Recall Estimation for Rare Topic Retrieval from Large Corpuses *

Praveen Bommannavar
Twitter, Inc.
praveen@twitter.com

Alek Kolcz
Twitter, Inc.
ark@twitter.com

Anand Rajaraman
Stanford University
anand@anandr.com

ABSTRACT

The problem of finding documents pertaining to a particular topic finds application in a variety of scenarios. Indeed, the demand for topically pertinent documents has led to myriad companies offering services to find and deliver them (perhaps along with sentiment analysis or clustering) to customers for any topics of interest. The methodologies used to uncover relevant documents range from manually curated keyword filters to trained classification models. Any serious topical analysis requires a sound understanding of key metrics behind the retrieval process, two of the most important being precision and recall. While precision can be easily and inexpensively measured by sampling from classified documents and utilizing (paid) human computation to mark incorrectly classified instances, it is not as straightforward to use the same approach for measuring recall. With most topics occurring relatively sparsely, an unbiased sampling approach becomes prohibitively expensive. In this paper, we introduce a recall measurement procedure requiring only relatively few human judgements. The technique makes use of pairs of sufficiently independent classifiers and the paper provides a detailed discussion of how such classifier pairs can be constructed in practice, with a focus on social media classifiers. We report the performance of the proposed method with simple keyword filters as well as with classifiers of progressive levels of complexity and show that under reasonable conditions, recall can be estimated to within 0.10 absolute error and 15% relative error, and often closer with a reduction of cost by a factor of as much as 1000x as compared with unbiased sampling.

Categories and Subject Descriptors

D.28 [**Data Mining**]: Metrics—*performance metrics*

General Terms

Data Mining, Text Mining, Recall Estimation

*Work done partially at Twitter, Inc.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'14, August 24 – 27, 2014, New York, NY, USA
Copyright 2014 ACM xxx-x-xxxx-xxxx-x/xx/xx ...$15.00.

Keywords

recall estimation, classifier evaluation, social media, Twitter, mechanical turk, human evaluation

1. INTRODUCTION

Uncovering topically pertinent documents from a large corpus has long been at the heart of the field of information retreival. Indeed, from web search to library lookup, a variety of methodologies have been proposed and implemented for obtaining a desired subset of documents [1].

In particular, the widespread creation of social media updates (e.g., Tweets, status posts) and their volume/availability has resulted in their usage for a variety of predictive tasks: analyzing the sentiment around products and brands (and resulting potential sales), inferring the audience for sporting events, estimating popularity of politicians, etc. This has made classification of social media streams into topics a valuable activity; indeed, providing the analysis to support these tasks has become the core offering of several internet companies. The methodologies we propose in this paper are broadly applicable, but due to the timeliness of social media applications, we describe our approaches with particular attention to them.

Before employing any analysis on the documents that are relevant to a topic of interest, a necessary step is to first retrieve the documents which are in fact about the topic of interest. The approaches taken for doing this vary in sophistication, which can be as simple as manually curating keyword filters to separate documents based on any of the keywords matching, and as involved as training a classification model and determining thresholds for which highly scored documents are judged to be "on topic" [2].

Regardless of the level of sophistication in the document retrieval process, two important associated metrics are the process's *precision*, the fraction of retrieved documents which are in fact on topic, and it's *recall*, the fraction of all topical documents which have been retrieved.

A commonly used method for estimating these metrics is to perform unbiased sampling followed by human evaluation. While in the case of precision one would sample from the retrieved documents (take a few hundred retrieved documents and pay for human labeling of them), in the case of recall one needs to sample from the entire universe of documents under consideration. That is, while only a few hundred documents need to be sampled in the case of precision estimation for a given margin of error and a particular confidence, orders of magnitude more documents need to be sampled to get the same number of on topic docu-

ments via sampling if the topics are rare. Indeed, in one of our Tweet datasets, the topic `Obama` occurs with frequency 0.0015; to get a few hundred on-topic documents by uniform sampling, approximately 100,000 documents need to be collected. Once these documents are sampled, paid human evaluation can be applied to determine which of these documents are truly on-topic, which even at a rate of pennies for a document can become too costly for regular evaluation (thousands of dollars for a single topic). Hence, although uniform sampling is a principled and effective technique, in the case of recall estimation it can become unreasonably expensive for topics which appear infrequently. While stratified sampling type approaches can be used for measurement of precision, its application to recall estimation is unclear. A common workaround is to rely on ad-hoc methods such as resorting to surveys asking human judges to manually provide examples of documents that should have been found but were not, but such heuristics offer no guarantees on the accuracy of the resulting recall measurement.

The main technical contribution of this paper is to offer an inexpensive and practical yet principled methodology by utilizing pairs of classifiers which approximately satisfy a conditional independence property, which is made precise in Section 5. Using conditionally independent classifiers is used extensively in machine learning, e.g., in co-training, as introduced in [3], where each classifier in a conditionally independent pair can use the other classifier's output to guide its own training. One way in which such classifiers are constructed is by splitting document features into two set sets which are as uncorrelated as possible. While we do not apply co-training in this work, we analogously make use of this property for a different purpose, namely to estimate the recall of each classifier in an inexpensive way when class distribution varies over time and tends to be highly skewed.

Estimating metrics in the face of skewed distributions and changing conditions has been the subject of a considerable amount of prior work. In [4], methods for counting positives when the class distributions vary with time are considered, but the methods assume a large enough corpus of training data such that the behavior of on-topic documents can be analyzed well enough to directly measure true positive rate (TPR). Similarly, [5] establishes confidence intervals for recall estimates, but again supposes a frequent enough class distribution so that unbiased sampling can produce enough on-topic documents. In this work, we develop methods with the specific problem of infrequent on-topic document occurrence in mind.

The issue of skewed class distributions in model calibration is touched upon in [6], where an argument is made for using false positive rate (FPR) rather than precision or F1 measure as a way to choose thresholds for a learned model. As in this paper, the goal is to minimize the cost of expensive human labeling procedures. Rather than focusing on threshold selection, we attack the challenging issue of estimating recall of a given classifier.

