

# High precision document classification in over-abundant domains: A case study of web pornography search

Pieter van den Wankendon

William Johnson

Randy Feller

*Wangbiz! Labs*

*Las Vegas, NV 89168*

PVDW@WANGBIZ.COM

WILLY@WANGBIZ.COM

RFELLER@WANGBIZ.COM

## Abstract

We present a small, but important, set of techniques that were found to boost precision in web pornography search, as judged by volunteer annotators. Both structured and statistical approaches yielded statistically significant improvements in document classification. We detail the initial corpus-development process, as well as empirical trials that demonstrate the efficacy of our proposed techniques.

## 1. Introduction

A widely-perceived holy grail for Internet service providers is an effective filter for certain common document classes, for example sexually explicit images (Bosson et al., 2002). This paper presents research into a separate but related question: How to retrieve only documents from this class. Internet pornography accounts for such a large percentage of on-line documents that any basic search technique will return many many true positives, which are virtually identical in the degree to which they meet the search criteria. Recall is not a critical issue. If, however, one is truly searching for documents of this class, false positives may dilute the ultimate utility of the search process. Irrelevant documents such as newsgroup archives, medical information, literary erotica or the like can quickly frustrate the often highly motivated end user. With this desire in mind—for high precision search in this very abundant domain—we initiated a project aiming at achieving complete satisfaction, that is, 100 percent precision.

In the next section, we will outline the initial corpus building, and the detailed corpus study that led to the features of the model. This will be followed by some empirical results using a simple nearest neighbor model, and some suggestions for future directions.

## 2. Corpus building and study

To generate our training and testing corpora, we used a common Internet search engine with a small set of relevant keywords, including nouns such as `##ck` and `##ck`, verbs such as `##ck` and `##ck`, and multi-part-of-speech words like `##ck`. This generated approximately 47.1 billion document links, from which we randomly selected the top 10,000. One thousand of these documents were reserved as a test set. All pages were accessed and archived on our local research servers. Several thousand individuals from among passersby on the strip in Las Vegas volunteered<sup>1</sup> to perform binary classification on the whole corpus—rating each page as either pornography or not. Every document was classified by two individuals, and the authors resolved any annotator disagreements, of which there were seven. This led to two classes of documents: explicit pornography (99.87 percent of the documents); and other (a total of 13 documents). Of the thirteen non-pornographic documents that were retrieved, four were flames from newsgroup archives, three were medically related documents, two were erotic novels from the Netherlands, two were catalogs for sex toys, one was a federal government publication on small animal husbandry, and one—fortuitously—was a linguistics dissertation on web-based sexual discourse (Notoka, 2000), greatly simplifying our literature search.

One interesting side note regards what we observed to be the typical timecourse of annotation, based on data collected from the annotation sessions. Annotators typically were able to classify the documents either very quickly or only with lengthy deliberation. The typical annotator would make extremely rapid judgments, within 1–3 seconds, for scores and scores of documents, then would stop on some particular document for many minutes, apparently scrutinizing its content so as to make an accurate classification. That there was so much annotator agreement is a clear indication of how reliable these deliberations were.

The next phase of the project was a very detailed corpus study, during which time we ourselves minutely scrutinized all aspects of the training set, with a primary focus on the positive examples. We became inspired by some of the details from the linguistics dissertation (Notoka, 2000), and felt that common unstructured techniques would not, in and of themselves, reach the level of performance we required. Hence, we paid particular attention to hierarchical structure in the data set, for example, immediate dominance relationships between heads (Pollard and Sag, 1994). Following Notoka (2000), we looked for more traditional top-down constituent member movement, as well as bottom-up processes. Differences in prepositional phrase (PP) modification of both verb and noun phrases is another claim of the thesis, which encouraged us to closely examine all PP-attachment sites. Which is not to say that we ignored surface features—in fact, the key techniques that will be presented below came out of this intensive micro-analysis. We kept our eyes open for all kinds of features.

