Shame to be Sham: Addressing Content-Based Grey Hat Search Engine Optimization

Fiana Raiber
Technion — Israel Institute of Technology
fiana@tx.technion.ac.il

Kevyn Collins-Thompson
Microsoft Research
1 Microsoft Way
Redmond, WA, USA 98052
kevynct@microsoft.com

Oren Kurland
Technion — Israel Institute of Technology
kurland@ie.technion.ac.il

ABSTRACT

We present an initial study identifying a form of content-based *grey hat* search engine optimization, in which a Web page contains both potentially relevant content and manipulated content: we call such pages *sham* documents, because they lie in the grey area between 'ham' (clearly normal) and 'spam' (clearly fake). Sham documents are often ranked artificially high in response to certain queries, but also may contain some useful information and cannot be considered as absolute spam. We report a novel annotation effort performed with the ClueWeb09 benchmark where pages were labeled as being spam, sham, or legitimate content. Significant inter-annotator agreement rates support the claim that there are sham documents that are highly ranked by a very effective retrieval approach, yet are not spam. We also present an initial study of predictors that may indicate whether a query is the target of shamming.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

Keywords: sham, spam, search engine optimization

1. INTRODUCTION

The Web constitutes an adversarial retrieval setting: in Web search, content creators try to have their pages ranked high in response to queries, to increase visibility and traffic to their sites. To that end, they often employ various search engine optimization (SEO) techniques [11]. Search engines, on the other hand, aim to rank pages in response to queries by the overall relevance of their *content* and strive to minimize the effect of potential adversarial methods.

A distinction is made between *white hat* SEO, which typically does not involve deceptive techniques and obeys search engine anti-spam guidelines, and *black hat* SEO, which uses deceptive content, links and other manipulation that ultimately may result in the site using them being penalized or banned by search engines. The black hat SEO area has attracted much research attention; e.g., there is a large body of work on spam classification [11, 5, 8, 19], where a page is deemed spam if it bears no useful content. *Grey hat* SEO, on the other hand, is an intermediate form of search ranking manipulation that yields neither a spam page, nor a (very) high-quality representative of relevant content for queries seeking information on that topic. Grey hat SEO has attracted very little research attention in the literature [13]. There is no rigorous definition of what makes for a grey hat SEO attempt: because grey hat documents can often contain relevant information as well as SEO content, identifying such documents is often a subjective decision.

In this paper we address the problem of identifying a specific, highly important type of grey hat SEO: pages whose *content* has been manipulated so as to have the page ranked higher than it should with respect to *some* queries. However, these pages are not spam in the absolute sense, in that they can still satisfy information needs. We use the term *sham* to refer to such pages[1]. For example, excessive use of query variants can turn an otherwise useful page to sham but not to spam, along with more obvious manipulations like keyword stuffing [11]. An example of a *sham* passage is shown in Figure 1.

By definition, sham pages can have considerable negative impact on the effectiveness of relevance ranking. For example, the frequency of query terms in them can be a misleading signal as to the actual information needs these pages can satisfy and the extent to which they can satisfy them. Our motivation for focusing on *content-based* shamming is threefold. First, relevance is typically determined by the ability to satisfy an information need using the page content. Second, content-based features are highly important when applying learning to rank methods for Web search [16]. Third, it is relatively easier to manipulate the content of pages compared to other derived relevance signals such as clickthrough rate, hyperlink structure, etc.

This paper provides two main contributions toward an initial understanding of the nature of *sham*. First, we describe the results of a sham annotation pilot effort applied to the publicly available ClueWeb09 benchmar[2] [7]. Documents that were highly ranked in response to queries by an effective learning-to-rank (LTR) method were classified into one of three categories: spam, sham, or legitimate. We found significant inter-annotator agreement for the task of labeling sham pages. Furthermore, many top-ranked pages were agreed to be sham: some were relevant to the queries, while most were not. Second, we present an initial study of predic-

[1]Appropriately, in English the word 'sham' can denote something fake or not genuine.
[2]www.lemurproject.org

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Your good accident lawyer will deal with whiplash injury claims, spinal injury claims, personal injury claims, and back injury claims. It is the job of the car accident lawyer to negotiate with the insurance company to get you a high auto accident insurance settlement as well as perform other duties such as...