Skewed class distribution as well as varying out-of-sample distributions are simultaneously considered in [7], but in the case of precision rather than recall estimation. A stratification procedure is developed and extended by which classifier precision "in the wild" can be estimated using labeled data from a different distribution. Biased sampling according to classifier output allows for correction of the distribution of documents in the test set. We review and extend this idea in our work.

An overview of evaluation for classification systems is given in Section 3, and the unbiased sampling approach, as well as the associated drawbacks in the case of recall, are detailed in Section 4. In Section 5, we derive a method for overcoming the main difficulty (costliness) of estimating the total volume of on-topic documents by utilizing pairs of classifiers that fire independently given that they are classifying on-topic documents. We continue in Section 6 by describing a number of classifier pairs that satisfy the requisite independence conditions, and zoom in on social media classifiers in particular. Section 7 presents the results of various experiments for evaluation of the proposed methods using three distinct data sources. Finally, Section 8 offers concluding remarks as well as directions for future work.

2. KEY CONTRIBUTIONS

In this paper we deliver the following:

1. An inexpensive method for measuring the recall of topic classifiers.
2. Theoretical justification as well as intuition for the proposed method.
3. Guidelines on how to construct pairs of conditionally independent classifiers.
4. Evaluation using Tweets and Open Directory Project (ODP) records [1].

3. CLASSIFIER METRICS

Let us formalize the problem of finding a subset of documents from a corpus according to some search criteria. The goal could be to find topical documents that would be of interest to a user, to identify not suitable for work (NSFW) content to avoid showing to users, or to generate a relevant set for yet another purpose. A *classifier* is a decision rule which, given a document, assigns it a label from the set {+1,-1}. A +1 label is given to "positive" decisions where the classification decision indicates a belief that the document in question satisfies the search criteria, while a -1 label is given to "negative" decisions. Classifiers make decisions using so-called *features* of the input document, such as the text in the document, authorship information or connections with other documents. They can be as simple and interpretable as rule based decisions ("all documents with the word 'basketball' is about the topic `sports`") or as opaque as neural networks with many layers and millions of input features.

There exist numerous methods for evaluating the quality of a classifier, such as precision, recall, false positive rate (FPR), area under the curve (AUC), discounted cumulative gain (DCG) and many others [8]. A long history of rigorous evaluation of information extraction systems provides the foundation for these choices of metrics [9] [10]. In the case of thresholded classifiers, a decision boundary is often chosen on the basis of optimizing some combination of these metrics, such as F-score (defined as the harmonic mean of precision and recall) [11].

Two of the most fundamental and widely used quantities of interest are the classification rule's *precision* and *recall*.

[1] http://www.dmoz.org

Precision is defined as the fraction of retrieved documents which are in fact on topic, or:

$$p := \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|},$$

whereas recall is the fraction of on-topic documents that are retrieved:

$$r := \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}.$$

In order to maintain an ongoing understanding of performance, any service utilizing a classification rule would be interested in tracking these metrics with some regularity. And as we shall see in the next section, a key part of understanding these quantities is the acquisition of human judgements on the topicality of documents.

4. SAMPLING AND HUMAN EVALUATION

In this section, we describe a simple sampling method for estimating the precision and recall of a topic classifier. A necessary component of this method is human evaluation for identifying on-topic documents [12]; the accuracy of the reported precision and recall depends on the amount of labeled data, as we describe in greater detail in the following sub-sections.

4.1 Mechanical Turk

Confirmation labels, or binary judgements on whether or not a document is on topic, can be obtained by using a crowdsourcing platform like Amazon's Mechanical Turk [2]. Current rates are generally inexpensive, between $0.01 and $0.10 per tweet. Longer documents can be more expensive.

A number of issues surround the reliability one can expect with regard to judgements received from Mechanical Turk, and sophisticated estimation procedures have been developed to determine a strategy for learning workers' reputations based on past performance [13].

In this work we take the simplest scheme possible and suppose that each labeled data point has been seen by three workers. Data collected in this way is the starting point from which all metrics in this paper are computed. The cost which we would like to minimize is the amount which is paid to workers through labeling tasks, or equivalently, the amount of data which needs to be judged for topicality.

4.2 Notation

Denote the possibly infinite set of all documents in the collection under consideration by U, and let C_1, C_2 refer to two classifiers which map a document $x \in U$ to $\{0, 1\}$ depending on whether or not it is classified as being on topic. Further let C_{12} be the joint classifier defined by taking documents which are positively classified by both C_1 and C_2.

In this paper, we are not only interested in classifiers for which a score is computed and on-topic documents are found by thresholding of this score, but we are also interested in keyword filters, whereby a document is marked as on topic if any one of a number of keywords appears in it. While keyword filters can be seen as a special case of the first kind of classifier, we will see that some methods apply more readily to one or the other type of retrieval mechanism. For notational purposes, we denote both of these classifier types the same way and explicitly state situations in which we wish to differentiate between them.

The sets of documents classified by each of C_1, C_2 and C_{12} as on topic are denoted A_1, A_2, and A_{12}, respectively. That is, $A_i = \{x \in U : C_i(x) = 1\}$.

For a given topic, we refer to the set of documents pertaining to that topic as T, and for any subset $S \subseteq U$, the probability that a uniformly sampled document is in S is written $P[S]$. Further we let the conditional probability $P[S_1|S_2]$ denote the probability that a randomly sampled document is in $S_1 \subseteq U$, conditional on being in $S_2 \subseteq U$.

Finally, the associated precisions of classifiers C_1, C_2, C_{12} are p_1, p_2, p_{12}, and the recalls of classifiers C_1 and C_2, which we are interested in estimating, are denoted r_1 and r_2.

This notation is summarized in Table 1.

