Fast Text Classification Using Sequential Sampling Processes

Michael D. Lee*

Department of Psychology, University of Adelaide

Abstract. A central problem in information retrieval is the automated classification of text documents. While many existing methods achieve good levels of performance, they generally require levels of computation that prevent them from making sufficiently fast decisions in some applied setting. Using insights gained from examining the way humans make fast decisions when classifying text documents, two new text classification algorithms are developed based on sequential sampling processes. These algorithms make extremely fast decisions, because they need to examine only a small number of words in each text document. Evaluation against the Reuters-21578 collection shows both techniques have levels of performance that approach benchmark methods, and the ability of one of the classifiers to produce realistic measures of confidence in its decisions is shown to be useful for prioritizing relevant documents.

1 Introduction

A central problem in information retrieval is the automated classification of text documents. Given a set of documents, and a set of topics, the classification problem is to determine whether or not each document is about each topic. This paper presents two fast text document classifiers inspired by the human ability to make quick and accurate decisions by skimming text documents.

2 Existing Methods

A range of artificial intelligence and machine learning techniques have been applied to the text classification problem. A recent and thorough evaluation of five of the best performed methods is provided in [1]. The classifiers examined are:

* This research was supported by the Australian Defence Science and Technology Organisation. The author wishes to thank Peter Bruza, Simon Dennis, Brandon Pincombe, Douglas Vickers, and Chris Woodruff. Correspondence should be addressed to: Michael D. Lee, Department of Psychology, University of Adelaide, SA 5005, AUSTRALIA. Telephone: +61 8 8303 6096, Facsimile: +61 8 8303 3770, E-mail: michael.lee@psychology.adelaide.edu.au, URL: http://www.psychology.adelaide.edu.au/members/staff/michaellee

M. Brooks, D. Corbett, and M. Stumptner (Eds.): AI 2001, LNAI 2256, pp. 309–320, 2001.
© Springer-Verlag Berlin Heidelberg 2001

- Support Vector Machines (SVM), which use a training set to find optimal hyperplanes that separates documents into those about a topic, and those not about a topic. These hyperplanes are then applied to classify new documents.
- k-Nearest Neighbor classifiers (kNN), which classify new documents according to the known classifications of its nearest training set neighbors.
- Linear Least Squares Fit classifiers (LLSF), which generate a multivariate regression model from a training set that can be applied to new documents.
- Neural Network classifiers (NNet), which learn the connection weights within a 3-layer neural network using a training set, and then applies this network to classify new documents.
- Naive Bayes classifiers (NB), which use a training set estimate the probabilities of words indicating documents being about topics, and uses a simple version of Bayes theorem with these probabilities to the classify new documents.

Different performance measures show different levels of relative performance for the five classifiers, although the SVM and kNN are generally the most effective, followed by the LLSF, with the NNet and NB classifiers being the least effective [1]. What is important, from an applied perspective, is the considerable degree of computation undertaken by each classifier, either during the training process, the process of classifying new documents, or both. SVMs, for example, require the solution to a quadratic programming problem during training, LLSF classifiers must solve a large least-squares problem, and NNets are notoriously time consuming to train.

In classifying new documents, most existing techniques consider every word in the document, and often have to calculate involved functions. This means that they take time to process large corpora. In many applied situations, analysts require fast 'on-line' text document classification, and would be willing to sacrifice some accuracy for the sake of timeliness. The aim of this paper is to develop text classifiers that emphasize speed rather than accuracy, and so the results in [1] are used as guides on acceptable performance, rather than benchmarks to be exceeded.

3 Some Insights from Psychology

As with many artificial intelligence and machine learning problems, there is much to be learned from examining the way in which humans perform the task of text classification. In particular, it is worth making the effort to understand how people manage to make quick and accurate decisions regarding which of the many text documents they encounter every day (e.g., newspaper articles) are about topics of interest.

3.1 Bayesian Decision Making

A first psychological insight involves the relationship between the decisions "this document is about this topic" and "this document is not this topic". When

people are asked to make this decision, they actively seek information that would help them make either choice. They do not look only for confirming information in the hope of establishing that the document is about the topic, and conclude otherwise if they fail to find enough information.

