0.50	38	4103.5	0.7710
0.50 - 0.75	10	8933.5	0.8430
0.75 - 1.00	53	5230.6	0.8204

Our results are shown in Table 7. Comments are much more effective on popular videos. For top-IDF queries, the *MRR* improvement ranges from 9% to 20%. For random queries, the *MRR* improvement ranges from 7% to 9%. Results are somewhat better for shorter queries and for top-IDF queries.

In Table 8, we again search for features that are correlated to the change in *MRR*. First, we notice that a greater percentage of videos are affected by the comments in the pop500 data set than in the rand3500 data set (about 47% versus 26%). Of the affected videos, 89 videos' *MRR*s worsened and 142 videos' *MRR*s improved with the use of comments.

We again see a correlation between the similarity between the comment and keyword fields and the change in *MRR*. The coefficient of correlation between these two variables is even greater than that of the rand3500 data set: 0.8295 versus 0.7589. The correlation between the length of the comment field and the change in *MRR* is 0.7404 with the pop500 data set versus 0.7214 with the rand3500 data set.

We summarize the performance results on the rand3500 and pop500 data sets in Table 9. We see that comments are clearly more effective on popular data. The change in *MRR* is greater and the number of videos whose *MRR* improves is greater. This is likely because of the similarity between the comment and keyword fields.

Table 9. Summary of the performance differences between experiments on rand3500 and pop500 data sets with length 3 top-IDF queries.

Metric \ Data Set	rand3500	pop500
MRR change	0.0175	0.1212
Pct of video *MRR*s improved	0.1306	0.2840
Pct of video *MRR*s worsened	0.1314	0.1780
Correl(*MRR* change, len(C))	0.7214	0.7404
Correl(*MRR* change, $\|C \cap K\| / \|K\|$)	0.7589	0.8295

Table 10. Query performance with and without comments with various query lengths on the rand10K data set with top-IDF queries.

Query Length	DT	DTC	Pct Change
2	0.6271	0.6442	2.65%
3	0.6842	0.7052	2.98%
4	0.7199	0.7388	2.56%

Table 11. Analysis of DTC results on the rand10K data set with length 3 top-IDF queries.

MRR Improvement	# of Videos	Avg(Len(C))	$\|C \cap K\| / \|K\|$
-1.00 - -0.75	121	838.7686	0.3614
-0.75 - -0.50	167	481.4551	0.3633
-0.50 - -0.25	421	411.7316	0.3627
-0.25 – 0.00	552	246.5326	0.2078
0	7248	291.2323	0.2947
0.00 - 0.25	538	480.7993	0.4043
0.25 - 0.50	461	553.7852	0.4892
0.50 - 0.75	121	724.9669	0.5224
0.75 - 1.00	349	874.0029	0.6521

5.2.1 Results on Larger Data Sets

To simplify our explication, in this section, we only report results using the top-IDF-type queries. Also, as done above, if no query length is specified, we use queries of length 3.

As shown in Table 10, comments also improve the query performance in the rand10K data set. *MRR* improvements are about 3%, which is similar to the improvements with the smaller, rand3500 data set (Table 6).

An analysis of the *MRR* change table for rand10K (Table 11) reveals that there is a again correlation between the length of the comments and the change in *MRR* (0.7515), and the similarity between the comment and keyword fields and the change in *MRR* (0.7064). In this case, most of the videos (72%) are unaffected by comments, however, while 13% have their *MRR*s worsened and 15% have their *MRR*s improved.

Table 12. Query performance with and without comments with various query lengths on the pop1500 data set with top-IDF queries.

Query Length	DT	DTC	Pct Change
2	0.5596	0.5991	6.59%
3	0.6228	0.6465	3.67%
4	0.6592	0.6818	3.31%

Table 13. Analysis of DTC results on the pop1500 data set with length 3 top-IDF queries.

MRR Improvement	# of Videos	Avg(Len(C))	$\|C \cap K\| / \|K\|$
-1.00 - -0.75	86	1966.686	0.6623
-0.75 - -0.50	49	2972.674	0.7259
-0.50 - -0.25	128	2573.914	0.7766
-0.25 – 0.00	78	2611.885	0.7254
0	706	2323.965	0.7587
0.00 - 0.25	165	2789.746	0.7826
0.25 - 0.50	121	2619.744	0.8201
0.50 - 0.75	24	3690.25	0.8609
0.75 - 1.00	136	3170.044	0.8430

Again, as shown in Table 12, the improvement in MRR with popular data is greater than that with random data. With the pop1500 data set, the percentage MRR improvement ranges from 3% to 7% compared with 3% for the rand10K data set. In this case, 47% of the videos MRRs are unaffected by the comments, 23% are worsened, and 30% are improved.

The coefficient of correlation between MRR change and comment length is 0.6769 and the coefficient of correlation between MRR change and similarity of comment and keyword fields is 0.9192. Again, long comment fields are able to substitute for keywords in search.