4.3 Unbiased Sampling

We begin by demonstrating an unbiased sampling approach for estimating precision and recall. It is reasonable to expect a 95% confidence interval with a margin of error at 0.10 in either direction for the measurement of either of these metrics. In the case of precision, the unbiased samples that are needed are the documents which are positively labeled by the classifier in question, whereas recall requires an unbiased sample of true positives. For each task, we can calculate the number of samples needed to be approximately 384, using the well known formula:

$$(Margin\ of\ Error) = 1.96\frac{s}{\sqrt{n}}$$

where s is the sample standard deviation and n is the number of samples [14]. Since the maximum value that s^2 can take is 0.25 (due to a proportion being a Bernoulli random variable), and using a margin of error of 0.05, we calculate the number of samples necessary to be 384. For a margin of error of 0.025, we require 1537 samples.

The thesis of this paper is that while obtaining samples for a measurement of precision can be done easily, obtaining a measurement of recall is intractably expensive via direct sampling. This is because obtaining positive examples for sparse topics requires human evaluation of many examples to determine if they are indeed on-topic.

4.3.1 Estimating Precision via Uniform Sampling

A simple procedure for estimating the precision of any topic classifier C_1 is:

1. Collect 384 samples from A_1, where A_1 is constructed by applying classifier C_1 to the document universe U.
2. Pay workers on Mechanical Turk to evaluate these samples. At approximately $0.05 per Tweet, this is about $19.
3. Report the fraction of documents which were correctly classified as the precision.

Hence, at a reasonable cost we are able to get a good estimate for the precision of the classifier using standard statistical sampling techniques.

4.3.2 Estimating Recall via Uniform Sampling

Although superficially similar to the procedure for estimating precision, in the case of recall one needs to first find positive examples of on-topic documents, resulting in a new human labeling step:

[2]http://mturk.amazon.com

Table 1: Notation used in this paper.

U	The set of all documents in a collection
T	The set of all on-topic documents ($T \subset U$)
$P[S]$	Probability that a uniformly sampled document from U belongs to the set $S \subset U$
$P[S_1\|S_2]$	Probability that a uniformly sampled document from $S_2 \subset U$ belongs to the set $S_1 \subset U$
C_1, C_2	Two classifiers which partition documents in U into on- and off-topic subsets
C_{12}	The joint classifier defined by $C_1 \wedge C_2$
A_1, A_2	The set of documents found by C_1 and C_2, respectively
A_{12}	The set of documents found by *both* C_1 and C_2
p_1, p_2	The precisions of classifiers C_1 and C_2, respectively
p_{12}	The precision of classifier C_{12}
r_1, r_2	The recalls of classifiers C_1 and C_2, respectively
δ_1, δ_2	The degree to which Assumptions 1 and 2 hold (on- and off-topic conditional independence)

1. Collect a large number of samples from U.
2. Pay workers on Mechanical Turk to evaluate these samples until 384 on-topic documents are found.
3. Report the fraction of these on-topic documents that are also in A_1.

Whereas in the case of estimating precision we could count on a relatively small number of human labeling tasks, estimation of recall via this straightforward sampling procedure requires first constructing a set of on-topic documents. For the same margin of error and confidence level as in the previous case of precision estimation, one would need to sample from U until 384 on-topic documents were found, paying for the labeling of each one, whether on-topic or not. If the topic occurs frequently, say $\frac{|T|}{|U|} = 0.1$, then approximately 3840 documents would need to be labeled, at a cost of \$190. For topics with lower frequency of occurrence, say $\frac{|T|}{|U|} \leq 0.01$, this cost becomes a prohibitive \$1900. Figure 1 shows how this cost grows as the rate of occurrence for the topic of interest decreases.

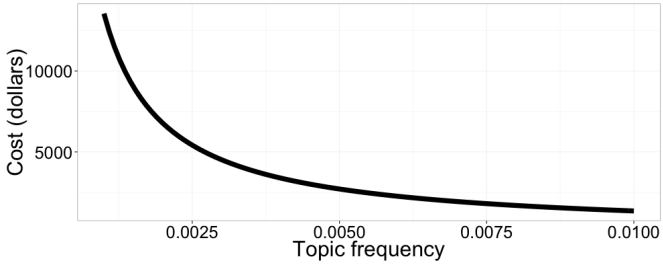

Figure 1: Cost of recall estimation via sampling as a function of prevalence rate.

5. RECALL ESTIMATION

As noted in the previous section, while unbiased sampling is a simple and inexpensive method for estimating the precision of a topic classifier, it becomes infeasible to use this method for estimating classifier recall of rare topics. Even if one is willing to pay a one-time cost to apply sampling, obtaining up-to-date information with any regularity becomes impossible. This paper proposes a practical method for obtaining an estimate of reasonable margins for the classifier recall of a sparse topic that only relies on the precision estimates, which we have seen to be inexpensive to estimate. The key requirement of the classifiers is that they are "sufficiently" independent of each other, which translates to a set of conditions that we make precise in the following subsection.

5.1 Conditional Independence

Let us now list the assumptions which drive the results in the remainder of this section.

ASSUMPTION 1. *(Conditional Independence 1) For the set of on-topic documents T, C_1 and C_2 are independent classifiers. That is,*

$$\delta_1 := \frac{P[A_1 \cap A_2|T]}{P[A_1|T]P[A_2|T]} \approx 1.$$

As we shall see in Section 7, this is a reasonable assumption in a variety of situations. For example, if C_1 and C_2 are using different features for classifying the documents, it is likely for this property to hold. An analogous assumption on the set of off-topic documents can be similarly stated as:

ASSUMPTION 2. *(Conditional Independence 2) For the set of off-topic documents T^c, C_1 and C_2 are independent classifiers. That is,*

$$\delta_2 := \frac{P[A_1 \cap A_2|T^c]}{P[A_1|T^c]P[A_2|T^c]} \approx 1.$$

Finally, we utilize the fact that the topics we are interested in tracking occur rarely. This is indeed true with many common topics of interest, especially niche topics. If this is not the case, unbiased sampling can be a feasible option.

ASSUMPTION 3. *(Sparsity) The number of on-topic documents T is small, as compared with the total universe of documents U. That is, $P[T] << 1$*

With these assumptions in hand, we can derive a set of relations which admit an inexpensive method for estimating the recall of C_1 or C_2, and hence the total number of on-topic documents.