For example, if people are asked whether a newspaper article is about the US Presidential Elections, consider the following three scenarios:

- The first word is "The". In this case, most people would not be able to make any decision with any degree of confidence.
- The first word is "Cricket". In this case, most people would confidently respond 'No'.
- The first word is "Gore". In this case, most people would confidently respond 'Yes'.

The fact that people are able to decide to answer 'No' in the second scenario suggests that they are actively evaluating the word as evidence in favor of the document not being about the topic (in the same way they actively evaluate the word "Gore" in the third scenario). This behavior suggests that people treat the choices "this document is about this topic" and "this document is not this topic" as two competing models, and are able to use the content of the document, in a Bayesian way, as evidence in favor of either model. Many established text classifiers, including the kNN, LLSF and NNet classifiers, do not operate this way. In general terms, these classifiers construct a measure of the similarity between the document in question, and some abstract representation of the topic in question. When the measure of similarity exceeds some criterion value, the decision is made that the document is about the topic, otherwise the default decision is made that the document is not about the topic. The text classifiers developed here, however, actively assesses whether the available information allows the decision "this document is not about this topic" to be made. Adopting this approach dramatically speeds text classification, because it is often possible to determine that a document is not about a topic directly, rather than having to infer this indirectly from failing to establish that it is about the topic.

At the heart of the Bayesian approach are measures of the evidence individual words provide for documents either being about a topic, or not being about a topic. The evidence that the i-th word in a dictionary provides about topic $\mathbf{T}$, denoted by $V_{\mathbf{T}}(w_i)$, may be calculated on a log-odds scale as follows:

$$V_{\mathbf{T}}(w_i) = \ln \frac{p(w_i \mid \mathbf{T})}{p(w_i \mid \bar{\mathbf{T}})} \approx \ln \frac{|w_i \in \mathbf{T}| \,/\, |\mathbf{T}|}{|w_i \in \bar{\mathbf{T}}| \,/\, |\bar{\mathbf{T}}|},$$

where $\mathbf{T}$ is "about a topic", $\bar{\mathbf{T}}$ is "not about a topic", $|w_i \in \mathbf{T}|$ is the number of times word w_i occurs in documents about topic $\mathbf{T}$, and $|\mathbf{T}|$ is the total number of words in documents about topic $\mathbf{T}$. Note that these evidence values are symmetric about zero: Words with positive values (e.g., "Gore") suggest that the document is about the topic, words with negative values (e.g., "cricket") suggest that the document is not about the topic, and words with values near zero (e.g., "the") provide little evidence for either alternative.

3.2 Non-compensatory Decision Making

A second psychological insight is that, when people decide whether or not a text document is about a topic, they often make non-compensatory decisions. This means that people are able to make a decision without considering all of the content of a document. For example, if asked whether a newspaper article is about the US Presidential Elections, and the first 11 words read are "Alan Border yesterday questioned the composition of the Australian cricket team ...", most people would choose to answer 'No', even if they were permitted to examine the remainder of the article. In making non-compensatory decisions,

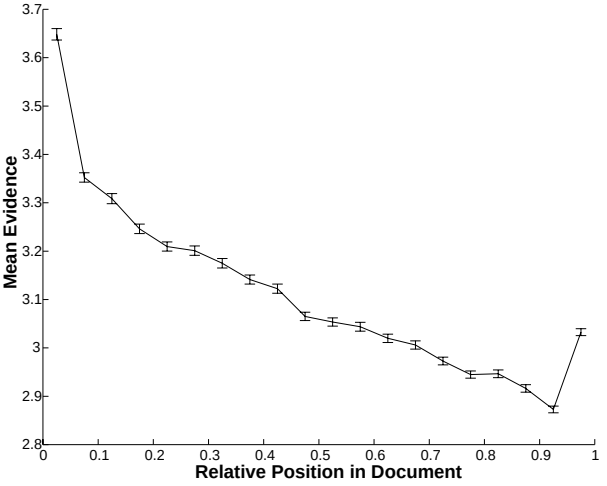

Fig. 1. The mean absolute evidence provided by words in the Reuters-21578 Corpus, as a function of their relative position in the document.

people rely on regularities in their environment [2]. In the case of text documents, they assume that words near the beginning will provide some clear indication of the semantic topic. This assumption is borne out by the analysis of the entire Reuters-21578 collection presented in Figure 1, which shows the mean absolute evidence provided by words according to their relative position in the documents. Words at the beginning of documents provide relatively more evidence than those in the middle or near the end, although there is a small increase for words at the very end, presumably associated with the 'summing up' of documents. The important point, for the purposes of fast text classification, is that it is possible to know *a priori* those words in a document that will be the most useful for making a decision. Figure 1 suggests that, at least for news-style documents, they will be words at or near the beginning of the document.