Figure 1: Example *sham* passage for the query `[auto injury attorney]`. In sham documents, content is manipulated, e.g., with excessive related queries (underlined), while retaining some degree of relevance.

tors for when a query is likely to be the target of shamming efforts. Our results show that query-specific features outperform an approach that uses spam classification for this task as well as content-based query-independent document quality measures [2].

2. RELATED WORK

There is much work on spam classification (e.g., [11, 14, 20, 8, 5, 19]) and on nullifying its effects on retrieval effectiveness [14, 9]. As already noted, the task of sham classification that we pursue here is different from spam classification by definition. Specifically, sham documents are manipulated for specific queries, and do bear useful content, while spam documents are absolutely useless. There has been some work on using query-dependent features to classify spam in its absolute query-independent sense [20]. We use query-dependent features for sham classification.

Using query-independent document quality measures is known to improve retrieval effectiveness [3, 2]. We note that sham documents are not necessarily of low quality. Furthermore, we show that the methods we study for the task of predicting whether a query is the target of shamming outperform methods that use content-based document quality measures which are very effective for search [2].

The realization that "not all content that complicates ranking is spam" was echoed in previous work [13]. However, we are not aware of any previous studies that study the difficulty of labeling Web pages as grey hat SEO, or using pre- and post-retrieval features to predict how likely a query is to be the target of such SEO attempts.

3. SHAM ANNOTATION PROCESS

As a first step in examining the shamming phenomenon, we conducted an annotation effort, wherein a total of 300 documents, 30 queries × 10 documents per query, were evaluated. (Further details regarding the set of documents and the queries are provided in Section 5.) The documents were evenly divided between five Ph.D. and M.Sc. students in the Information Retrieval field. Each document was labeled by two annotators as being either spam, sham, or legitimate page. In case of a disagreement between the first two annotators, a third annotator was asked to label the document. Thus, for each document we have either two or three labels.

Annotators were instructed to label a document as sham if any of the following conditions was met. (i) The document contained text that did not seem to satisfy any information need, including information needs satisfied by other parts of the same document (ii) The page contained text that appeared to be related to information needs satisfied by this page, but the text contained many artificial, repeated, or otherwise unnecessary extra words, phrases or sentences added to the page solely for the purpose of promoting the page in the result lists; and (iii) The page contained sections of content that were copied from other pages (e.g., from a Wikipedia page) for presumably the sole purpose of promoting the page. A document was labeled as spam if it contained no useful information *at all* for satisfying *any* information need.

The annotators were asked to base their decision only upon the *content* of the document, and to ignore non-content elements such as incoming and outgoing links, page URL, site domain, or ads and sponsored links that might appear on the page. The documents were presented to the annotators in a random query-independent order and the query itself was not presented.

4. PREDICTORS FOR QUERY SHAMMING

We next examine the problem of predicting which queries are more susceptible to sham, as measured, e.g., by the fraction of sham documents in the top 10 results. To this end, we study the use of various predictors which were developed to estimate retrieval effectiveness [4]. As we show below, some of these are negatively correlate with the percentage of sham, typically because most sham documents are non-relevant, while others have a positive correlation.

Pre-retrieval predictors are based on properties of the query and the corpus. Post-retrieval predictors utilize, in addition, information about the result list of documents, $\mathcal{D}_{res}$, which was returned in response to the query [4]. Details regarding the retrieval model used in our experiments are provided in Section 5. In what follows we present the different classes of predictors analyzed for our prediction task.

Pre-retrieval predictors. The **SumSCQ** predictor measures the similarity between the query and the collection using the sum of the TF.IDF values of the query terms [21]. **SumVar** [21] is the sum over query terms of the variance of the TF.IDF values of the term in documents in the corpus in which it appears. The **SumIDF** predictor is the sum of the IDF values of the query terms [10]. Another pre-retrieval predictor that we use is **QueryLength**. Since sham documents are assumed to negatively affect relevance ranking, a lower pre-retrieval predicted value is assumed to be correlated with higher percentage of sham.

