Term Selection for Filtering based on Distribution of Terms over Time

Avi T. Arampatzis **Th.P. van der Weide** **C.H.A. Koster** **P. van Bommel**

Faculty of Mathematics and Computing Science, University of Nijmegen,
Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands.
tel: +31 24 3653147, fax: +31 24 3553450

E-mail: {avgerino|tvdw|kees|pvb}@cs.kun.nl

In: *Proceedings of RIAO'2000*, 12-14 April 2000, Paris.

Abstract

In this article we investigate the use of time distributions in retrieval tasks. Specifically, we introduce a novel term selection method, namely *Term Occurrence Uniformity (TOU)*, based on the hypothesis that terms which occur uniformly in time are more valuable than others. Our empirical evaluation so far has neither proved nor disproved this hypothesis. However, results are promising and suggest the need for a deeper theoretical and empirical investigation. Our current concern is filtering, but this line of research may easily be extended to other retrieval tasks which involve temporally-dependent data.

1 Introduction

Information Filtering is the process of searching in large amounts of data for information which matches a user information need. The filtering task is usually described as the inverse of the traditional retrieval task. In retrieval, a one-time user request (called query) is matched to a static collection of information objects. In filtering, users issue a long-term request (called *profile*) which is compared to a dynamic collection, for instance, a *stream* of arriving information objects. Filtering may also be seen as a *binary classification/categorization* task where each arriving object has to be classified under one of two categories: relevant, or not.

Information objects and requests are represented by some characterization language, allowing to compute comparisons of aboutness between them. Representations are usually made of bags of weighted *terms* (also called *features* in classification) derived from information objects. The long-term nature of user needs in filtering may be exploited to build more complete and precise profiles than the corresponding queries in retrieval. This is done by accounting for user's (explicit or implicit) relevance feedback on selected objects from a stream. A collection of objects with user relevance judgments is called *training data* and may be used to build or update profiles.

The process of constructing profiles mainly consists of *term selection* from training objects, and *term weighting* within profiles[1]. Even moderate-sized training data may contain tens of thousands of terms; nevertheless, not all of them are suitable or necessary for representing an information need. Moreover, suitable terms differ in their ability to represent a need, thus they are weighted accordingly.

[1] in contrast to term weighting within documents, e.g. *tf*×*idf* (term frequency × inverted document frequency) weighting.

The extensive research heritage of retrieval and the close similarity of filtering and retrieval tasks have led researchers to see filtering as an attractive application for techniques that developed for retrieval [Belkin and Croft, 1992]. As a result, qualitative differences of filtering, which may influence the effectiveness of proven retrieval techniques when these are adapted, are usually overlooked.

Current term selection and weighting techniques have been originally developed in the traditional retrieval context. These techniques mostly consider training data as unordered sets of objects, totally disregarding their time of arrival. To our knowledge, temporal information has not been widely explored in retrieval environments. The most closely related subject is *event detection*, i.e. the identification of novel events in news streams [Yang et al., 1998, Allan et al., 1998].

In this paper we investigate ways to incorporate temporal information into profile construction by assuming that terms which are distributed uniformly, either in the series or in actual time-line of relevant training objects, are more valuable than others. In the next section, we elaborate on this hypothesis, which we have until now tested in a term selection context. Section 3 summarizes the two temporally-dependent term selection methods we experimented with. Both were compared to the baseline of selecting terms with *document frequency thresholding*. Document frequency thresholding has proved to be more than an *ad hoc* approach and quite powerful in feature selection in categorization environments [Yang and Pederson, 1997]. In section 4, we define a *term occurrence uniformity* measure derived from a variant of *Kolmogorov-Smirnov* discrepancy test, namely, *Kuipers'* statistic. All term selection schemes are evaluated in a batch text filtering framework based on Rocchio's relevance feedback method. As a dataset we used the Reuters-21578 text categorization test collection. The experimental setup, properties of the dataset and pre-processing are discussed in section 5. We give some experimental evidence on how our proposed schemes perform in section 6. The evaluation so far has been rather inconclusive and suggests the need for a deeper theoretical and empirical investigation.

Currently, our concern is filtering but the results of this line of research, in principle, apply or can be interpreted for all retrieval tasks which may involve temporally dependent data, e.g. routing, or classification. With the large amounts of digital information that become available daily, the importance of more effective and efficient retrieval environments is obvious.

2 The Term Occurrence Uniformity Conjecture

Term selection and term weighting techniques developed for retrieval tasks usually consider object collections as unordered sets. Thus, the arrival time of objects in filtering is totally disregarded when selecting and weighting terms using traditional retrieval techniques. Quoting David Hull from TREC-7's Filtering Track [Hull, 1998] (the last TREC proceedings published by the time of writing of this article):

> *"... no one has yet explored whether the distribution of a feature over time is related to its usefulness as a discriminator for relevance."*

Changes in the distribution of a feature over time may indicate several things, for example:

- A slow monotonic change in the occurrence rate of a term in relevant objects may indicate a *topic drift*, i.e. a slow shift in the focus of a user's interest over time [Allan, 1996], or even a *concept drift*, where the meaning of a concept changes [Brodley and Rissland, 1993]. Either case means that arriving relevant objects tend to be different than training data, and this difference is becoming greater over time. The quality of filtering will slowly degrade, unless an adaptive filter responses adequately to these changes.

- A sudden increase of the occurrence rate of a term in relevant objects may indicate a *temporal event*. For instance, *NYC subway bombing* is an event relevant to the topic *terrorism*. Such important events are usually associated with bursts of incoming documents for some period of time. A fast-responding filter, trained for topic *terrorism*, could deceptively be adapted as a result of the very frequent occurrence of terms *NYC* and *subway*, which in general are not characteristic terms of terrorism.

Topic and concept drift are related, in the sense that the idea of relevance changes. [Allan, 1996] demonstrated that such drifts can be handled readily by phasing out old context. However, in a topic with temporal events the idea of relevance remains unaltered, but it is the content of relevant objects that changes temporarily.