3.3 Complete Decision Making

A final psychological insight is that when people decide whether or not a text document is about a topic, they undertake a decision process that generates more information than just a binary choice. People give answers with a certain level of accuracy, having taken a certain period of time, and are able to express a certain level of confidence in their decision. An automatic text classification system capable of providing the same sort of response outputs seems likely to have advantages in many applied situations.

4 Sequential Sampling Process Models

Within cognitive psychology, the most comprehensive accounts of human decision making are provided by sequential sampling models. In particular, a number of 'random-walk' and 'accumulator' models have been developed, and demonstrated to be successful in a variety of experimental situations [3,4]. These models are based on the notion of accruing information through the repeated sampling of a stimulus, until a threshold level information in favor of one alternative has been collected to prompt a decision.

Both random walk and accumulator models naturally capture the three psychological insights into the text classification problem. Both models use a Bayesian approach to model selection, in the sense that they establish explicit thresholds for both of the possible decisions. The use of thresholds also means that non-compensatory decisions are made, since the stimulus is only examined until the point where the threshold is exceeded. Furthermore, by examining the words in a text document in the order that they appear in the document, those words that are more likely to enable a decision to be made will tend to be processed first. Finally, both models are able to generate measures of confidence in their decisions.

This integration of the psychological insights suggests text classifiers that operate by examining each word in a text document sequentially, evaluating the extent to which that word favors the alternative decisions "this document is about the topic" and "this document is not about the topic", and using the evidence value to update the state of a random-walk or accumulator model.

4.1 Random Walk Text Classifier

In random walk models, the total evidence is calculated as the difference between the evidence for the two competing alternatives, and a decision is made once it reaches an upper or lower threshold. This process can be interpreted in Bayesian terms [5], where the state of the random walk is the log posterior odds of the document being about the topic. Using Bayes' theorem, the log posterior odds are given by

$$\ln \frac{p\left(\mathbf{T} \mid \mathbf{D}\right)}{p\left(\bar{\mathbf{T}} \mid \mathbf{D}\right)} = \ln \frac{p\left(\mathbf{T}\right)}{p\left(\bar{\mathbf{T}}\right)} \frac{p\left(\mathbf{D} \mid \mathbf{T}\right)}{p\left(\mathbf{D} \mid \bar{\mathbf{T}}\right)},$$

where $\mathbf{D}$ is the document being classified in terms of topic $\mathbf{T}$. Assuming the document is appropriately represented in terms of its n words $w_1, w_2, \ldots, w_n$, which is probably the most justifiable assumption, although it is certainly not the only possibility, this becomes

$$\ln \frac{p\left(\mathbf{T} \mid \mathbf{D}\right)}{p\left(\bar{\mathbf{T}} \mid \mathbf{D}\right)} \approx \ln \frac{p\left(\mathbf{T}\right)}{p\left(\bar{\mathbf{T}}\right)} \frac{p\left(w_1, w_2, \ldots, w_n \mid \mathbf{T}\right)}{p\left(w_1, w_2, \ldots, w_n \mid \bar{\mathbf{T}}\right)}.$$