The fact that MRR is better for popular data has been shown in other work (e.g., [9]). This is clearly due to the fact that popular data have more comments. This result is significant as it shows that increasing the number of comments does not only increase the ability for videos to naively match queries, but also increases the ability for queries to distinguish the relevant videos.

5.3 Improving our Results

Our next goal is to improve on our improvement-in-MRR results based on our observations. If we detect a correlated feature, we use the correlation in our indexing strategy.

Our main observation is that as the length of the comments field increases, so does its effectiveness in search. Therefore, we should only index comments if they are above a certain length.

We also acknowledge that there is a correlation between the change in MRR and the similarity between the comment and keyword fields. However, as there is also a correlation between comment length and similarity, we roughly cover both features by considering just the comment length.

Our first strategy is to index comments only if they are above a given length threshold. We refer to this strategy as "length-selective indexing." We show the experimental results in Table 14, where the threshold is in terms of number of terms (words).

The performance of length-selective indexing is negative. MRR consistently decreases with increasing thresholds. The problem with this strategy is that it creates a situation where certain videos are too eager to match queries. In other words, videos that have their comments indexed are ranked higher than other videos compared with the base case regardless of whether they are relevant to the query or not. Because videos are only relevant to a single query (by definition of MRR), MRR must decrease with this type of indexing.

Table 14. Percentage change in MRR with length-selective indexing on the rand3500 data set.

Len(C) Threshold	Pct Change in MRR
0	0
50	-0.19%
100	-0.40%
150	-0.64%
200	-0.97%
250	-1.09%
300	-1.25%
350	-1.38%
400	-1.60%
450	-1.76%
500	-1.95%

This was not a problem in the case where all videos' comments were indexed because the "eager matching" problem is offset by the fact that all videos (with comments) have additional terms associated with them. We expect that the additional terms contained in the comments are more likely to help match relevant queries.

5.3.1 Comment Pruning

The problem with length-selective indexing is that it un-uniformly adds "noise" to the description of videos making them match irrelevant queries. If noise were applied uniformly to all videos, then such a problem would be attenuated. The problem is that noise still causes the incorrect matching of query to results.

This observation inspires a solution whereby we index each video with its comments, but then prune away noise from the comments, leaving only the most relevant terms in the description of each video. This solution is expected to do two things:

1. Reduce the irrelevant matches of a query, and
2. Decrease the size of the index.

The technique we use to prune the comment field is that which was proposed to shrink indices in [4], known as document-centric pruning. With document-centric pruning, each term in each document is ranked based on its contribution to the *KL*-divergence [13] of the document to the background language model. The lower-ranked terms are removed from the documents before they are indexed. This technique was shown to be able to shrink the index by up to 90% with little loss in precision.

In these experiments, we prune a percentage of the comments of each video. We assume that there is a "fixed rate" at which terms that are useful to search accuracy appear in the comments. If this rate is *r*, then a comment field of length len(C) will have *r*len(C) useful terms. If we pick a pruning rate of *r*, then all of the terms left in the comment field will be useful.

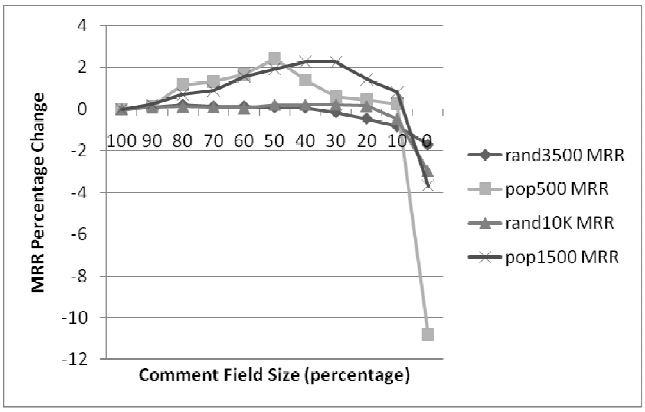

Figure 3. Percentage change in MRR for different comment field sizes for various data sets.

In Figure 3, we see the effect that comment pruning has on MRR. The data on the left of the figure corresponds to a complete comment fields, whereas the data on the right corresponds to no comments. We see that pruning initially increases the MRR for all data sets. MRR then drops dramatically as the comment field size decreases to zero.

The effect of pruning is more pronounced for the popular data sets than for the random data sets. With the random data set, the maximum MRR percentage increase is about 0.7% (60% pruning on the rand10K data set), while with the popular data set, the maximum MRR percentage increase is 2.4% (50% pruning with the pop500 data set).

The reason for this is that the random data sets' comment fields contain so few comments in the first place. They are therefore less likely to contain terms that make eagerly match irrelevant results. Second of all, the *MRR* improvement with using comments with random videos is low in the first place, suggesting the marginal impact that such comments have. We do not expect there to be much of an increase in performance with pruning.

Based on these results, a pruning rate of 50% is reasonable choice. We are able to eliminate half of the index overhead introduced by the comments and are safe from losing *MRR*

performance. *MRR* starts to decrease first with the rand3500 data set with 70% pruning.

