Document Filtering as an Adaptive and Temporally-dependent Process

Avi Arampatzis Th.P. van der Weide

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.

avgerino@cs.kun.nl
http://www.cs.kun.nl/~avgerino

Abstract

The filtering task has traditionally been defined as a special case of the information retrieval task, and undeniably, it can be performed by applying retrieval techniques. This theoretical study summarizes our experiences in viewing filtering as an adaptive and temporally-dependent process. A process that, in contrast to traditional retrieval, takes into account the dynamic nature of relevance and its temporal aspects. We investigate the nature of user interests, formulate useful types of adaptivity, and discuss the effectiveness of those types in relation to user interests. To deal with drifts, we introduce the notion of the half life of documents. Furthermore, we discuss potential dangers for effectiveness such as selectivity traps. We pay special attention to practical efficiency issues by discussing term selection and incrementality.

1 Introduction

The digital and networking revolution over the last decade has made large amounts of digital information available. This tremendous increase in digital information has led to a new challenge in *information seeking*. Currently, users everyday find themselves confronted with large amounts of information in the form of news, e-mail messages, and especially World-Wide Web pages. Although users have access to a rich body of information, only a small fraction of this is actually relevant to the interests of any particular user.

Information retrieval, and especially text retrieval, is an information seeking process with an extensive research heritage. Given the shared similarities between many information seeking processes, the filtering task has been seen as a special retrieval case, treated by retrieval techniques. In some cases, the filtering and retrieval tasks have even been seen as "two sides of the same coin" [1]. We do not question the similarity of the tasks; the filtering task can indeed be performed with slightly modified retrieval techniques. However, we point out a few important differences in the nature of data involved. Taking these differences into account is beneficial for effectiveness.

This article is influenced by the work of several researchers. We have found especially useful the conceptual framework for filtering described in [2], and the adaptivity issues discussed in [3]. We additionally refresh and revise the most important parts of the work described in [4] and [5]. In the following Sections, document filtering systems are addressed by:

- classifying user interests with respect to how the idea of relevance changes over time (Section 3). As we will see, relevance may be disturbed by user-triggered and world-triggered factors.

- classifying user interests with respect to the occurrence patterns of relevant documents in time (Section 4). We introduce a measure which enables the temporal classification of interests. Moreover, we outline how such information may be used in filtering.
- classifying forms of adaptivity (Section 5).
- discussing implementation issues, such as incrementality (Section 6).
- discussing the performance of different forms of adaptivity on different kinds of user interests (Section 7).
- discussing term selection for adaptive filtering tasks (Section 8).
- discussing potential dangers for effectiveness, such as selectivity traps (Section 9).

This study is the result of the bottom-up approach we have followed to deal with filtering in the last years. Guided by the experiments we have performed — in the context of the TREC-9 adaptive filtering tasks and elsewhere — we will try to formulate what we believe lies on the top and is important for effectiveness. For completeness, we will start from the definition of the filtering task.

2 Document Filtering

Document filtering is an information seeking process that searches through a dynamically generated document collection, e.g. a *stream* of arriving documents, for documents which match a user interest. The user interest is assumed to be long-term, in contrast to one-time queries in retrieval, and we will call it a *topic*. Filtering may also be seen as a *binary classification/categorization* task where each new document has to be classified under one of two categories: relevant, or non-relevant.

Document filtering, and similarly other information seeking processes, can be broken down into three sequentially-performed sub-tasks or modules: *collection*, *selection*, and *display* of documents. The overall picture is depicted in Figure 1.

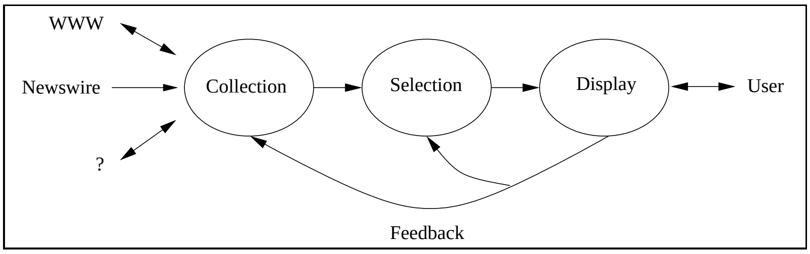


Figure 1: Sub-tasks of a filtering process.

Collection is concerned with providing a document stream. Two ways of collecting documents may be distinguished: *passively* collected e.g. from a newswire [6], or *pro-actively* collected e.g. with autonomous intelligent agents going out to find new documents in the World-Wide Web [7]. The combination of both actively and passively collected documents in one stream is also possible. The display module is responsible for the interaction between users and the system. It does not only display the selected documents, but it interacts with users and accounts for their reactions on the presented output to guide the collection and selection processes. In this study, we focus our attention on the selection module. The collection and display tasks are already rich research areas in their own right, and will be considered here as black-boxes, where the former provides a document stream and the latter provides relevance judgments for some of the selected documents.

