Notes on Term Selection

Endre Boros

November 25, 2002

Feature selection is a standard approach to reduce the dimension of data sets, as well as to eliminate noise (see [1] for an overview of this area as well as for related results). In our setting "term selection" can be an effective preprocessing step to achieve both of these goals. Appropriately selected terms may also provide a comprehensible and condensed summary of the differences between two sets $\mathbb{R}$ and $\mathbb{I}$ of documents (messages, emails, etc.). The sets $\mathbb{R}$ and $\mathbb{I}$ may represent documents belonging to two "small" time windows in a "large" document stream, or may simply be the sets of known **r**elevant and **i**rrelevant documents for a particular topic.

In our study we focus on computationally fast methods to select a small number of terms which distinguish well documents belonging to $\mathbb{R}$ from those in $\mathbb{I}$. ("Term" can stand for any binary feature one can imagine and use to represent documents; in our study a "term" is simply a word which is considered either present or absent from a particular document, i.e. we disregard multiplicities, ordering, etc.).

Let us associate with the pair $(\mathbb{R}, \mathbb{I})$ a relevance function $\chi : \mathbb{R} \cup \mathbb{I} \longrightarrow \{0, 1\}$ defined by $\chi(d) = 1$ for $d \in \mathbb{R}$ and $\chi(d) = 0$ for $d \in \mathbb{I}$. Let $R = |\mathbb{R}|$, $I = |\mathbb{I}|$, and let $\mathbb{T} = \mathbb{T}(\mathbb{R}, \mathbb{I})$ denote the sets of all terms appearing in $\mathbb{R} \cup \mathbb{I}$. For a term t and document d let $\chi(t, d) = 1$ if $t \in d$. Finally, for a subset S of the terms and for two documents x and y let us denote by $h_S(x, y)$ the *Hamming distance of x and y with respect to S*, that is the number of those terms from S which belong to exactly one of the two documents x and y. Clearly, from the set S only these terms can help us to differentiate x from y.

To every term $t \in \mathbb{T}(\mathbb{R}, \mathbb{I})$ we associate two integers, $a(t) = \sum_{d \in \mathbb{R}} \chi(t, d)$ and $b(t) = \sum_{d \in \mathbb{I}} \chi(t, d)$, that is the numbers of relevant, resp. irrelevant documents t belongs to. These parameters can efficiently be computed, and updated even for document streams. Based on these parameters, we associate a score $\rho(a(t), b(t))$ to every term, where ρ is measuring in some sense how much help term t provides to our objective of distinguishing documents in $\mathbb{R}$ from those in $\mathbb{I}$.

In our approach we rank the terms according to their $\rho(a(t), b(t))$ scores, and choose the top k terms $S = S(k, \rho)$. We also compute the average Hamming distance between $\mathbb{R}$ and $\mathbb{I}$ with respect to this set $S(k, \rho)$ of terms, i.e.

$$H(k, \rho) = \frac{1}{|\mathbb{R}||\mathbb{I}|} \sum_{d \in \mathbb{R}} \sum_{d' \in \mathbb{I}} h_{S(k,\rho)}(d, d').$$

Clearly, $0 \leq H(k,\rho) \leq k$, and the higher $H(k,\rho)$ is, the better we can distinguish documents in $\mathbb{R}$ from those in $\mathbb{I}$ using only terms form $S(k,\rho)$. We study this average Hamming distance both as a function of k, the number of terms used, and ρ, the scoring measure.

For the measure ρ we considered many different choices, all of which we scaled to be nonnegative and increasing in the sense that larger values supposedly correspond to terms differentiating better $\mathbb{R}$ from $\mathbb{I}$. We considered many simple measures from the literature, including the *correlation* of the appearances of a term with the relevance function χ of the document collection $(\mathbb{R},\mathbb{I})$, the so called *Gini index*, the χ^2 statistics, the change of *entropy*, etc. Some of these in fact are equivalent in the sense that they will rank the terms in the same order (e.g. the absolute value of the correlation and the χ^2 statistics, etc.) There are however many other possible measures based on the parameters $a(t)$ and $b(t)$, not (yet) considered in the literature. To narrow down the choices, we assumed certain basic "axiomatic" properties, such as $\rho(|\mathbb{R}|,0)$ and $\rho(0,|\mathbb{I}|)$ both should be as large as possible, while $\rho(|\mathbb{R}|,|\mathbb{I}|) = \rho(0,0) = 0$, as well as some natural monotonicity. Examples include

$$\mu(a(t),b(t)) = \frac{a(t)|\mathbb{I}| + b(t)|\mathbb{R}| - 2a(t)b(t)}{|\mathbb{R}||\mathbb{I}|}$$