In this paper we assume that topics are stable (no drifts of any kind) but have temporal events. As an example consider a document stream, e.g. an electronic Newswire issuing on a daily basis articles about politics, sports, entertainment, etc. With respect to its power in characterizing a certain topic (e.g. *football*), each term can be classified as: *relevant*, *non-relevant*, *spurious*, or *indifferent*. Words like *the, are*, and all other common function words (*stop-words*) are indifferent. Stop-words have very low semantic content and occur in almost all documents and topics, thus are incapable of characterizing anything in particular. Others words occur too sparsely to make any difference in accuracy, e.g. *Jenkins* (the goal-keeper of Rising Hope FC). A word like *dollars*, which has semantic content and may be taken as relevant (think of dollars spent for player transfers), may actually be spurious if it is also related to other topics of the stream[2] (think of government funds in politics articles). The word *moonshine* is definitely non-relevant. *Game* is a relevant term since it occurs very frequently in documents about football. The less frequent but relevant, *injury* and *world-cup* will also occur. Every few games an injury worth mentioning happens, and every four years for a period of a few weeks most games are played for the world-cup. The former term occurs in a regular manner, while the latter occurs in a temporally clustered manner[3]. Thus, relevant terms may further be classified as: *regularly* or *temporally relevant*. Table 1 summarizes the term relevance classification we consider.

classification		**examples** (for topic "football")
relevant	regular	*football, game, injury*
	temporal	*world-cup*
non-relevant		*moonshine*
spurious		*dollars*
indifferent	stop-terms	*and, the, are*
	sparse	*Jenkins*

Table 1: Term relevance with respect to a topic

Streams in filtering tasks, like the one we have just considered, usually contain too many terms. Thus it is not uncommon to end up with thousands of terms in the indexing vocabulary of a filtering system. Fortunately, most of the terms can safely be discarded as non-discriminating for a topic, reducing dramatically the dimensionality of the indexing space. This reduction of dimensionality is highly desirable, mainly for efficiency reasons. In the weighting phase, a large number of terms is difficult to handle for learning algorithms. For instance, few neural networks can handle a large number of nodes,

[2] note that *spuriousness* is both topic and stream related.

[3] in fact, in the case of *world-cup* the distinction between temporal and regular occurrence depends on the time-scale under consideration; for very large time scales ($\gg$ 4 years), *world-cup* occurs *regularly* every four years.

and probabilistic models will be computationally intractable unless term independence[4] is assumed.

Term selection and term weighting schemes disregarding time make no distinction between regularly and temporally relevant terms. Should these terms indeed be distinguished and treated differently? We can speculate that taking into account distributions of term occurrences over time may be useful:

> **Term Temporal Locality Hypothesis:** *Terms occurring frequently over a short period of time, rather than distributed over the whole time-line, do not have lasting predictive value.*

If this hypothesis is true,

- it can at least be used as a term selection mechanism. The corresponding terms can be removed without a negative impact in filtering effectiveness, but with a desirable benefit for efficiency. In fact, effectiveness may also improve slightly for the same reasons it is improved in classification tasks (see section 3) or if these terms happen to be *noise* terms. Alternatively,

- terms with temporal occurrence characteristics can be down-weighted by learning algorithms, hopefully reducing *classification noise* and gaining effectiveness. However, classification noise is related to the nature of topics and streams, and usually occurs for narrow topics when the stream contains closely-related but irrelevant objects.

In this paper we investigate the hypothesis in the term selection context. We will report on our attempts to create or alter profile term weights elsewhere. We consider two term selection approaches, the first is based on the *order of arrival (time-order)*, and the second on the actual *time of arrival (time-stamp)* of relevant objects. The approaches are identical in case the relevant objects arrive with a constant rate of objects per time unit. Given a stream of objects relevant to a topic, for each occurring term we define a quantity we call *Term Occurrence Uniformity (TOU)* simply as:

> **Term Occurrence Uniformity:** *the degree to which the term occurrences are fairly distributed in every possible interval of the stream with respect to interval's length.*

3 Term Selection Methods

The goal of term selection is to reduce the dimensionality of the indexing space without reducing classification accuracy. The removal of most indifferent terms is straightforward. The most common techniques use a *stop-list* for removal of stop terms, and *document frequency (DF)* thresholding for sparse terms. Part-of-speech tagging has been also used for removal of common function words [Rüger, 1998, Arampatzis et al., 2000]. However, after the removal of indifferent terms, a large number of non-discriminating terms still remain. Automatic term selection methods can remove more of these terms according to training data statistics.

Applying feature selection techniques to text classification tasks was found not to impair classification accuracy even for reductions up to a factor of ten. In fact, feature selection techniques slightly improve classification [Lewis, 1992, Yang and Pederson, 1997, Ragas and Koster, 1998]. Possible reasons for these improvements are — despite the fact that less information is actually used — the prevention of over-fitting a classifier into the training data, and the decrease in violations of the feature independence assumption of probabilistic models[5].

[Yang and Pederson, 1997] performed a comparative evaluation of the most popularly used feature selection methods: document frequency thresholding, expected mutual information, χ^2 statistic, term

[4] usually, a false statement.

[5] as the size of the feature set grows, the number of stochastically dependent features grows as well.

strength, and information gain. In this study, it turned out that the supposedly *ad hoc* DF thresholding presents a performance comparable to the theoretically justified and best performing schemes like χ^2 and information gain, for term removal up to 90%. The term scores of the latter three methods were found to be strongly correlated, so DF thresholding can be used instead of the others where these are computationally too expensive.

In our experimental setup we consider three approaches for term selection: relevant document frequency thresholding, uniformity in time-order, and uniformity in actual time.

RDF (Relevant Document Frequency): that is the total number of relevant documents in which the term occurs[6]. In classification tasks, learning is applied to a single pool of terms which serve to separate objects belonging to different classes. In filtering, each topic is assumed to be filtered independently of others; thus it utilizes its own pool of terms. DF thresholding on term statistics of the whole stream could hurt topics with a few relevant documents by eliminating too many of their relevant terms. Therefore, DF thresholding should be applied individually for each topic with respect to the size of its relevant training data. Consequently, the RDF approach is more suitable than DF in filtering contexts.

$U_{time\text{-}order}$ (Uniformity in time-order): that is the term occurrence uniformity in the order (rather than actual time) of arrival of relevant objects. Note that in this case the time-line is discrete, since relevant objects are seen as arriving with a constant arrival rate (objects per time unit).

$U_{time\text{-}stamp}$ (Uniformity of term time-stamps): that is the term occurrence uniformity in the continuous time-line of the stream of relevant objects. Each term is associated with a list of time-stamps, the actual times of arrival of relevant objects it occurs in. In principle, relevant objects can arrive as temporally close to each other as physically possible, therefore the *RDF* of terms in a certain time period is not bounded.

The difference between considering time-order or time-stamps will become clearer in section 4.