If it is further assumed that each word provides independent evidence, which is more problematic, but is likely to be a reasonable first-order approximation, the log posterior odds becomes

$$\begin{aligned} \ln \frac{p\left(\mathbf{T} \mid \mathbf{D}\right)}{p\left(\bar{\mathbf{T}} \mid \mathbf{D}\right)} &= \ln \frac{p\left(\mathbf{T}\right)}{p\left(\bar{\mathbf{T}}\right)} + \ln \frac{p\left(w_1 \mid \mathbf{T}\right)}{p\left(w_1 \mid \bar{\mathbf{T}}\right)} + \ln \frac{p\left(w_2 \mid \mathbf{T}\right)}{p\left(w_2 \mid \bar{\mathbf{T}}\right)} + \ldots + \ln \frac{p\left(w_n \mid \mathbf{T}\right)}{p\left(w_n \mid \bar{\mathbf{T}}\right)} \\ &= \ln \frac{p\left(\mathbf{T}\right)}{p\left(\bar{\mathbf{T}}\right)} + V_{\mathbf{T}}\left(w_1\right) + V_{\mathbf{T}}\left(w_2\right) + \ldots + V_{\mathbf{T}}\left(w_n\right). \end{aligned}$$

This final formulation consists of a first 'bias' term, given by the log prior odds, that determines the starting point of the random walk, followed by the summation of the evidence provided by each successive word in the document.

Once a random walk has terminated, and a decision made according to whether it reached an upper or lower threshold, a measure of confidence in the decision can be obtained as an inverse function of the number of words examined. For documents that require many words to classify, confidence will be low, while for documents classified quickly using few words, confidence will be high.

Figure 2 summarizes the operation of the random walk classifier on a document from the Reuters-21578 collection that is about the topic being examined. The state of the random-walk is shown as the evidence provided by successive words in the document are assessed. A threshold value of 50 is shown by the dotted lines above and below. This example highlights the potential of non-compensatory decision making, because the first 100 words of the documents allow the correct decision to be made, but the final state of the random-walk, when the entire document has been considered, does not lead to the correct decision being made.

4.2 Accumulator Text Classifier

The accumulator text classifier is very similar to the random walk version, except that separate evidence totals are maintained, and a decision is made when either one of them reaches a threshold. This means that evidence provided by each successive word $V_{\mathbf{T}}\left(w_i\right)$ is added to the "is about topic" accumulator $A_{\mathbf{T}}$ if it is positive, and to the "is not about accumulator" $A_{\bar{\mathbf{T}}}$ if it is negative. Once either $A_{\mathbf{T}}$ reaches a positive threshold, or $A_{\bar{\mathbf{T}}}$ reaches a negative threshold, the corresponding decision is made. The confidence in this decision is then measured according to the difference between the evidence totals accumulated for each decision, as a proportion of the total evidence accumulated, as follows: $\left(A_{\mathbf{T}} - |A_{\bar{\mathbf{T}}}|\right) / \left(A_{\mathbf{T}} + |A_{\bar{\mathbf{T}}}|\right)$.

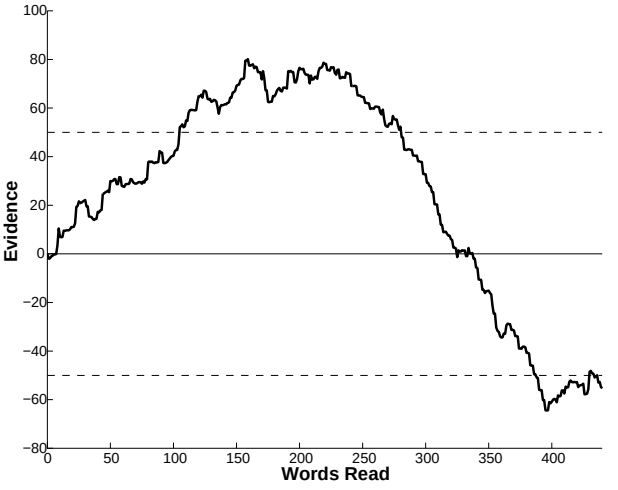

Fig. 2. Operation of the random walk text classifier in a case where the document is about the topic in question

5 Evaluation against Reuters-21578

5.1 Standard Information Retrieval Measures

The random walk and accumulator classifiers were evaluated using the ModApte training/test split detailed in [6] to enable comparison with the benchmark results presented in [1]. In the interests of ensuring speed, the corpus was not pre-processed to the same extent as [1]. In particular, no word-stemming was undertaken. The only pre-processing was to filter the documents into lower case characters {a...z} together with the space character. The performance of the text classifiers was measured in five standard ways [7]: recall, precision, macro F1, micro F1, and error rate.