Although the assumptions as stated may seem strong ("how can one obtain such classifier pairs?"), we provide a number of examples in Section 6 for which these conditions hold in an approximate sense.

5.2 Projecting Classifiers

In this section we describe a method for estimating the recall of a classifier C_1 by pairing it with another classifier C_2 such that the pair C_1, C_2 abides by the properties above. We first give some intuition and then rigorously derive the key results of this paper.

5.2.1 Intuition

We can intuitively see from the Venn diagram in Figure 2 that if A_1 and A_2 do not behave "badly" (in a sense that we shall make precise), we can think of the recall of C_1 as the ratio

$$r_1 = \frac{|T \cap A_1 \cap A_2|}{|T \cap A_2|}$$

That is, we can measure how much of A_2 the classifier C_1 is able to find. If C_1 and C_2 are sufficiently independent, this estimate reliably conveys the recall r_1.

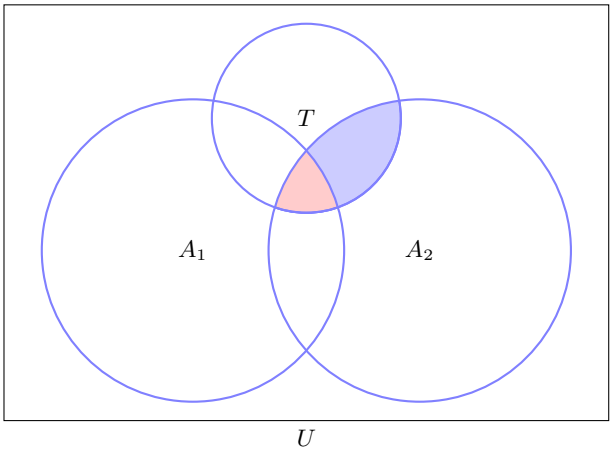

Figure 2: Document space and Venn diagram of classification output.

Rather than independence over random draws from the entire set of documents, it is in fact independence over the on-topic and off-topic subsets of U which makes this intuition valid. In the following section, we utilize our assumptions from the previous section to show that these are enough to guarantee that forming an estimate based on this ratio is justified.

5.2.2 Proof of Correctness

Let us begin by noting that the recall r_1 can be written using our notation as

$$r_1 = P[A_1|T] = \frac{P[A_1|T]P[A_2|T]}{P[A_2|T]} \approx \frac{P[A_1 \cap A_2|T]}{P[A_2|T]}$$

by Assumption 1. Continuing, we can write

$$\begin{aligned} r_1 \approx \frac{P[A_1 \cap A_2|T]P[T]}{P[A_2|T]P[T]} &= \frac{P[A_1 \cap A_2 \cap T]}{P[A_2 \cap T]} \\ &= \frac{P[T|A_1 \cap A_2]P[A_1 \cap A_2]}{P[T|A_2]P[A_2]} \\ &= \frac{p_{12}|A_{12}|}{p_2|A_2|} \end{aligned}$$

Therefore, if the precisions p_{12} and p_2 are known, the recall can be written as

$$r_1 \approx \frac{p_{12}|A_{12}|}{p_2|A_2|} \tag{1}$$

Note that for this to be true, we have used only Assumption 1. In particular, this result holds regardless of the topic sparsity.

If the joint precision p_{12} cannot be measured, for example because the evaluation data are too sparse, we can make use of Assumptions 2 and 3 by proceeding as follows:

$$r_1 \approx \frac{P[T|A_1 \cap A_2]|A_{12}|}{p_2|A_2|} = \frac{(1 - P[T^c|A_1 \cap A_2])|A_{12}|}{p_2|A_2|}$$

Let us now rewrite $P[T^c|A_1 \cap A_2]$ in a different form. By Bayes' Rule and Assumption 2,

$$\begin{aligned} P[T^c|A_1 \cap A_2] &= \frac{P[A_1 \cap A_2|T^c]P[T^c]}{P[A_1 \cap A_2]} \\ &\approx \frac{P[A_1|T^c]P[A_2|T^c]P[T^c]}{P[A_1 \cap A_2]}. \end{aligned}$$

Once again applying Bayes' Rule, we can write the right hand side as

$$\frac{P[A_1|T^c]P[A_2|T^c]P[T^c]}{P[A_1 \cap A_2]} = \frac{\frac{P[T^c|A_1]P[A_1]}{P[T^c]} \frac{P[T^c|A_2]P[A_2]}{P[T^c]} P[T^c]}{P[A_1 \cap A_2]}$$

and after applying Assumption 3 we can suppose $P[T^c] \approx 1$ so that

$$P[T^c|A_1 \cap A_2] \approx \frac{(1 - p_1)(1 - p_2)|A_1||A_2|}{|U||A_{12}|}.$$

We finally note that $P[T^c|A_i] = (1 - p_i)$ to get the following approximation for r_1:

$$r_1 \approx \frac{|A_{12}|}{p_2|A_2|}\left[1 - \frac{(1 - p_1)(1 - p_2)|A_1||A_2|}{|U||A_{12}|}\right]. \tag{2}$$

As in (1), the recall is expressed purely in terms of precisions, but in this case without knowledge of p_{12}. Unlike (1), however, further assumptions were required. While the conditional independence assumptions that were used in this section appear to be strong, in the following section we give a number of ways in which one might construct such pairs of classifiers.

6. CONSTRUCTING CONDITIONALLY INDEPENDENT CLASSIFIER PAIRS

Until this point, we have conducted all analysis under the assumption that pairs of classifiers which find on-topic documents are readily available. Even further, we have been relying on the conditional independence assumptions of Section 5.1 to obtain cost-efficient ways of estimating recall. We now discuss a few ways to generate pairs of topic classifiers that satisfy these assumptions.

For an arbitrary document collection, one can exploit the fact that documents tend to be redundant - the same information is expressed in many different ways throughout the document. There are several ways to leverage this redundancy to create conditionally independent classifiers. One possibility is to have the classifiers use non-overlapping portions of the document. It is easy to imagine that different sections of an article would all contain enough of a signal

to determine topicality. Another possibility is exploit document structure, as is common in information retrieval systems: anchor text, headers and linked URL's all provide different signals that can be leveraged into conditionally independent classifiers. Co-training frequently makes use of these types of splits in the data, and collections that exhibit natural divides in their structure are seen to be better suited for co-training [15].