We have compared the above techniques in a batch filtering framework. The early part of the stream was considered as training data from which profiles were constructed. For each topic, all potential terms (the ones which occurred in relevant documents) were ranked by each of the above techniques. Term selection was performed by selecting the top k fraction of the rank for every technique. Different values of k were investigated from $k = 1$ (no term selection) down to $k = 0.01$ (99% of all topic terms were eliminated). Because for aggressive cut-offs (small k's) it is possible that some relevant training documents become empty, we applied term selection only down to the lowest k which resulted in non-empty relevant training documents.

4 A Measure for Term Occurrence Uniformity

Let us consider a *normalized training data time-line* $S = [0, 1]$, where the bootstrapping of a filtering task is located at 0 and the present time is at 1. Each term occurrence can now be represented by a point in that interval, and the occurrence pattern of a term which occurs in f objects by a list of points $x_1, \ldots, x_f$. Measures of (non)uniformity of point-lists are called *discrepancies* [Kuipers and Niederreiter, 1974]. Such measures have the structure of statistics to measure the overall difference between an estimated probability distribution and a conjectured probability distribution.

A list of f occurrence points can be converted to an unbiased estimator $S_f(x)$ of the *cumulative* distribution function of the probability distribution function from which it was drawn: $S_f(x)$ is the

[6] in the fashion of DF: the total number of documents a term occurs in.

function giving the fraction of occurrences to the left of x. The cumulative distribution function of the *uniform distribution* is $P_U(x) = x$. Different lists of points have different cumulative distribution function estimates. However, all cumulative distributions agree for $x = 0$ and $x = 1$ where they are zero and one respectively. As a consequence, it is the behaviour between 0 and 1 of their cumulative distribution functions that distinguishes distributions.

There are many statistics to measure the overall difference between two cumulative distributions. We have chosen a variant of the generally accepted *Kolmogorov-Smirnov* (K-S) test, namely *Kuipers'* statistic, which is the sum of the maximum distance of $S_f(x)$ above and below $P_U(x)$ (figure 1):

$$V = D_+ + D_- = \max_{0<x<1}[S_f(x) - P_U(x)] + \max_{0<x<1}[P_U(x) - S_f(x)] \tag{1}$$

This statistic guarantees equal sensitivities at all values of x, in contrast to the original K-S test which

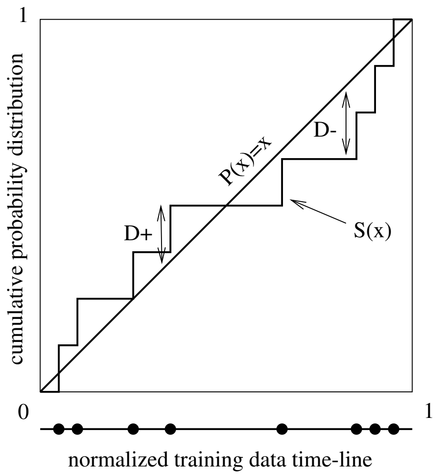

Figure 1: Kuipers' statistic

tends to be more sensitive around the median value where $P_U(x) = 0.5$ and less sensitive where $P_U(x)$ is near 0 or 1. It is also invariant under re-parameterizations of x and shifts on the circle created by glueing points zero and one of the time-line. K-S-like statistics have a computational complexity linear to f. More details on how to compute them can be found in [Press et al., 1992].

Based on equation 1, a term occurrence uniformity measure is defined as:

$$U = 1 - V = 1 - \left(\max_{0<x<1}[S_f(x) - x] + \max_{0<x<1}[x - S_f(x)]\right). \tag{2}$$

U takes values in [0, 1) and the largest the U, the most uniformly a term occurs.

Equation 2 and has the following properties:

1. $f = 1 \Rightarrow U = 0$ (easily deduced from figure 1), and

2. $\lim_{f\to\infty} U = 1$ (generally not true for all distributions of x_i's, e.g. $x_i = c, \forall i$, but provable in our context since there is always some distance between consecutive x_i's).

The first property will conveniently score at the bottom of the rank, terms which occur only once (too sparse). The latter property implies that the expected value of U is lower for $f = m$ than for $f = m + 1$

for all m. This effect may be desirable for small f's (e.g. $f < 5$) in order to devalue more sparse terms, but it also indicates a certain bias of U to f.

Figure 2 gives values of U for 10,000 randomly generated term occurrence patterns with up to 200 occurrences. Note that by putting term occurrences randomly in a time-line, the resulting occurrence pattern should have high term occurrence uniformity (close to 1 for large numbers of occurrence, because of property 2). The left plot corresponds to the case of taking into account the time-order of

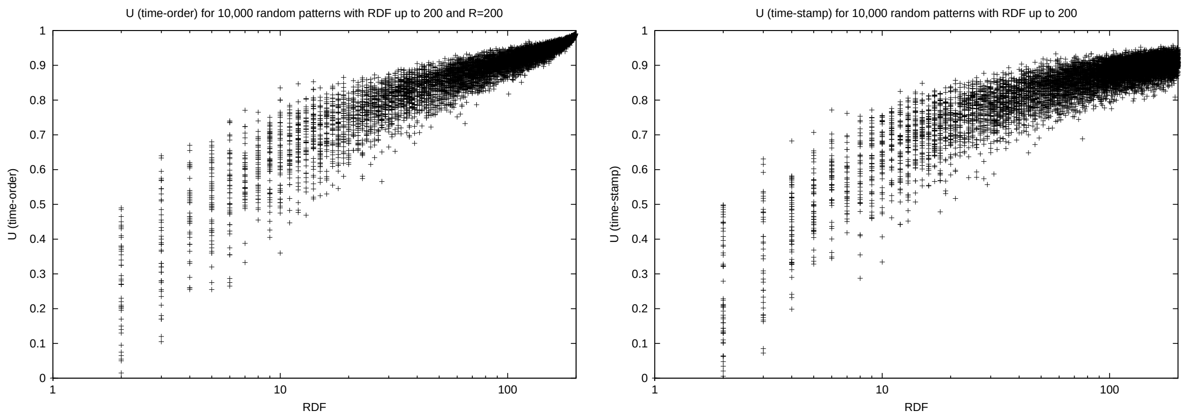

Figure 2: Correlations between RDF and U for random (uniform) input

relevant objects. In this case, the time-line is considered as being discrete, since R relevant objects will be seen as arriving at points $i/R, i = 1, \ldots, R$. Consequently, terms can occur only at these points. Obviously, the correlation between *RDF* and U becomes stronger as *RDF* gets close to R (the spread of crosses in the figure becomes thinner).