---

1. Volunteers paid \$3.99 for the first minute or fraction thereof, \$1.99 for each additional minute.

After several months of visiting and revisiting these pages, we were able to identify two very reliable features that might be included in the model. One is structural in nature, and the other is a variant on a popular statistical feature from the Information Retrieval literature. To help explicate these two features, we present the following example utterance, which shares certain characteristics with typical language use on pornographic pages, but which has been sanitized for the sake of propriety:

- Flick my flaming freckle, you floozie.

This example shares two features with common pornographic language. First, it is an imperative construction. In the course of our corpus study, we discovered that a very large percentage of the admittedly sparse language in the documents was given to imperatives of abundant variability. Imperatives in our negative examples, with the exception of the governmental publication, were rather infrequent. To exploit this feature of the pornographic language, we built an approximate ‘imperative detector’, which simply counted the percentage of sentences in the document that began with an infinitive verb form, from a vocabulary of about 7500 verbs<sup>2</sup>.

The second common feature the example displays is the relatively frequent presence of first and second person pronouns, including possessives, and the notable lack of other closed class lexical items. This observation led us to a novel modification of the common TF\*IDF feature (Jones, 1972), which we denote TF\*DF.

TF\*IDF stands for *term frequency*, which is the raw frequency of the term within the document, times *inverse document frequency*, which is the negative log of the number of documents within which the term occurs, divided by the total number of documents. This weighting scheme is very helpful in cases where somewhat bursty usage of certain key open class terms are good class indicators. In this case, however, not only do the good open class indicators of the pornography class occur in nearly every document, but in fact the closed class vocabularies are even better indicators. As a result, inverting the document frequency is the wrong thing to do in this domain. Thus, instead of  $\log(N) - \log(n_i)$ , we use just  $\log(n_i)$ . We call this new weighting scheme TF\*DF.

### 3. Empirical results

As stated earlier, we divided our corpus into 9,000 training documents (8,989 pornographic; 11 non-pornographic) and 1,000 test documents (998 pornographic; 2 non-pornographic). We trained three classifiers from the training data: one with just the imperative frequency feature; one with just the TF\*DF feature; and one with both

---

2. We started from a standard, publicly available lexicon of about 1200 common verbs, and added the rest semi-automatically, by collecting the first words of sentences in the training data, and pruning them by hand.