Post-retrieval predictors. The **Clarity** [10] prediction value is the KL divergence between a relevance language model [15], $\mathcal{R}$, induced from the documents in the result list $\mathcal{D}_{res}$ and the (unsmoothed) language model of the collection. Formally, let $p(w|d)$ and $s(d)$ be the probability assigned to term w by an unsmoothed language model induced from document d, and the score assigned to d by the search algorithm that was used to create $\mathcal{D}_{res}$, respectively. Then, the probability assigned to w by $\mathcal{R}$ is defined as $p(w|\mathcal{R}) \stackrel{def}{=} \sum_{d \in \mathcal{D}_{res}} p(w|d) \frac{s(d)}{\sum_{d_i \in \mathcal{D}_{res}} s(d_i)}$. The assumption is that sham documents for a given query are similar to one another; for example, due to the repeated terms that they contain. Thus,

these documents presumably form a cluster whose language model is focused with respect to that of the corpus.

We also consider a variant of **WIG** [22] which is simply the average retrieval score in $\mathcal{D}_{res}$: $\frac{1}{|\mathcal{D}_{res}|}\sum_{d\in\mathcal{D}_{res}} s(d)$. The premise is that sham documents are likely to be assigned high retrieval scores, as by definition they have been manipulated to be promoted in result lists.

Query-independent document-quality measures. The entropy of the term distribution in a document, **Entropy**, the stopwords to non-stopwords ratio in a document, **SW1**, and the percentage of stopwords that appear in the document, **SW2**, were shown to be highly effective content-based query-independent quality measures for Web search [2], and for predicting query performance over the Web [17]. We used INQUERY's stopword list [1]. The document **PageRank** score [3] is another measure widely used for improving retrieval effectiveness. We use these measures for our sham prediction task. The prediction value for the result list $\mathcal{D}_{res}$ is the sum of the per-document values assigned by a measure. While Entropy, SW1 and SW2 are based on the content of a document, PageRank is based on the hyperlink structure. The content-based measures quantify the presumed textual content breadth in the document. We hypothesize that the content breadth in sham documents is low, as there is focus on a small subset of terms that are the target of shamming. Conversely, since by definition sham documents can still satisfy information needs, we assume that the PageRank score of such documents is likely to be high.

Spam-based predictor. To study the connection between spam and sham documents, we use the recently proposed **NS** predictor [17]. Documents in $\mathcal{D}_{res}$ that are classified as spam by Waterloo's spam classifier are treated as "non-relevant"; the rest are treated as "relevant". Then, the mean average precision (MAP) at cutoff $|\mathcal{D}_{res}|$ is computed based on these artificially created "relevant" and "non-relevant" labels. This score is then multiplied by the total number of "relevant" (non-spam) documents.

Query-log-based predictors. For each query, we derived the following features from the query logs of a commercial search engine. **RawImpressions**: the total frequency the query (with stopwords) was used to search; **ClicksOnResult**: percentage of clicks on the search engine results page (SERP) that were on a search result *vs.* other parts of the page, such as query suggestions; **ClicksForPaging**: the percentage of clicks on the SERP used to get to the next or previous page of results. For all features we used one month of traffic from January 2013 originating from the U.S. locale.

5. EVALUATION

Our experiments were conducted using the ClueWeb09 Category B collection, which contains about 50 million English Web pages. We used queries 1-150 from TREC 2009-2011. The Indri toolkit (www.lemurproject.org/indri) was used for experiments. We applied Krovetz stemming upon queries and documents, and removed stopwords on the INQUERY list [1] only from queries.

To create a highly effective result list, $\mathcal{D}_{res}$, we did the following. First, for each query q, the documents in the collection were ranked using the negative cross entropy between the unsmoothed unigram language model induced from q, and the Dirichlet-smoothed unigram language model induced from a document (with the smoothing parameter μ set to 1000). Then, following previous work [9], documents assigned by Waterloo's spam classifier with a score below 50 were removed top to bottom from that ranking until 100 (presumably non-spam) documents were accumulated. These documents were then re-ranked using SVM^{rank} [12] applied with 130 features. Ten-fold cross validation was performed.