The right plot corresponds to the case of considering the actual time of arrival (i.e. time-stamp) of relevant objects. In principle, the time-line is now continuous, since relevant objects can arrive as temporally close to each other as one may think. The *rate* at which the correlation between *RDF* and U becomes stronger (as *RDF* tends to R) is now lower than the discrete case (the spread of crosses does not become considerably thinner for large *RDF*'s, but becomes asymptotically thin only when $RDF \rightarrow \infty$). In practice, the arrival rate of objects is always bounded due to processing power and network speed limitations. Thus, this correlation will be somewhat stronger.

In general, it could be proved that any term occurrence uniformity measure is correlated in some way to relevant document frequency, and the correlation becomes stronger as relevant document frequency becomes larger. Especially $U_{time\text{-}order}$ tends to produce the same rank of terms as *RDF*, for $RDF/R \rightarrow 1$ (this usually happens at the top of the rank, but not necesarily). Therefore, *RDF* and $U_{time\text{-}order}$ are expected to result in comparable effectiveness at aggressive cut-offs, something that is not guaranteed for $U_{time\text{-}stamp}$. In any case, the terms we are looking for, according to the temporal locality hypothesis, occur with frequencies just above the sparsity level and much less than R where the correlation is expected to be low.

5 Experimental Setup

Initially, we performed a pilot experiment using the LCS classification system developed in the Esprit DoRo project [Ragas and Koster, 1998]. However, we found it inefficient to simulate a filtering situation (i.e. binary classification) with LCS, since we had to run the system for every topic individually.

Therefore, we decided to run the final experiments reported in this paper with the same filtering system we used in [Arampatzis et al., 2000]. This system filters all topics in one pass over the collection. In this section we briefly describe the algorithms used, evaluation measures, the dataset, and pre-processing.

5.1 Filtering System

Our experimental system is based on the vector space model where documents and profiles are represented by vectors of weights [Salton, 1975]. Each weight corresponds to an indexing term and denotes its importance within a document or a profile. Indexing terms may be words, phrases, n-grams, or other linguistic entities, and are assumed to be stochastically independent. For these experiments we used single-word terms.

Words in documents were weighted in a *tf* $\times$ *idf* fashion, in specific, by the Cornell *ltc* variant commonly used in text retrieval [Buckley et al., 1994]. *ltc* has also found to perform better than other weighting schemes (e.g. atc, lnc, bnn) in categorization tasks on the Reuters data, and in topic detection on data from Reuters and CNN [Yang and Pederson, 1997, Yang et al., 1998]. We have also tried binary weighting, and tf-thresholding before weighting, but all these resulted in worse performance than *ltc* in the Reuters dataset. If d $(= \langle d_1, d_2, \ldots, d_n \rangle)$ a document, the *ltc* formula weights the ith term as:

$$d_i = \frac{(\log(f_i) + 1) \times \log(N/n_i)}{||d||}, \tag{3}$$

where N is the total number of documents, n_i is the total number of documents in which the ith term occurs, and f_i is the number of occurrences of the ith term in d. The denominator $||d||$ is the geometrical length of vector d, that is $||d|| = \sqrt{\sum_{i=1}^{n} d_i^2}$. The quantities N and n_i were estimated on the training data.

Term weights for profiles were calculated by the Rocchio relevance feedback method [Rocchio, 1971]. Rocchio was developed in the vector space model and classifiers based on it have proven to be quite effective in filtering and classification tasks [Ittner et al., 1995, Schapire et al., 1998, Ragas and Koster, 1998]. Given a set of documents to be ranked for a topic, an ideal classifier should rank all relevant documents above the non-relevant ones. Such an ideal classifier might just not exist. Therefore, Rocchio settles for a classifier that maximizes the difference between the average score of relevant and the average score of non-relevant documents. Rocchio specifies that the *optimal* classifier $p = \langle p_1, p_2, \ldots, p_n \rangle$ should have the ith term weighted as:

$$p_i = \frac{1}{R} \sum_{d:d \in \text{relevant}} d_i - \frac{1}{N} \sum_{d:d \notin \text{relevant}} d_i, \tag{4}$$

where R and N are the total numbers of relevant and non-relevant documents respectively. Its a common practice to set all negative Rocchio weights to zero.

The similarity between a profile and a document is computed by the dot product of their weighted vectors. Filtering in the vector space model can be done by thresholding the similarity between documents and profiles. Threshold selection is an issue which should get special attention by itself. In order to abstract away from the threshold selection problem we allow the system to return a traditional *ranked list* of documents for every profile: most relevant first, least relevant last.

5.2 Evaluation Measure

Evaluation is done on the ranked list using 11-point interpolated average precision. First, the recall and precision are calculated at every rank of the list. If any relevant documents score zero, they are ranked at the bottom of the list *below* all non-relevant which score zero. Then, the pairs of recall-precision are interpolated at 11 standard recall levels $R_s = s, s = 0, 0.1, \ldots, 1$, using the interpolation method described in [van Rijsbergen, 1990]. According to this method, a set of recall-precision pairs $G = \{(R, P)\}$ is interpolated as:

$$P_s = \{\max P : R' \geq R_s \text{ and } (R', P) \in G\},$$

where P_s is the precision at the standard recall level R_s. This interpolation method estimates at R_s the best possible precision achieved by the system. Average precision is calculated as $\frac{1}{11}\sum_s P_s$.

5.3 Dataset

The Reuters-21578 (distr. 1.0) text categorization test collection is a resource freely available for research in Information Retrieval, Machine Learning, and other corpus-based research[7].

We re-produced the *ModApte* split, which consists of documents about economic topics, such as *income*, and *gold*. Note that documents can be relevant to more than one topic. We used only the topics which have at least 100 relevant training documents (16 topics in total). All training documents which did not belong to any of these topics were screened out. We did not remove any of the test documents. Table 2 gives some statistics of the resulting dataset.

topic	**trn. docs.**	**test docs.**	$U(topic)$
earn	2861	1087	0.813
acq	1648	719	0.827
money-fx	534	179	0.769
grain	428	149	0.881
crude	385	189	0.794
trade	367	117	0.739
interest	345	131	0.794
wheat	211	71	0.866
ship	191	89	0.859
corn	180	56	0.831
money-supply	132	34	0.815
dlr	131	44	*0.581*
sugar	125	36	0.815
oilseed	124	47	0.824
coffee	111	28	0.806
gnp	101	35	0.771
total	6909	3299	

training stream time-span: 40.42 days (from 1st training to 1st test document)
total unique words: 22,102 (after pre-processing)

Table 2: Dataset statistics

The training stream covers a period of 40.42 days, calculated from the time of arrival of the first training document of any topic to the first test document. The rightmost column gives the *topic occurrence uniformity* in the training stream. The closest this number is to one, the more constant the delivery