One might initially call into question the utility of the described methods if only certain classifier pairs are able to make use of them. We note, however, that whatever classifier we have in hand, two new classifiers can be constructed as per the suggestions given here. Having the recalls of each of those classifiers, an estimate on the total number of positives, $\hat{n}$, can be formed. That is, we can produce the estimate $\hat{n}$ by computing

$$\hat{n} = \frac{p_1|A_1|}{\hat{r}_1}$$

where r_1 is the recall of one of our conditionally independent classifiers, p_1 is the corresponding precision and A_1 is the set of documents retrieved by this classifier.

With this estimate in hand, an estimate of the recall for the classifier of interest can be obtained, since the precision can be inexpensively computed as per Section 4. More concretely,

$$\hat{r}_{new} = \frac{p_{new}|A_{new}|}{\hat{n}}.$$

Although these approaches are effective guidelines for creating classifiers for general document collections, we utilize the remainder of this section to discuss classifier pairs that are particularly convenient to construct in the context of social media streams. It is worth noting, however, that the proposed strategies can also find application for other collections of documents, a point we demonstrate using ODP data (Figure 3).

www.criticalsoftware.com - Develops and markets software products for business and mission critical information systems, and provide consulting and engineering services for enterprises.

Figure 3: ODP entry for topic Software Engineering: link and description.

Inspecting the form of this record, we see that it greatly mirrors that of stories on Twitter: description text is short and conveys a summary, much like tweet text, and linked web documents are exactly the same as in stories.

6.1 Social Media

Due to the special form of the data generated in social media applications, constructing conditionally independent classifier pairs can be accomplished by leveraging this structure.

For the case of Twitter data, Tweets contain at most 140 characters of text. Sometimes a tweet contains a link to a web document; in this work, we make use of three main features of such Tweets: (status, link, author). Figure 4 gives an example of a tweet that also contains a link to a web document.

In this paper we perform experiments on three corpora: (1) 800K Tweets from August 6, 2012 (2) 10.5M stories from

Figure 4: Status update by author @jack with link.

March 1-7, 2013 and (3) ODP data. The ODP data have the form (description - link), which is analogous to the (status - link) pair for stories (Figure 3).

The remainder of this section discusses how to leverage the structure of social media updates to obtain pairs of classifiers that roughly satisfy Assumptions 1 and 2.

6.1.1 Seed terms and high similarity neighbors

We first consider a pair of status keyword filters that tend to exhibit the conditional independence properties of interest, reminding the reader that a keyword filter is a special type of classifier for which a document is marked as on-topic if *any* of the words in the filter appear in the document.

Before describing the two classifiers we wish to use, let us define the notion of a *seed* term to be a single word or phrase that captures the topic of interest in the broadest generality possible. Some examples of seed terms would be *sports*, *mars* or *obama*.

The classifiers are defined as follows: the first classifier in the pair, C_{seed}, filters on the terms

$$\left[\ seed, \quad \#seed\ (in\ the\ case\ of\ Tweets)\ \right],$$

while the second filter, $C_{neighbors}$, filters on "similar neighbors" of the seed terms. A number of different scores can be used for measuring overlap. Letting A_{seed} and A_y denote the documents found by the seed term filter and a candidate keyword y, respectively, one such similarity score is

$$S_{jaccard}(A_y, A_{seed}) = \frac{|A_y \cap A_{seed}|}{|A_y \cup A_{seed}|},$$

known commonly as the Jaccard similarity. Another possibility is the asymmetric overlap measure we define as

$$S_{overlap}(A_y, A_{seed}) = \frac{|A_y \cap A_{seed}|}{|A_y|}.$$

PMI, Tversky similarity or many other measures can be chosen, but regardless of which is used, the top scoring neighbors (leaving out seed terms) are included in the filter defining $C_{neighbors}$. Some examples of seed terms and corresponding partial sets of high similarity neighbors are given in Table 2. We can see that although some terms appearing in the list of neighbors are not on-topic, both filters have overall precisions that are 'reasonable'. Indeed, our results from Section 5 require only that the precision in the denominators of Equations 1 and 2 are not so small as to make the computation unstable (in addition to the conditional independence assumptions).

Intuitively one can see that the seed terms, for well behaving topics, might occur independently of other terms, given an on-topic document. For instance, for topic `olympics`, one would expect that the occurrences of the terms *100m*, *swimming* and *javelin* occur independently of the seed term *olympics*, given that a document is drawn from the set of on-topic documents.

Table 2: Examples of seed terms and high Jaccard similarity neighbors (Twitter statuses).

$[apple, \#apple]$	[#ios6,#ipad3,#iphone,hack,macintosh, iphones, #siri, ios, macbook, icloud, ipad, samsung, #ipodtouch, 4s, itunes, cydia, cider, #gadget, #tech, #tablet, app, connector, #mac, ...]
$[mars, \#mars]$	[rover, nasa, #curiosity, #curiousity, image, mission, surface, #curiosityrover, bruno, milky, budget, @marscuriosity, gale, crater, orbiting, successfully, lands, landing, breathtaking ...]
$[obama, \#obama]$	[@barackobama, barack, bush, #mitt2012, #obama2012, obamas, #dems, #gop, #military, romney, #idontsupportobama, potus, #president, administration, pres, #politics, voting ...]
$[olympics, \#olympics]$	[medalist, gold, london, gb, kirani, gymnast, kate, sprinter, winning, won, #boxing, soccer, watch, watching, #usa, #teamgb, #canada, #london, javelin, nbc, match, 2012, 400m ...]

6.1.2 Status and Hyperlinked Article

Next we focus threshold-based classifiers and present a social media classifier pair satisfying the conditional independence property. Here we are primarily concerned with classifiers induced over a training corpus that produce a real-valued score when evaluated over an out-of-sample datum. For scores above a set threshold, a document is labeled as on-topic. Many commonly used machine learning techniques fall into this category, including logistic regression, SVM or Naive Bayes.