We used a subset of the features used by Microsoft's learning to rank datasets[3]. Our features do not include the Boolean Model, Vector Space Model, LMIR.ABS, and the number of outgoing links. The SiteRank score, the two quality measures, QualityScore and QualityScore2, and the click-based features were also not included as these information types are not available for the ClueWeb09 collection. Three additional features that we used, and which were found to be highly effective for Web retrieval [2], are Entropy, SW1, and SW2 which were described in Section 3. As is the case with the features mentioned thus far, these three features were also computed separately for all the text in the document, its anchor text, URL, body, and title. Waterloo's spam classifier score [9] is another feature that was used. The BM25 score was computed with $k1$=1 and b=0.5, following experiments with $b \in \{0.5, 0.75\}$ and $k1 \in \{1, 1.2, 1.5, 1.7\}$. LMIR.DIR and LMIR.JM, the language-model-based features, were computed with μ=1000 and λ=0.9, respectively. Lastly, 30 queries were randomly sampled, and the top 10 non-Wikipedia documents were used to form $\mathcal{D}_{res}$. We assume that Wikipedia pages are not sham.

We used two approaches to aggregate the labels assigned by the annotators to the documents. According to the first approach, a document is considered sham if it was labeled as such by at least one of the annotators. This 'weak' label definition is henceforth referred to as **AtLeastOne**. The second approach, denoted **AtLeastTwo**, is more strict: a document is considered sham if it was marked as sham by at least two annotators.

To evaluate the quality of the various predictors, we report Pearson's correlation [4] between the values assigned by a predictor to each of the tested queries, and the percentage of sham documents computed for a query, according to the AtLeastOne and AtLeastTwo approaches described above.

The number of terms used to construct the relevance language model, which is used by the Clarity predictor, is set to 50. Given that documents assigned with a score below 50 by Waterloo's spam classifier are removed in the process of creating $\mathcal{D}_{res}$, to determine the non-spam documents that are used by the NS predictor we use a threshold. The threshold is selected from $\{60, 70, 80, 90\}$ so as to optimize the prediction quality of NS. Documents assigned with a score above that threshold are considered as non-spam ("relevant").

Annotation Results. The inter-annotator agreement on the label pairs from the two main judges for each document was 0.69, computed using the free-marginal multi-rater kappa measure [18]. Overall, 79% of the 300 label pairs were in agreement. The average percentage of sham documents per query was 30% based on AtLeastOne labels, and 21% based on AtLeastTwo labels.

Query shamming predictor analysis. The correlations of the predictors described in Section 4 with the per-

[3]www.research.microsoft.com/en-us/projects/mslr

Predictor Type	Predictor Name	Sham Label Defn.	
		AtLeastOne	AtLeastTwo
Pre-retrieval	SumSCQ	−0.500	−0.401
	SumVar	−0.398	−0.337
	SumIDF	−0.424	−0.363
	QueryLength	−0.553	−0.432
Post-retrieval	Clarity	+0.276	+0.235
	WIG	+0.361	+0.351
Query-Independent	Entropy	−0.324	−0.242
	SW1	−0.001	−0.022
	SW2	−0.024	+0.046
	PageRank	+0.226	+0.129
Spam	NS	+0.099	+0.183
Query Log	RawImpressions	+0.178	+0.231
	ClicksOnResult	−0.216	−0.320
	ClicksForPaging	+0.277	+0.387

Table 1: Pearson's correlation (r) of various predictor variables (Section 4) with % sham documents in the top 10 results, for weak (AtLeastOne) and strict (AtLeastTwo) sham label definitions.

centage of sham pages retrieved for a query are presented in Table 1. We can see that the pre-retrieval predictors are negatively correlated with shamming. As these predictors were designed to predict the retrieval effectiveness of using the query for search, we can conclude that sham documents are prevalent in queries whose performance is predicted to be low. Indeed, we found that *about* 75% *of sham documents are non-relevant*, for both AtLeastOne and AtLeastTwo.

QueryLength is negatively correlated with the percentage of sham documents, for both weak (AtLeastOne) and strict (AtLeastTwo) sham label definitions. That is, a shorter query has a higher chance of being the target of shamming. This might be attributed to the fact that short queries are more ambiguous, and as a result documents returned in response to these queries might be promoting the page with respect to different information needs, not necessarily the information need that the current query expresses.