[7] The collection and its documentation is available from: `http://www.research.att.com/~lewis/`

rate (relevant documents per time unit) is in the training data for that topic. Obviously, documents about *dlr* (dollar) arrive in bursts, a possible consequence of temporal events concerning e.g. a dollar's exchange rate drop in Tokyo. Figure 3 gives the temporal histograms of document arrivals for topics *dlr* and *money-supply*. We can see that many of *dlr* documents arrive in the period of days 25 and 37, while

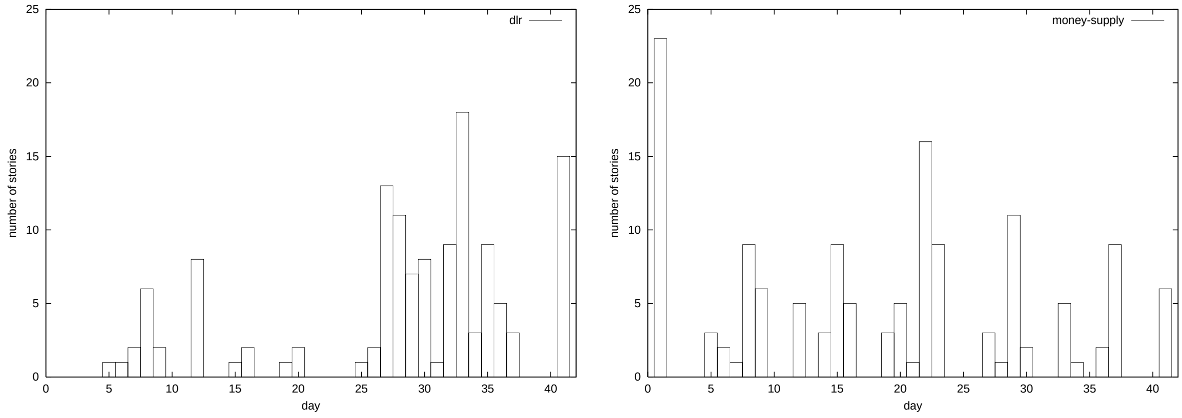

Figure 3: Temporal histograms of document arrivals for 2 topics

the document distribution in the time-line for topic *money-supply* is more uniform[8]. A comparison of their cumulative distribution functions to the cumulative distribution function of the uniform distribution is given in figure 4. In section 7, we will come back to the subject of interpreting topic uniformity in a

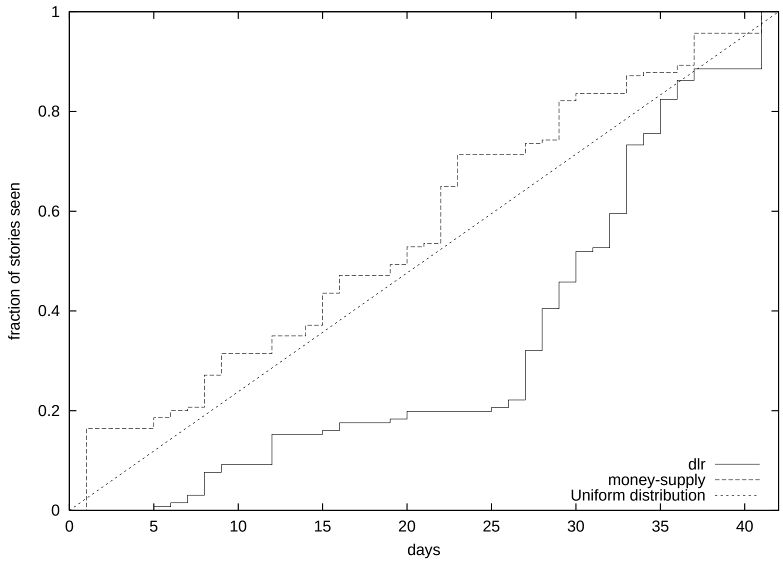

Figure 4: Cumulative distribution functions of document arrivals for 2 topics

[8]The 2-day gaps (e.g. days 10-11, 17-18, etc.) in *money-supply* correspond to weekends where no economic news is made. Note that this match is not exact since the days in the plots do not correspond to real days; they are successive 24-hour intervals taken from the *arrival time* of the first document of the training stream. Considering also the different closing times of international stock-markets, it should explain why few economic documents seem to occur in weekends (especially for *dlr*). These 2-day gaps are partly responsible for the apparently upper-bounded values of topic uniformity. In a way, this boundness propagates also to term occurrence uniformity and is visible at the right plot of figure 7, further in this article.

filtering context.

5.4 Pre-processing

The pre-processing phase was performed in four stages: tokenization, Part-Of-Speech (POS) tagging, removal of common function words, and morphological normalization of the remaining words. Tokenization consisted of detection of sentence boundaries, followed by division of sentences into words. Detection of sentence boundaries was necessary since we used a POS tagger.

Brill's rule-based tagger[9] [Brill, 1994] was employed to obtain POS information for the words of the dataset. The tagger comes with a lexicon derived from both the Penn Treebank tagging of the Wall Street Journal (WSJ) and the Brown Corpus. Conveniently, the WSJ articles are, like the Reuters documents, about economic topics and this increased the reliability in tagging the Reuters corpus. We used a *POS stop-list* to remove all common function words. In fact, we removed all words except: nouns, adjectives, verb-forms, and adverbs.

Morphological normalization of the remaining words was performed by means of *lemmatization* (which can be seen as a form of POS-directed stemming), using WORDNET's v1.6 [Miller, 1995] morphology library functions[10].

6 Experimental Results and Discussion

Figures 5 and 6 give our experimental results per topic. Topic plots appear in decreasing order of their training set size. The 11-point average precision is plotted as a function of the fraction of term selected from 1 (no term selection) down to 0.01 (99% of all terms were eliminated). In fact, for each term selection method and topic, cutoff thresholds were applied only down to the lowest point for which there were no empty training documents. All terms with single occurrences in a topic were eliminated in advance.

Our term selection results agree with previous research [Lewis, 1992, Yang and Pederson, 1997, Ragas and Koster, 1998]: most of the terms in classification environments can be eliminated without impairing classification effectiveness (as this measured by average precision); even slightly improving it for some topics. It is worth mentioning that average precision increases drastically for topics *wheat*, *sugar*, and *coffee* for aggressive thresholds. This result seemed counter-intuitive at first glance, but after further investigation it was found that these topics have words which occur in almost all of their relevant documents (unique identifiers). These words unsurprisingly are *wheat*, *sugar* and *coffee* and occur in 97%, 96%, and 100% of the relevant to the respective topic documents. Each of these words together with a few others were capable to achieve the best result for the respective topic. For these topics, average precision was maximized for profiles with 9–12 words, while a larger number of words was likely to introduce noise, rather than improve effectiveness.