Additionally, we redefine the universe of documents U to be the set of documents which also contain links to other documents. In the case of Twitter, we constrain ourselves to Tweets which also contain links to documents on the web (see Figure 4). ODP listings also take on a similar structure, where there exists a description for each link and the linked document itself (Fig 3).

Under these circumstances, a conditionally independent pair can be constructed as follows: one classifier is trained and run entirely on tweet text (ODP description) while the other is based entirely on the web documents that the tweet (ODP listing) links to. We note that it is important for the processes generating each description/web-document pair to be independent. Indeed, if a tweet or ODP description is composed only of snippets of the linked article, one cannot expect the independence properties to hold. In practice, we only need the conditional independence to hold in an approximate sense.

It can be observed that as the thresholds for classifiers C_1 and C_2 are lowered, our conditional independence assumptions become increasingly valid. The tradeoff for choosing thresholds at a lower level, however, is that obtaining accurate estimates of the respective precisions p_1 and p_2 becomes more expensive. Furthermore, Equations 1 and 2 suggest that our estimate of recall r_1 becomes increasingly unstable as p_2 becomes smaller.

6.1.3 Status and Social Signals

Finally, returning to our original universe of documents U, we consider a classifier pair which allows the first classifier, C_1, to be either filter based or threshold based. The second classifier, C_2, is based entirely on social signals. These signals would primarily be composed of authorship (an author who is primarily known for technology articles would have all of his Tweets classified by C_2 as on-topic for the topic `technology`). On the Twitter platform, @mentions can provide additional social signals. For instance, knowing that a tweet has been sent as a reply to @barackobama can provide a signal that the tweet is about `politics`.

Due to the fact that the signals used by each of these classifiers have independent origins, we would hope that conditional independence holds here as well. It should be noted, however, that there are many ways in which this strategy can fail; for instance, a classifier C_2 for the topic `politics` filtering only on @accounts on one end of the political spectrum (e.g., just liberal views) might not be conditionally independent of C_1.

7. EVALUATION / NUMERICAL RESULTS

To demonstrate the effectiveness of constructing conditionally independent classifier pairs for recall estimation, we provide empirical results using three different datasets. For the first one, we construct conditionally independent keyword filters for four topics over a day's worth of Tweets. With the second dataset we train classifiers for four different topics over a day's worth of Twitter stories. Finally, with the third dataset we report results for classifiers trained over top-level topics of the ODP dataset. In all of the following examples, in order to measure precision of a classifier, we sample approximately 1000 data points, even when the full ground truth is known.

7.1 Sampled Tweets

Our first data set is comprised of approximately 800,000 English language Tweets sampled from August 6, 2012, where we have normalized the collection so that all words are in lowercase and non-ascii characters have been removed. For each tweet, we have utilized mTurk to obtain a judgement on whether it pertains to any of four topics: `apple` (the technology company), `mars` (the planet), `obama` or `olympics`. This exhaustive labeling, while expensive, allows us to perform unbiased evaluation of our techniques.

The chosen topics are interesting for a number of reasons. On August 6, 2012, two of these topics (`olympics` and `mars`) were prominent news topics, and hence frequently appeared in Twitter statuses. Additionally, `mars` and `apple` as topics are ambiguous without further specification; `apple` can also refer to a fruit, and `mars` has other interpretations as well. Hence, mTurk workers were told to specifically focus on the interpretations that we wished to study. As we shall see, the classifier pairs we construct work regardless of this complication.

Using the methodology of Section 6.1.1, our first experiment consists of constructing seed-based and high similarity keyword-based classifiers via Jaccard similarities, which produces keyword lists (partially listed in Table 2). We suppose that the precisions given can be easily estimated using the standard techniques overviewed in Section 4 and focus instead on estimation of recall.

Table 3 summarizes our numerical results for the 800,000

Table 3: Experimental results: tweet keyword filters. Both recall estimation schemes are within 0.10 absolute error and 15% relative error of the true recall for all topics.

Topic	$\|A\|$	$\|A_{seed}\|$	$\|A_{kw}\|$	$\|A_{joint}\|$	$\hat{p}_{seed}$	$\hat{p}_{kw}$	$\hat{p}_{joint}$	$\hat{r}^{(1)}_{seed}$	$\hat{r}^{(2)}_{seed}$	r_{seed}	$\hat{r}^{(1)}_{kw}$	$\hat{r}^{(2)}_{kw}$	r_{kw}
`Apple`	3038	676	10217	420	0.655	0.247	0.774	0.129	0.166	**0.146**	0.734	0.943	**0.830**
`Mars`	2372	1783	7703	1433	0.904	0.264	0.938	0.661	0.704	**0.680**	0.834	0.889	**0.857**
`Obama`	1253	851	7400	513	0.984	0.116	0.994	0.596	0.599	**0.668**	0.609	0.613	**0.683**
`Olympics`	23126	4595	45705	2688	0.986	0.330	0.989	0.176	0.178	**0.196**	0.587	0.593	**0.653**

Table 4: Experimental results: story text and webpage logistic regression classifiers. Both recall estimation schemes are within 0.10 absolute error of the true recall for all topics and most topics are within 15% relative error.