The post-retrieval predictors, Clarity and WIG, are positively correlated with the percentage of sham documents, as hypothesized in Section 4. However, the prediction quality is lower than that for the pre-retrieval predictors.

We can also see that the correlation between the query-independent document-quality measures and shamming is relatively low. Furthermore, the NS predictor is not correlated with sham. These findings suggest that sham documents might be of either high or low quality, and that there is no evident connection between sham and spam documents.

For query log features, the positive correlation of RawImpressions with the level of sham is consistent with the fact that frequent queries are more susceptible to shamming. ClicksForPaging is positively correlated with the percentage of sham, while the ClicksOnResult feature is negatively correlated. These findings resonate with the fact that shamming degrades retrieval performance.

Finally, for the task of predicting sham percentage for a given query at low (< 0.3), medium ($\in [0.3, 0.6)$), or high (≥ 0.6) levels, we performed ordinal regression [6] using *all the predictors* with default regression settings. The mean absolute error (MAE), the absolute difference between the predicted class ordinal and the true ordinal, averaged over 100 randomized trials with a 2:1 train/test split, was 0.532. For comparison, an oracle run using one rater's sham labels to predict sham levels derived from other raters' labels, had MAE 0.37; while always guessing category 'medium' had MAE 0.67. These initial results show that different predictors can be effectively integrated for predicting per-query sham levels.

6. CONCLUSION

We have provided an initial study on the identification of *sham* documents: a form of grey-hat SEO using *content-based* feature manipulation, along with an analysis of features associated with low- or high-sham queries. Even search engines that use highly effective ranking methods still suffer from sham documents in the top-most rankings that adversely affect retrieval quality. These findings call for a principled treatment of query-specific grey hat SEO.

Acknowledgments. We thank the reviewers for their comments. This work was supported by and carried out at the Technion-Microsoft Electronic Commerce Research Center.

7. REFERENCES

[1] J. Allan, M. E. Connell, W. B. Croft, F.-F. Feng, D. Fisher, and X. Li. INQUERY and TREC-9. In *Proc. of TREC*, pages 551–562, 2000.

[2] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proc. of WSDM*, pages 95–104, 2011.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of WWW*, pages 107–117, 1998.

[4] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool, 2010.

[5] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, 2006.

[6] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *J. Machine Learning Res.*, 6:1019–1041, 2005.

[7] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Proc. of TREC*, 2009.

[8] G. V. Cormack. TREC 2007 spam track overview. In *Proc. of TREC*, 2007.

[9] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465, 2011.

[10] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. of SIGIR*, pages 299–306, 2002.

[11] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. of AIRWeb*, pages 39–47, 2005.

[12] T. Joachims. Training linear SVMs in linear time. In *Proc. of KDD*, pages 217–226, 2006.

[13] T. Jones, R. S. Sankaranarayana, D. Hawking, and N. Craswell. Nullification test collections for web spam and SEO. In *Proc. of AIRWeb*, pages 53–60, 2009.

[14] V. Krishnan and R. Raj. Web spam detection with Anti-Trust rank. In *Proc. of AIRWeb*, pages 37–40, 2006.

[15] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proc. of SIGIR*, pages 120–127, 2001.

[16] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3), 2009.

[17] F. Raiber and O. Kurland. Using document-quality measures to predict web-search effectiveness. In *Proc. of ECIR*, pages 134–145, 2013.

[18] J. J. Randolph. Online kappa calculator (2008). Retrieved February 16, 2013, http://justus.randolph.name/kappa.

[19] D. Sculley and G. Wachman. Relaxed online SVMs for spam filtering. In *Proc. of SIGIR*, pages 415–422, 2007.

[20] K. M. Svore, Q. Wu, C. Burges, and A. Raman. Improving web spam classification using rank-time features. In *Proc. of AIRWeb*, 2007.

[21] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proc. of ECIR*, pages 52–64, 2008.

[22] Y. Zhou and B. Croft. Query performance prediction in web search environments. In *Proc. of SIGIR*, pages 543–550, 2007.