In a comparison between *RDF*, $U_{time\text{-}order}$, and $U_{time\text{-}stamp}$, all methods presented a comparable performance ($<$ 5% difference in average precision) for reductions up to 70%. Actually, for most topics the methods are comparable up to 90%. At aggressive cutoffs ($>$ 90% reduction), *RDF* performs generally better that uniformity-based term selection. Nevertheless, even here the difference in average precision is less than 10% (with as an exception the topic *ship*). Thus, all methods seem to hold up

[9] Eric Brill's tagger V1.14 and a description are available by anonymous ftp from: `ftp://ftp.cs.jhu.edu/pub/brill` in the `Programs` and `Papers` directories.

[10] Specifically, we called the `morphstr()` function which tries to find the base-form (lemma) of a word or collocation, given its part-of-speech. WORDNET is created by Cognitive Science Laboratory, Princeton University, 221 Nassau St. , Princeton, NJ 08542. It is available for anonymous ftp from `clarity.Princeton.edu` and `ftp.ims.uni-Stuttgart.de`.

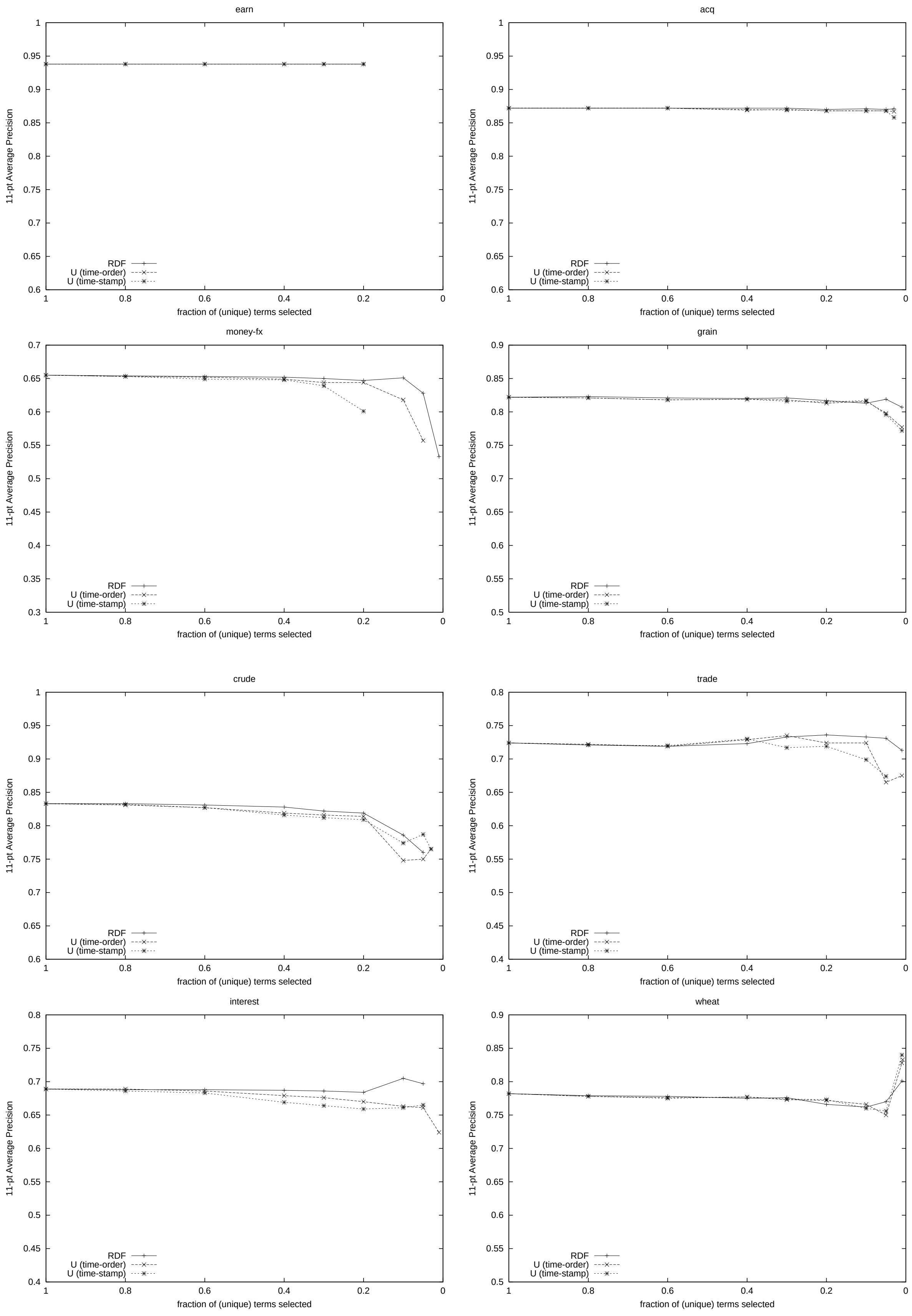

Figure 5: Average precision vs. fraction of terms selected for topics *earn* to *wheat*