Topic	$\|A_{tw}\|$	$\|A_{web}\|$	$\|A_{joint}\|$	$\hat{p}_{tw}$	$\hat{p}_{web}$	$\hat{p}_{joint}$	$\hat{r}^{(1)}_{tw}$	$\hat{r}^{(2)}_{tw}$	r_{tw}	$\hat{r}^{(1)}_{web}$	$\hat{r}^{(2)}_{web}$	r_{web}
`Ads/Marketing`	42073	76771	4369	0.825	0.698	0.900	0.073	0.077	**0.075**	0.113	0.120	**0.145**
`Education`	93292	76535	21426	0.827	0.868	0.873	0.282	0.319	**0.206**	0.242	0.275	**0.214**
`Real Estate`	42841	31978	12411	0.836	0.918	0.989	0.418	0.420	**0.413**	0.343	0.346	**0.380**
`Food`	42376	218507	20493	0.875	0.842	0.898	0.100	0.110	**0.122**	0.496	0.546	**0.522**

sampled Tweets. In our notation, $\hat{r}_i^{(j)}$ is the recall estimate for classifier i made by using Equation j, and r_i is the true recall as measured by using the mTurk provided labels (and analogously for true precisions). We can see that estimates from Equation 2 (unknown joint precision) are about as reliable as estimates from Equation 1 (joint precision known). In both cases we can see that for all topics, recall is measured to within a margin of approximately 0.10 absolute error, and many times to within 0.05, while the relative error is 15% or lower. In the following examples, we show that the conditional independence required holds not only for keyword matching rules, but also trained regression models.

7.2 Sampled Twitter Stories

In our second illustration, we utilize a collection of 10.5M Twitter stories (Tweets containing hyperlinked articles) from March 7, 2013 and focus on the high level topics `ads and marketing`, `education`, `real estate` and `food`. Due to the fact that these topics are somewhat more general than those of the previous section and that stories tend to contain less 'chatter' than Tweets without hyperlinked documents, sampling approaches can be used to verify the methodologies proposed in this paper. Even though the prevalence of these topics is greater than the niche topics studied in the previous numerical illustration, their overall frequency is still low enough that Assumption 3 remains true in an approximate sense.

Rather than keyword filters, here we have constructed classifiers for each of our topics using a 1 vs all logistic regression model for each topic. Story text is classified using hashed character 4-grams as features, while linked webpages use hashed unigrams as features. Thresholds for each of the classifiers have been selected in such a way as to obtain classifier precisions of 0.70 or greater. Further, we can see that even those these are 'rare' topics when considering the size of the entire Twitter stream, each of these topics have a significant absolute volume, allowing for enough training and test data for statistical effects to take hold.

As in the previous section, we observe in Table 4 that the recall estimates are consistently within 0.10 absolute error of the true recall (obtained via a sampling procedure) and most of the time within 15% relative error. The estimates produced by Equation 2, where the joint precision is estimated indirectly, are seen to have higher estimates than those produced by Equation 1 where the joint precision is estimated directly. This is in part due to the fact that the assumption of topic rarity (Assumption 3) is only approximately true here.

7.3 ODP Entries

In our final numerical example, we utilize a dataset comprised of ODP entries for 12 top level categories. Other top level categories, such as `News`, `Regional` and `World`, from the original data set were removed since they are actually loose aggregation of other topics which appear in the top level. Specifically, each of these categories contains its own entries for topics such as `Arts`, `Business` and `Shopping`, making classification an ambiguous task.

As shown in Figure 3, the structure of these entries mirrors that of stories on Twitter: there is a short description of the document and a linked web document with a larger body of text. A training and test set are constructed using a random 70%-30% split; each host does not appear more than once in sample documents, so as to eliminate overfitting to host names which may appear in the document text [16]. The classifiers trained on description text and linked web text are multinomial logistic regression models with up-sampling to balance class distributions in the training set. Hashed character 4-grams were used as features for the classifiers, and the estimated precisions of each classifier used 1000 samples from the total number of documents which were retrieved.

Let us turn to the problem of estimating the recall of each classifier using precision and coverage measurements of the other. Since there are 12 topics in this collection and each record corresponds to some topic, Assumption 3 is violated by these circumstances for several topics, and hence we focus on estimating recall using Equation 1. Indeed, while on Twitter there are many available stories related to topics which comprise a small fraction of the overall collection, in order to be able to train classifiers and run these experiments, we must here consider categories with enough true positives to allow the statistics to play out. Nonetheless, this open data set validates our intuition that $\delta_1 \approx 1$ is a reasonable assumption when training classifiers on text generated in different fields of a document.

Table 5: Experimental results: prevalence and recall estimation in ODP records. Using joint precision directly gives high fidelity recall estimates for most topics, but attempting to approximate it results in poor recall estimates.

Topic	$\|A\|$	$\|A_{desc}\|$	$\|A_{web}\|$	$\|A_{joint}\|$	$\hat{p}_{desc}$	$\hat{p}_{web}$	$\hat{p}_{joint}$	$\hat{r}^{(1)}_{desc}$	$\hat{r}^{(2)}_{desc}$	r_{desc}	$\hat{r}^{(1)}_{web}$	$\hat{r}^{(2)}_{web}$	r_{web}
Adult	1470	10080	2815	757	0.089	0.430	0.710	0.624	0.878	**0.593**	0.839	1.180	**0.814**
Arts	13811	13719	12469	5523	0.524	0.766	0.865	0.581	0.671	**0.510**	0.771	0.891	**0.692**
Business	32304	18766	23888	10849	0.748	0.870	0.938	0.520	0.554	**0.443**	0.770	0.820	**0.646**
Computers	11235	11802	11756	4111	0.431	0.706	0.804	0.498	0.620	**0.453**	0.814	1.012	**0.746**
Games	2146	4245	3723	764	0.223	0.441	0.754	0.465	0.616	**0.439**	0.807	1.070	**0.766**
Health	5986	6180	6881	2991	0.576	0.667	0.872	0.649	0.744	**0.602**	0.837	0.960	**0.766**
Home	1546	7565	3616	643	0.118	0.299	0.574	0.594	1.035	**0.543**	0.717	1.249	**0.705**
Recreation	10846	9022	10712	4488	0.626	0.713	0.899	0.586	0.652	**0.505**	0.792	0.881	**0.703**
Science	5540	6005	7710	1706	0.380	0.462	0.658	0.474	0.720	**0.417**	0.739	1.123	**0.640**
Shopping	12386	13534	14610	3865	0.335	0.642	0.773	0.419	0.542	**0.375**	0.868	1.122	**0.757**
Society	12925	8397	11289	4071	0.627	0.720	0.922	0.501	0.543	**0.408**	0.774	0.839	**0.622**
Sports	6049	6929	6775	3197	0.550	0.708	0.910	0.664	0.729	**0.638**	0.835	0.918	**0.778**

Table 5 shows that in most cases, our recall estimates using Equation 1 are within 0.10 absolute error of the true recall, and often closer, with the relative error at 15% or lower. On the other hand, recall estimates from Equation 2 are in many cases gross overestimates of the true recall. It is also possible that the assumption that $\delta_2 \approx 1$ does not hold for this dataset, which could also be a consequence of the amount of data available for each topic.