6. RELATED WORK

In [14], the authors consider the impact that indexing blog comments have on query recall. Their conclusion is that recall is boosted by the comments, that they are useful. This result is expected, but little consideration was given to the precision of the results.

In [17], it was shown the comments do have discriminating power. The authors clustered Blog data and by using high weights for comments, were able to improve the purity and decrease the entropy of their clusters significantly.

Much of the work on "social" Web sites – where users are free to modify the metadata associated with shared data – focus on "tag" analysis, where a tag is a keyword that a user can associate with data to, say, make it easier to index. Findings related to tag analysis are they indicate data popularity and are useful in describing content [15][16][18][19]. This is somewhat orthogonal to our goal of determining if *casual* user comments can help improve search accuracy.

Table 15. Summary of potential improvement with comments.

Data Set	No Comments	Best *MRR*	Percent Change
rand3500	0.6645	0.6775	1.96%
pop500	0.5984	0.6872	14.84%
rand10K	0.6842	0.7068	3.30%
pop1500	0.6228	0.6612	6.17%

7. CONCLUSION

Our results show that comments indeed improve the quality of search compared with just using titles and descriptions to describe videos. They are particularly useful with popular videos, where the *MRR* is lower than with random videos (Table 15).

This result is not a given, however, as some queries actually do worse with comments. The reason for these cases of decreased accuracy is that the videos with fewer comments become "buried" by those with more comments in search results.

The problem of skew in result sets toward videos with larger comment fields can be addressed by well-known index pruning techniques – which also shrink the size of the index. Index pruning technique work by removing terms deemed less distinguishing or relevant to the particular "document." Applying index pruning to the comments further improves accuracy by up to about 2% (with a decrease in index size of up to 70%). Overall, accuracy improved by up to about 15% as shown in Table 15.

Our ongoing work includes further analyses and characterizations of comment terms and their impact on search accuracy. For example, our observation that comments work best when they contain query terms (Section 5.2) and when the title and description fields do not may suggest that we should only index comments when they are "different" than the title and description.

8. REFERENCES

[1] Jindal, N. and and Liu, B, "Identifying Comparative Sentences in Text Documents," In *Proc. ACM SIGIR*, 2006.

[2] Li, X., Guo, L. and Zhao Y., "Tag-based Social Interest Discovery," In *Proc. WWW*, 2008.

[3] Heymann, P., "Can Social Bookmarks Improve Web Search?" In *Proc. ACM Conf. Web Search and Data Mining* (WSDM), 2008.

[4] Buttcher, S. and Clarke, C. L. A., "A Document Centric Approach to Static Index Pruning in Text Retrieval Systems," In *Proc. ACM CIKM*, 2006.

[5] Ponte, J. M. and Croft, W. B., "A language modeling approach to information retrieval," In *Proc. ACM SIGIR*, 1998.

[6] Terrier Search Engine Web Page. http://ir.dcs.gla.ac.uk/terrier/

[7] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. A., Frieder, O. "Hourly analysis of a very large topically categorized web query log." In Proc. *ACM SIGIR*, 2004, pp. 321-328.

[8] Yee, W. G., Nguyen, L. T., and Frieder, O., "A View of the Data on P2P File-sharing Systems." In Jrnl. Amer. Soc. of Inf. Sys. and Tech (JASIST), to appear.

[9] Azzopardi, L., de Rijke, M., Balog, K. "Building Simulated Queries for Known-Item Topics: An Analysis Using Six European Languages." In *Proc. ACM SIGIR*, 2007.

[10] Porter Stemming Web Site. http://tartarus.org/~martin/PorterStemmer/

[11] Jenkins, M.-C., Smith, D., "Conservative Stemming for Search and Indexing." In *Proc. ACM SIGIR*, 2005.

[12] Atkinson, K. SCOWL Word List. http://wordlist.sourceforge.net/

[13] Kullback, S., "Information theory and statistics." John Wiley and Sons, NY, 1959.

[14] Mishne, G., Glance, N. "Leave a Reply: An Analysis of Weblog Comments." In *Third Workshop on the Weblogging Ecosystem*, 2006.

[15] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z. "Optimizing Web Search Using Social Annotations." In *Proc. WWW*, 2007.

[16] Zhou, D., Bian, J., Zheng, S., Zha, H., Giles, C. L. "Exploring Social Annotations for Information Retrieval." In *Proc. WWW*, 2008.

[17] Li, B., Xu, S., Zhang, J. "Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments." In *Proc. ACM Southeast Regional Conf.*, 2007.

[18] Dmitriev, P. A., Eiron, N., Fontoura, M., Shekita, E. "Using Annotations in Enterprise Search." In *Proc. WWW*, 2006.

[19] Heymann, P., Koutrika, G., Garcia-Molina, H.."Can Social Bookmarks Improve Web Search?" In *Proc. Int'l. Conf. on Web Search and Web Data Mining*, 2008.

[20] Bischoff, K., Firan, C. S., Nejdl, W., Paiu, R. "Can All Tags be Used for Search?" In *Proc. ACM CIKM*, 200