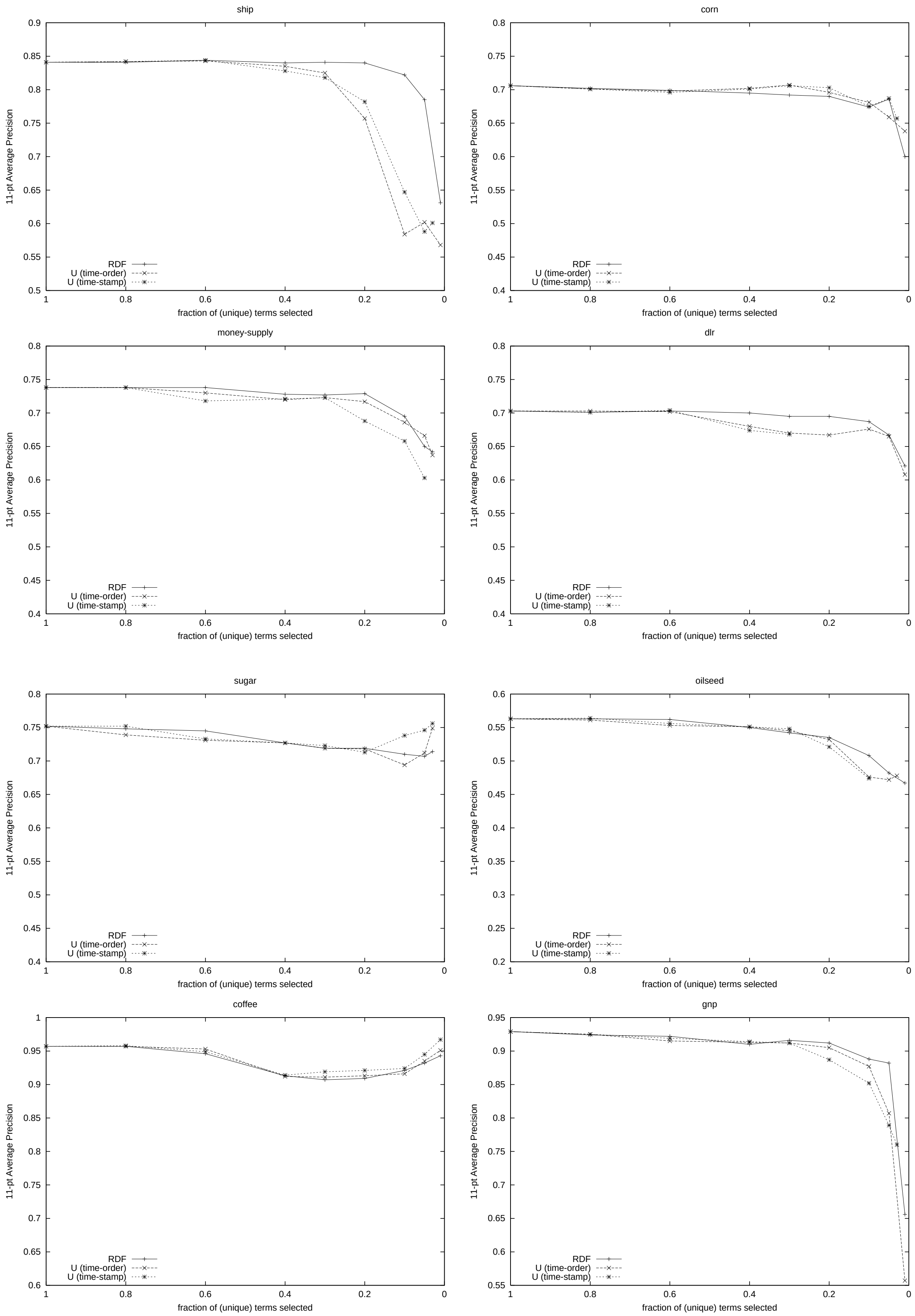

Figure 6: Average precision vs. fraction of terms selected for topics *ship* to *gnp*

comparably at aggressive cutoffs. It is important to note that for the topics with unique identifiers (*wheat*, *sugar* and *coffee*) our methods even performed better than *RDF* at aggressive cutoffs, suggesting that they selected more discriminating words to accompany the unique identifiers in profiles.

The fact that $U_{time\text{-}order}$ reached for all topics equal or more aggressive cutoffs than $U_{time\text{-}stamp}$ (meaning that it does not result in empty relevant documents) is a consequence of its stronger correlation to *RDF*. The correlations of $U_{time\text{-}order}$ and $U_{time\text{-}stamp}$ to *RDF* for Reuters are given in figure 7 (we normalized *RDF* as RDF/R per topic, so the plots make more sense when $U_{time\text{-}order}$ values for all topics are plotted together). The obvious upper bound of $U_{time\text{-}stamp}$ values is partly a consequence

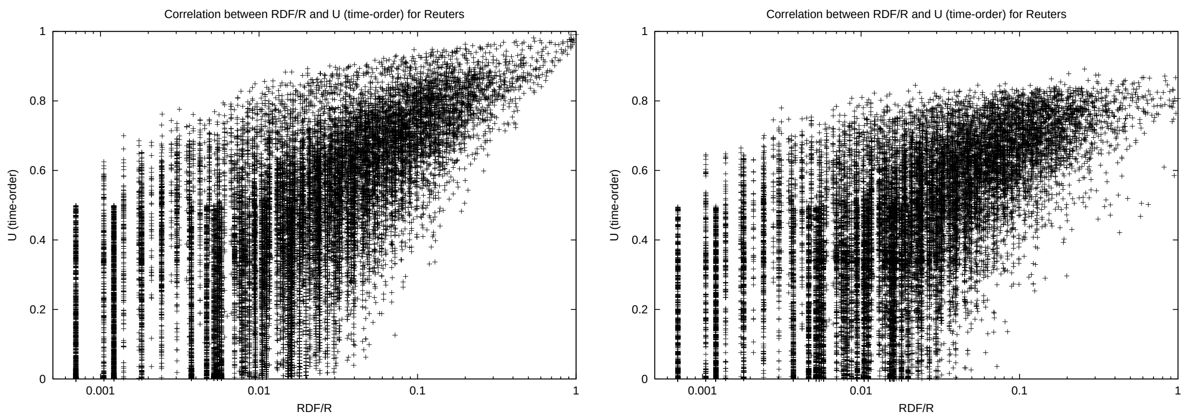

Figure 7: Correlations between RDF and U in Reuters

of the lack of documents arriving in weekends, as we mentioned earlier. It is also because of the continuous time-line considered, a consideration which produces in general lower values than $U_{time\text{-}order}$. Although both TOU methods present a strong correlation to RDF/R as the latter tends to one, these correlations rather diminish for frequency characteristics with which most of the terms occur, e.g. for $RDF/R < 0.01$. This observation suggests that our methods are indeed novel, since they threw away quite different sets of terms than the *RDF* method, for reductions up to 70%. Nevertheless, the fact that we saw no improvements in performance at these thresholds implies that we were looking for local events where their recognition is not of great importance for classification, e.g. in the Reuters collection.

On one hand, the fact that our TOU term selection methods showed a performance comparable to *RDF* for reductions up to 90% appears promising, since document frequency thresholding is known to be a powerful method for term selection. On the other hand, if most of the terms are to be thrown away, what matters most for a term selection method is to achieve high accuracy at very aggressive thresholds. At these thresholds, while there was no sharper decrease in effectiveness with our methods, document frequency thresholding has been more reliable.

All the above suggest that a wise integration of a TUO method and some other powerful time-disregarding term selection method may acquire the benefits of both approaches. Temporal events should be taken into account only when they exist, and their special treatment is expected to make a difference in classification. If this is not the case, the new method should turn into a time-disregarding one. We believe that $U_{time\text{-}stamp}$ is a better candidate for such an integration, since it reflects more the actual event identification. We have not yet worked out a scheme like that. At any rate, our results so far are inconclusive, so we have not been able to prove or disprove the term temporal locality hypothesis.