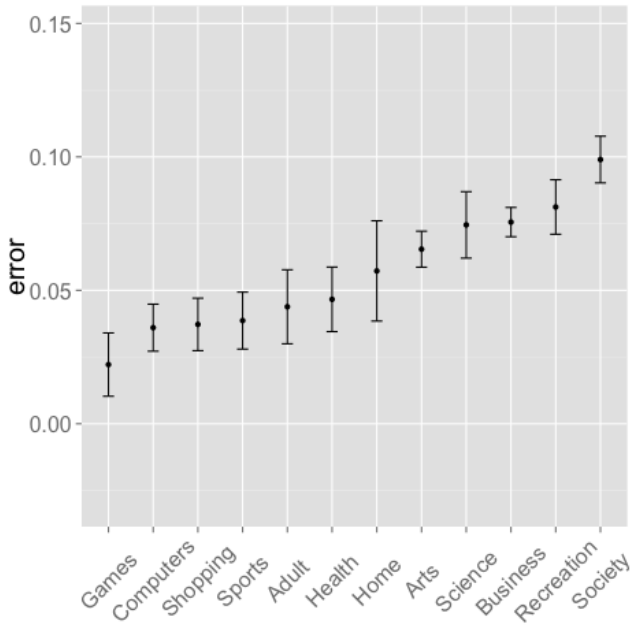

Figure 5: Distribution of recall estimation error for the description text classifier over 100 random training-test splits. In almost all cases the 95% confidence interval lies fully below 0.10.

We can further validate these results by considering 100 random splits of the data and rerunning this experiment over each split. Figure 5 shows the mean recall estimation errors of the description text classifier for each topic with their 95% confidence intervals.

The mean error for all topics is within 0.10, and the confidence intervals for almost all topics sit below 0.10. Further, we can see that the 95% intervals tend to be larger for those topics with fewer samples. Although not shown in the axis, these error intervals also correspond to relative errors within 10-15% of the true recall.

7.4 Cost Analysis

In all of the given examples, precisions of each classifier as well as the join classifier made use of approximately 500 samples, which at $0.05 per classification is $25. Depending on the frequency of the topics, recall estimation via sampling methods can be much more expensive. Indeed, for the topics occurring with frequency of 0.01 or less, this cost increasing by a factor of more than 100 times. In order to run such analysis with any regularity, this quickly becomes infeasible.

8. CONCLUSIONS AND FUTURE WORK

In this paper we have described the difficulties of using a sampling based approach for estimating classifier recall and introduced a more efficient method by making use of pairs of conditionally independent classifiers. While dependent on a few key assumptions, our method is able to produce high fidelity recall estimates without relying on a massive number of human judgements. We have described redundancies that often exist in document collections which would allow such independent classifier pairs to be created, and have provided a number of examples that are particularly relevant to social media updates. Numerical results showing the efficacy of our techniques have been presented with keyword based classifiers on Tweets, trained regression models on Twitter stories, and trained regression models on ODP data.

Many avenues remain open for investigation in this line of research. One possibility is to develop and experiment with ways to use multiple classifiers to get better recall estimates. This could be done either by averaging estimates from pairs of classifiers being used for prevalence estimation, or by using a third classifier in our current scheme to estimate the deviation from the assumptions given in Section 5.1. Still another direction to pursue could be an exploration of techniques which require even weaker conditions that the conditional independence utilized throughout this work.

9. ACKNOWLEDGMENTS

Thanks to Twitter and @WalmartLabs for providing data and resources with which to conduct experiments. Thanks also to Shuang-Hong Yang for providing suggestions on the text classifiers used in numerical experiments as well as Dong Wang and Pankaj Gupta for feedback on initial drafts of this work.

10. REFERENCES

[1] C. Manning P. Raghavan and H. Schuetze. *Introduction to Information Retrieval*. Cambridge University Press, 2011.

[2] Shuang Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. Large-scale high-precision topic modeling on twitter. *paper pending review*, 2014.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *COLT: Proceedings of the Workshop on Computational Learning Theory*, pages 92–100, 1998.

[4] George Forman. Counting positives accurately despite inaccurate classification. *ECML*, 2005.

[5] William Webber. Approximate recall confidence intervals. *CoRR*, abs/1202.2880, 2012.

[6] Xiaofeng He, Lei Duan, Yiping Zhou, and Byron Dom. Threshold selection for web-page classification with highly skewed class distribution. *WWW*, pages 1081–1082, 2009.

[7] P.N. Bennett and V. Carvalho. Online stratified sampling: Evaluating classifiers at web-scale. *Short-paper in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*, October 2010.

[8] N. Japkowicz and M. Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 1st edition, 2008.

[9] C. W. Cleverdon. The cranfield tests on index language devices. *Aslib Proceedings*, 19:173–192, 1967.

[10] Timothy G Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 601–610. ACM, 2009.

[11] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252, 1999.

[12] E. Law and L. von Ahn. *Human Computation (Synthesis Lectures on Artificial Intelligence and Machine Learning)*. Morgan & Claypool Publishers, 1st edition, 2011.

[13] Marios Kokkodis and Panagiotis G Ipeirotis. Have you done anything like that?: predicting performance using inter-category reputation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 435–444. ACM, 2013.

[14] J. Rice. *Mathematical Statistics and Data Analysis*. Thompson/Brooks/Cole, 2nd edition, 2007.

[15] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, pages 86–93, New York, NY, USA, 2000. ACM.

[16] A. Kolcz, G. Hulten, and J. Szymanski. Topical host reputation for lightweight url classification. In *Microsoft Research Technical Report*, 2012.