Precision, p, measures the proportion of documents the classifier decided were about a topic that actually were about the topic. Recall, r, measures the proportion of documents actually about a topic that were identified as such by the classifier. The two versions of the F1 measure, $\mathrm{F1} = 2rp/(r+p)$ were obtained by different forms of averaging. The first was obtained by 'micro-averaging', where every decision made by the classifier was aggregated before calculating recall and precision values. The second was obtained by 'macro-averaging', where recall and precision values were calculated for each topic separately, and their associated F1 values were then averaged. As argued in [1], it is important to consider both approaches when using corpora, such as Reuters-21578, where the distribution of topics to documents is highly skewed. The error was simply measured as the percentage of incorrect decisions made by the classifier over all document-topic combinations. These measures were based on modified 'forced-

choice' versions of the random walk and accumulator classifiers, where a decision was made even when no threshold had been reached at the end of the document. For the random walk classifier, this decision was made on the basis of whether the final state was positive or negative. For the accumulator, the larger of the two accumulated totals was used to make a decision.

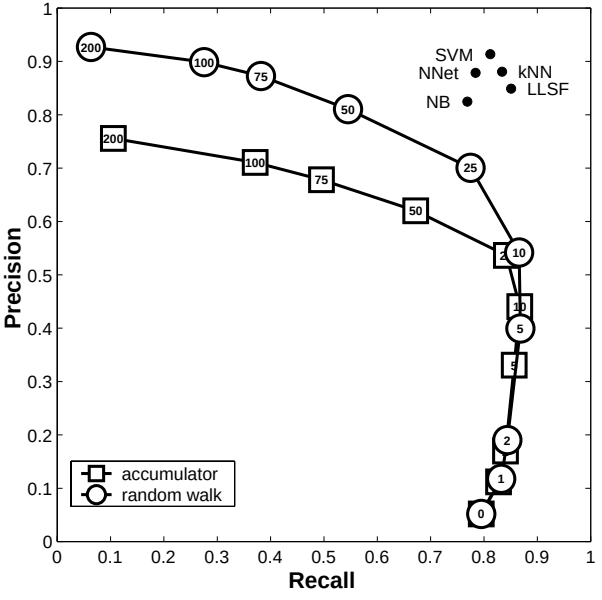

Fig. 3. Precision-recall performance of the random walk and accumulator text classifiers, together with existing benchmarks.

Figure 3 shows the precision and recall performance of the random walk and accumulator text classifiers for a range of different threshold values, together with the benchmark performances reported in [1]. While different applied settings can place different degrees of emphasis on recall and precision, the best balance probably lies at about the threshold value of 25. In terms of the existing benchmarks, both classifiers have competitive recall performance, but fall short in terms of precision. In practical terms, this means that the random walk and accumulator classifiers find as many relevant documents, but return 3 or 4 irrelevant documents in every batch of 10, whereas benchmark performance only return 1 irrelevant document in every batch of 10.

Figure 4 shows the micro F1 and macro F1 performance of the random walk and accumulator text classifiesr for the threshold values up to 25, together with the benchmarks. On these measures, both classifiers are very competitive and, in fact, outperform some existing methods on the macro F1 measure.

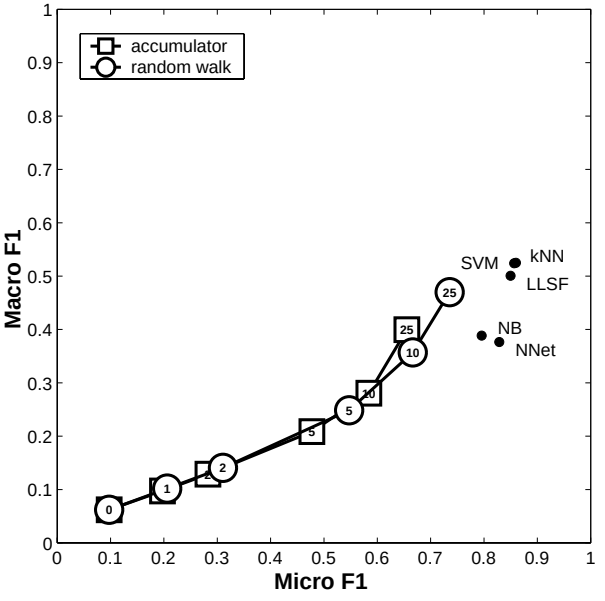

Fig. 4. Micro and Macro F1 performance of the random walk and accumulator text classifiers, together with existing benchmarks.