7 Conclusions and Directions for Further Research

We have in this article taken up the challenge by David Hull [Hull, 1998] and investigated the use of time distributions in retrieval environments with temporally-dependent data. The hypothesis of temporal locality of terms was neither proved nor disproved, since our results were inconclusive. Nevertheless, we have introduced *Term Occurrence Uniformity (TOU)* as a novel term selection method with a performance comparable to document frequency thresholding. We regard this result as promising, since document frequency thresholding is known to be more than just an *ad hoc* approach for term selection, and quite powerful in text categorization environments [Yang and Pederson, 1997]. The subject indeed merits deeper theoretical and empirical investigation.

To keep the ball rolling, we will point out what we believe has gone wrong. First, the Reuters-21578 collection is improper for this kind of research. The training period is short, covering slightly over 40 days, which gives little scope for temporally local events and non-uniformity. We will have to repeat the test with material collected over longer periods of time.

Second, the distribution of a topic in time can provide useful information about its nature. For instance, *terrorism* is an *event-driven* topic, in the sense that documents about terrorism occur mostly when a related event happens, e.g. a NYC subway bombing. Compare this to the topic *football* which is rather an *event-irrespective* interest. Football developments are reported on a regular basis, irrespective from whether something really important has happened. These observations suggest the need for a better modeling of the nature of topics and streams. Our best try so far has considered distributions of relevant terms in time, which refers only indirectly to the distribution of the topic. Topic occurrence uniformity may encapsulate useful information and should be taken into account.

Third, while our intention is to develop temporally-dependent term selection and term weighting schemes for filtering, we have tested our approach in a rather static situation, namely batch filtering, with clearly defined training and test phases. A real-world filtering task is usually an adaptive process. Adaptive filtering is a much more difficult and especially sensitive task. Therefore, the application of such temporally-dependent term selection and term weighting schemes in adaptive filtering is expected to show larger variations in effectiveness.

At any rate, the issue of using time distributions in retrieval tasks is not settled and we will report our progress in forthcoming publications.

Acknowledgments

The main author would like to thank André van Hameren for fruitful discussions and proof-reading, and Jean Beney for running a preliminary helpful experiment.

References

[Allan, 1996] Allan, J. (1996). Incremental Relevance Feedback for Information Filtering. In Frei, H.-P., Harman, D., Schauble, P., and Wilkinson, R., editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 270–278, Zurich, Switzerland. ACM Press.

[Allan et al., 1998] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic Detection and Tracking Pilot Study — Final Report. In *Proceedings of the Broadcast News Transcription and Understranding Workshop*.

[Arampatzis et al., 2000] Arampatzis, A. T., van der Weide, T. P., Koster, C. H. A., and van Bommel, P. (2000). An Evaluation of Linguistically-motivated Indexing Schemes. In *Proceedings of the BCS-IRSG'2000*. To appear.

[Belkin and Croft, 1992] Belkin, N. J. and Croft, W. B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, 35(12):29–38.

[Brill, 1994] Brill, E. (1994). Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Wa.

[Brodley and Rissland, 1993] Brodley, C. E. and Rissland, E. L. (1993). Measuring Concept Change. In *AAAI Spring Symposium: Training Issues in Incremental Learning*, pages 98–107.

[Buckley et al., 1994] Buckley, C., Salton, G., and Allan, J. (1994). The Effect of Adding Relevance Information in a Relevance Feedback Environment. In Croft, W. B. and van Rijsbergen, C. J., editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, Dublin, Ireland. ACM Press.

[Hull, 1998] Hull, D. A. (1998). The TREC-7 Filtering Track: Description and Analysis. In Voorhees, E. M. and Harman, D. K., editors, *The Seventh Text REtrieval Conference (TREC-7)*, pages 33–46, Gaithersburg, Maryland. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-242.

[Ittner et al., 1995] Ittner, D. J., Lewis, D. D., and Ahn, D. D. (1995). Text Categorization of Low Quality Images. In *Symposium on Document Analysis and Information Retrieval*, pages 301–315, Las Vegas, NV. ISRI; University of Nevada.

[Kuipers and Niederreiter, 1974] Kuipers, L. and Niederreiter, H. (1974). *Uniform Distribution of Sequences*. Wiley, New York.

[Lewis, 1992] Lewis, D. D. (1992). Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of Speech and Natural Language workshop, Harriman, New York, February 23–26*, pages 212–217, San Mateo, CA. Morgan Kaufmann.

[Miller, 1995] Miller, G. A. (1995). WORDNET: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

[Press et al., 1992] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing, 2nd ed.* Cambridge University Press, Cambridge, UK.

[Ragas and Koster, 1998] Ragas, H. and Koster, C. H. A. (1998). Four Text Classification Algorithms Compared on a Dutch Corpus. In Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 369–370, Melbourne, Australia. ACM Press, New York.

[Rocchio, 1971] Rocchio, J. J. (1971). Relevance Feedback in Information Retrieval. In *The SMART Retrieval System — Experiments in Automatic Document Processing*, pages 313–323, Englewood Cliffs, NJ. Prentice Hall, Inc.

[Rüger, 1998] Rüger, S. M. (1998). Feature Reduction for Information Retrieval. In Voorhees, E. M. and Harman, D. K., editors, *The Seventh Text REtrieval Conference (TREC-7)*, pages 409–412, Gaithersburg, Maryland. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-242.

[Salton, 1975] Salton, G. (1975). A Vector Space Model for Information Retrieval. *Communications of the ACM*, 18(11):613–620.

[Schapire et al., 1998] Schapire, R. E., Singer, Y., and Singhal, A. (1998). Boosting and Rocchio Applied to Text Filtering. In Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 215–223, Melbourne, Australia. ACM Press, New York.

[van Rijsbergen, 1990] van Rijsbergen, C. J. (1990). *Information Retrieval*. Butterworths, London, United Kingdom.

[Yang and Pederson, 1997] Yang, Y. and Pederson, J. (1997). A Comparative Study on Feature Selection in Text Categorization. In Engels, R., Evans, B., Herrmann, J., and Verdenius, F., editors, *Proceedings of the Fourteenth International Conference on Machine Learning '97 (ICML 97)*, Vanderbilt University, Nashville, TN.

[Yang et al., 1998] Yang, Y., Pierce, T., and Carbonell, J. (1998). A Study on Retrospective and Online Event Detection. In Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36, Melbourne, Australia. ACM Press, New York.