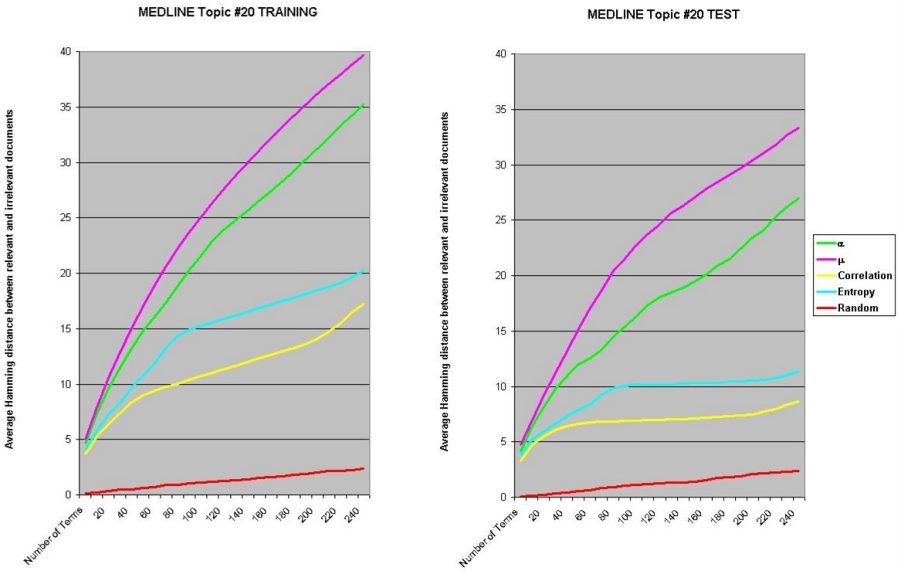
measuring exactly the contribution of term t to the average Hamming distance, or

$$\alpha(a(t),b)(t)) = \left| \frac{a(t)}{|\mathbb{R}|} - \frac{b(t)}{|\mathbb{I}|} \right|$$

a simplified version of μ, etc.

Among the many possibilities we chose 16 different measures, including those from the literature, and tabulated $H(k,\rho)$ for all these measures ρ and for $k = 10, 20, ..., 250$. We repeated our experiments for a large number of topics and for different document collections. To evaluate the *differentiating power* of the obtained term sets, in each case we used the same set of $k = 10, 20, ..., 250$ terms, and computed the average Hamming distance with respect to this set of terms of a (typically larger) independent *test collection* $(\mathbb{R}',\mathbb{I}')$ of relevant and irrelevant documents.

As an illustration we include here the graphs of these results for a topic from the MEDLINE collection, both for the training and the test sets. To simplify the picture, we include only 4 selected measures together with a random ranking of the terms, which we included as a baseline. These results are quite typical across topics and collections.



What is not surprising in these results: The fact that μ provides the best separation on the training set, since choosing the top ranked terms according to this measure amounts to maximizing the average Hamming distance between $\mathbb{R}$ and $\mathbb{I}$ (see definition of μ); the fact that the random ranking is the poorest separator, since most terms are just "noise" for any given topic, and most terms appear only in a few of the documents.

What is surprising in these results: The fact that the same set of terms provide reasonably high separation in the independent test collection, in particular in case of μ; the facts that such usual measures as correlation, or entropy select terms which provide substantially weaker separation on the test set: for instance the top 100 hundred terms selected with μ provides a Hamming distance of 24 on average on the training set, and 22 on the test set – a loss of less than 10%, while if we select the top 100 terms with correlation, we get only around 10 as the average Hamming distance on the training set, and about 7 on the test set, a loss of about 30%.

Validation

To test the above findings, we used these results within the TREC 2002 batch filtering task (see [3]).

We chose 5 measures which seemed to provide robust separation in our experiments (such as μ, α, etc.), and selected the set S of top 50 terms by the majority ranking of these 5 measures. These computations were done on the (small) training sets for each of the 100 topics.

Next, to each term in $t \in S$ we assigned the weight

$$\omega(t) = \frac{\frac{a(t)+0.5}{|\mathbb{R}|+1}}{\frac{b(t)+0.5}{|\mathbb{I}|+1}}$$

which can be seen to be the Bayesian weight of evidence (based on the training set $(\mathbb{R}, \mathbb{I})$), and to each test document d we assigned the score

$$\Omega(d) = \sum_{t \in S \cap d} \log(\omega(t)).$$

That is, each document projected onto the selected feature set S is assigned a score equal to the sum of the logarithms of the Bayesian weights of evidence for the terms it contains.

The batch filtering task requires the definition of a static classification rule which specifies whether each document in the test set should be considered relevant and retrieved, or irrelevant and ignored. We considered document d relevant if $\Omega(d) \geq \tau$ for some threshold $\tau \in \mathbb{R}$. The threshold τ was selected so as to optimize the utility measure $TU11 = 2 * RR - RI$ over the training set, where RR is the number of retrieved relevant documents and RI is the number of retrieved irrelevant documents.

Despite the simplicity of these computations, our system performed quite well. On the 100 topics our TU11 score was less than the median of all competing systems 16 times, equal to the median 26 times, more than the median 58 times, and attained the maximum 21 times.

The results of our batch filtering seemed to provide some evidence that employing this very simple and fast term selection can improve simple text classification methods and can be the foundation of a reasonably performing system.

References

[1] Endre Boros, Takashi Horiyama, Toshihide Ibaraki, Kazuhisa Makino and Mutsunori Yagiura, Finding Essential Attributes from Binary Data, *Annals of Mathematics and Artificial Intelligence*, to appear, 2002.

[2] Endre Boros and David J. Neu, Rank Based Feature Selection in Information Retrieval, RUTCOR, Rutgers University, in preparation, 2002.

[3] Endre Boros and David J. Neu, Feature Selection and Batch Filtering: TREC 2002, RUTCOR, Rutgers University, 2002.