Table 1. Mean number of words examined, mean percentage of words examined, and mean percentage error of the forced choice random walk and accumulator text classifiers.

	Random Walk			Accumulator		
Threshold	Words	Percentage	Error	Words	Percentage	Error
0	1.06	0.8%	18.5%	1.06	0.8%	18.5%
1	1.59	1.3%	8.1%	1.54	1.3%	8.5%
2	1.99	1.7%	4.7%	1.88	1.6%	5.4%
5	3.45	2.9%	1.8%	2.96	2.5%	2.4%
10	6.72	5.6%	1.1%	5.48	4.6%	1.6%
25	15.7	13.1%	1.0%	12.8	10.7%	1.2%
50	29.0	24.2%	1.0%	24.9	20.8%	1.1%
75	40.4	33.6%	1.0%	35.9	29.9%	1.1%
100	49.9	41.6%	1.0%	45.3	37.8%	1.1%
200	75.1	62.6%	1.0%	71.2	59.4%	1.1%

Table 1 presents the mean number of words examined by each of the classifiers at each threshold, this mean count as a percentage of the average document length of the test set, and percentage error of the classifiers. It is interesting to note that the accumulator classifier generally requires fewer words than the random walk classifier. More importantly, these results demonstrate the speed with which the classifiers are able to make decisions. At a threshold of 25, only 10–13% of the words in a document need to be examined on average for classification at a 1% error rate. Given the computational complexity of existing methods, it seems reasonable to claim that the random walk and accumulator classifiers would have superior performance on any 'performance per unit computation' measure.

5.2 Confidence and Prioritization

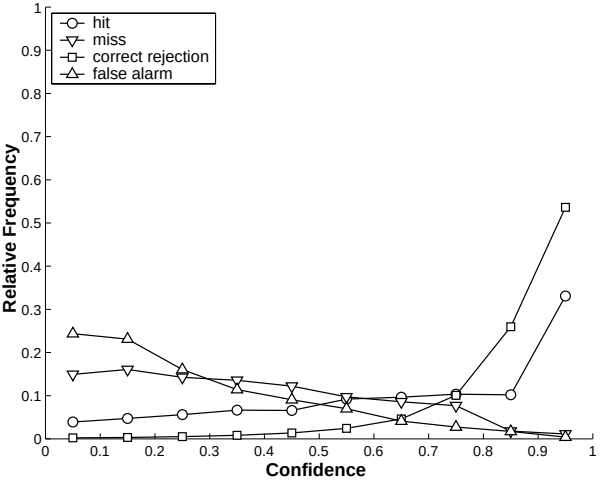

Fig. 5. Confidence distributions for the forced-choice version of the accumulator classifier.

For the forced choice versions of the classifiers, it is informative to examine the distribution of confidence measures in terms of the standard signal detection theory classes of 'hit', 'miss', 'correct rejection' and 'false alarm'. These distributions are shown at a threshold of 25 for the accumulator classifier in Figure 5, and for the random walk classifier in Figure 6. The measures of confidence generated by the accumulator are meaningful, in the sense that hits and correct rejections generally have high confidence values, while misses and false alarms generally have low confidence values. The random walk confidence measures, in contrast, do not differ greatly for any of the four decision classes and, in fact, the classifier is generally more confident when it misses than when it hits.