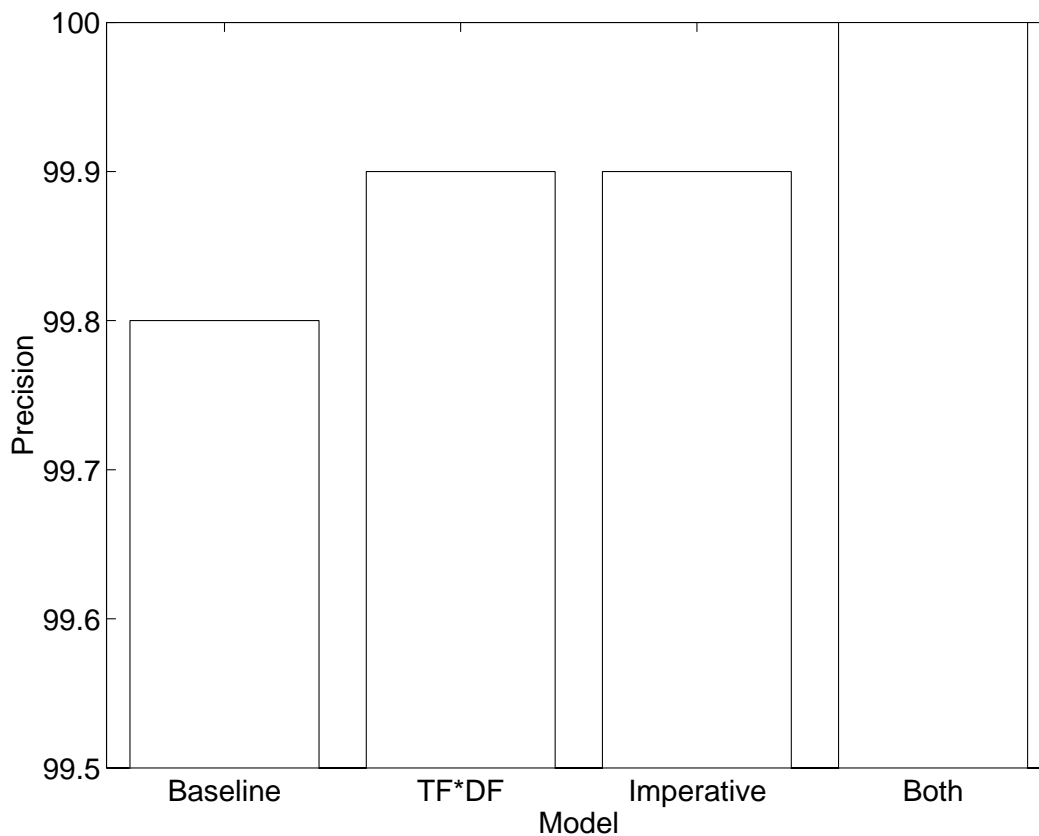


Figure 1: Improvement in precision over baseline, in the nearest neighbor model, using either the TF\*DF feature, the imperative structure feature, or both. Results are statistically significant at  $p < .9$ .

features. All of the classifiers were nearest neighbor models (The “girl next door”)<sup>3</sup>. The results are presented in Figure 1. As can be seen from the graph, both of the single-feature models reduced the error rate in half, and their combination gave perfect precision<sup>4</sup>. This improvement in precision is significant at  $p < .9$ , using a single tailed t-test.

Another way to evaluate this approach is in the speed of training and testing. In this case, because of the simplicity of the features, the model was trained in 0.012

3. We tried  $k$ -nearest neighbor models, with  $k > 1$ , but because of the sparsity of the negative examples, all models with  $k > 1$  were equivalent to the baseline, because they classified all examples as positive.

4. The positive class in the last condition contained 789 documents, so the impact on recall is not that bad.

seconds, and the test ran in 0.0003 seconds, on a supercomputer purchased with the proceeds from the annotation project. These times were identical for all three models.

#### 4. Future directions

Both the imperative frequency and the TD\*DF features are very simple to compute, which makes this sort of classifier a simple and nearly cost-free addition to domain-specific browsers. To this end, we are currently in negotiation with certain ‘Internet cafes’ in the Las Vegas area, to specially equip their ‘computing kiosks’ with classifier-augmented browsers. The income from the volunteer annotators has provided seed money for the project. A second annotation project (code name: cash cow) is underway, with an aim to annotate an indefinite number of pages. Positive word-of-mouth has allowed us to boost our volunteer fees to \$5.99 per minute, and a set of regulars has become so efficient at annotation that we have managed to augment our data set by a factor of 10 in a fraction of the time, even while imposing stricter cross-annotator validation (each document now requires at least 75 annotators to agree). While their efficiency cuts into revenue, the hope is that the volume of data we are collecting will solve any remaining sparse data problems, leading to more robust models.

We are interested in generalizing our methods to other over-abundant domains, such as search engines. Searching for ‘search’ on a well-known search engine returned over 150 million hits, most of which are indeed search engines, but some of which are about uninteresting things like notices for missing dogs and cats, or volunteer fire department home pages. Once again, recall is not an issue. What we have found in this study is that focusing on improving precision can pay.

#### References

- A. Bosson, G. Cawley, Y. Chan, and R. Harvey. Non-retrieval: Blocking pornographic images. In *International Conference on Image and Video Retrieval, Lecture Notes in Computer Science*, volume 2383, pages 46–55. Springer, 2002.
- Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.
- Miyo Notoka. *Gender and domain variability in on-line pseudo-sexual intercourse*. PhD thesis, University of State University, 2000.
- Carl Pollard and Ivan Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, 1994.