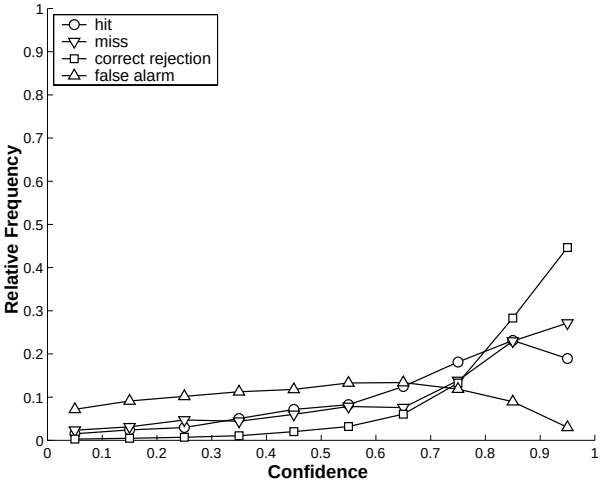

Fig. 6. Confidence distributions for the forced-choice version of the random walk classifier.

The ability of accumulator models to provide more realistic confidence measures than random walk models has been observed within psychology [4], and has practical implications for text classification. In particular, the confidence measures can be used as 'relevancy' scores to order or prioritize the decisions made by the classifiers. The obvious way of doing this is to return all of the documents that were classified as being about topics first, ranked from highest confidence to lowest confidence, followed by the documents not classified as being about topics, ranked from lowest confidence to highest.

This prioritization exercise was undertaken for both of the classifiers on all of the possible document-topic combinations, and the results are summarized by the 'effort-reward' graph shown in Figure 7. The curves indicate the proportion of relevant documents (i.e., the reward) found by working through a given proportion of the prioritized list (i.e., the effort). It can be seen that both classifiers return almost 90% of the relevant documents in the first 5% of the list, but that the accumulator then performs significantly better, allowing all of the relevant documents to be found by examining only the top 20% of the list.

6 Conclusion

This paper has presented two text classifiers based on sequential sampling models of human decision making. Both techniques achieve reasonable levels of performance in comparison to established benchmarks, while requiring minimal computational effort. In particular, both classifiers are capable of making extremely fast decisions, because they generally need to examine only a small proportion

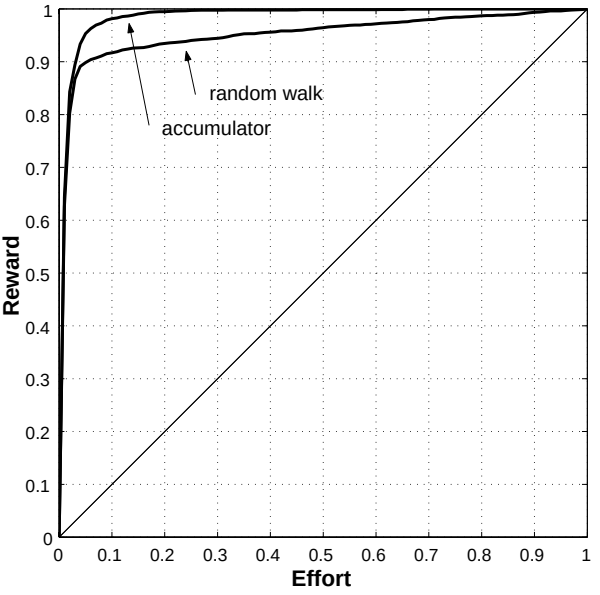

Fig. 7. Effort-reward performance for priorization using the forced-choice accumulator and random walk classifiers.

of the words in a document. The ability of the accumulator classifier to generate meaningful confidence measures has also been demonstrated to be useful in presenting prioritized lists of relevant text documents.

References

[1] Y Yang and X Liu, "A re-examination of text categorization methods," in *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkley, CA, 1999, pp. 42–49, ACM.

[2] G Gigerenzer and P M Todd, *Simple Heuristics That Make Us Smart*, Oxford University Press, New York, 1999.

[3] P L Smith, "Stochastic dynamic models of response time and accuracy: A foundational primer," *Journal of Mathematical Psychology*, vol. 44, pp. 408–463, 2000.

[4] D Vickers and M D Lee, "Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module," *Non-linear Dynamics, Psychology, and Life Sciences*, vol. 2, no. 3, pp. 169–194, 1998.

[5] R E Kass and A E Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[6] D D Lewis, "Reuters-21578 text categorization test collection," 1997, Available at http://www.research.att.com/∼lewis/reuters21578/readme.txt.

[7] C J Van Risjbergen, *Information Retrieval*, Butterworths, London, 1979.