The selection module does the actual filtering of the collected documents, selecting the relevant or rejecting the irrelevant ones, with respect to a topic. It uses some internal *representation* for documents and topics, called *profiles*. Representation allows, by means of a *filtering function*, the calculation of the aboutness of each document with respect to a topic, so as to decide whether to select or reject it. Two sources of deducing representations have been dominating the research in filtering, distinguishing two types of filtering: *collaborative* (or *social*), and *content-based* (or *cognitive*) [6, 8].

In collaborative filtering, documents are represented by annotations made by their prior readers. By exchanging these annotations, groups of users with shared interests can automatically be identified. Collaborative filtering can provide a basis for selection of documents regardless of whether or not their content is represented. Content-based filtering assumes that each user operates *independently*. There is no exchange of information of any kind, thus document representations can only be derived from their content. Of course, both approaches may be combined in a way that annotations and content both contribute to estimate the aboutness. In this study, we are concerned with content-based filtering.

Filtering systems can exploit the long-term nature of topics to improve the filtering model over time. A system may continuously monitor the stream accumulating different kinds of statistical data, and using them to produce better representations for profiles. Moreover, as documents are filtered for a topic, the user may give relevance judgments for some of the selected documents. Judged documents can be used to adapt the topic profile and the filtering function. The choice between exploiting the long-term nature of topics or not, distinguishes between two types of systems. Systems that do not change the way they filter over time are called *batch*[1] or *non-adaptive*. One-pass filtering systems that alter their filtering model in response to the history are called *adaptive*. In TREC-9, adaptivity has been proven important: the effectiveness of adaptive runs (initiated only with very little relevance information) has been comparable to this of batch runs (initiated with full training sets).

3 A Relevance Classification of Topics

A filtering task begins with a user interest and a stream of documents. With respect to a stream of N documents, and assuming binary relevance, we will define as topic T the substream of all documents relevant to the user's interest, e.g., $T = D_1, \ldots, D_n,\ n \leq N$. This definition of topic quantifies the user interest in terms of the document stream. We will assume that the topic is *persisting* in the stream, that is, as the stream grows ($N \to \infty$) the topic grows as well ($n \to \infty$).

Adopting this point of view, only 2^N different topics may be distinguished for a certain N, however, an infinite number of interests may be thought of. When two or more different interests translate to the same substream of relevant documents, we will not distinguish between those interests; the idea is that you cannot get anything more than what is actually present in the stream.

Let us assume an *abstract distance measure* $d(D_i, D_j) \in [0, +\infty)$ between any two documents D_i, D_j. Small distance values mean that two documents are about similar *subjects*. We will also introduce a *fuzziness parameter* ε which denotes the maximum distance allowed for two documents to be considered as being about the same subject.

A topic T may be classified with respect to the values of the distance $d(D_i, D_j)$, for all relevant documents D_i, D_j, as $n \to \infty$:

- **stable:** All distances between the documents are less than or equal to ε:

$$\forall i, j: \;\; d(D_i, D_j) \leq \varepsilon \; .$$

- **drifting:** All distances between consecutive documents are less than or equal to ε, but some distances of non-consecutive documents are not:

$$\forall i: \;\; d(D_i, D_{i+1}) \leq \varepsilon \quad \text{and} \quad \exists i, j : d(D_i, D_j) > \varepsilon \; .$$

[1] Adhering to the TREC convention.

- **multimodal:** There are consecutive document distances greater than ε, but the topic can be broken down to a finite number of k stable disjoint subtopics:

$$\exists \text{ stable } T_1, \ldots, T_k : T = \oplus_i T_i \quad \text{and} \quad \forall D_i \in T_l, D_j \in T_m, l \neq m : \; d(D_i, D_j) > \varepsilon \; ,$$

where $\oplus$ means that T_i's are *exclusive partitions* of T: they have no documents in common but their union amounts to T.

- **vagrant:** the same as multimodal, but the number k of subtopics is infinite.

- **white noise:** the same as vagrant, but $k \to \infty$ faster than for vagrant topics.

This classification is rough, but sufficient for our analysis. A topic may exhibit in reality a more complex behaviour in time by switching between two or more of the above types. For example, a topic is at first stable, but then starts drifting; or even a subtopic T_i of a vagrant topic is drifting.

Note that the fuzziness parameter ε determines the limits of the classes: a very large fuzziness will classify all topics as stable, while an infinitesimal one will classify everything as white noise. However, for a given reasonable fuzziness, what classifies a topic under one of the above categories depends on *user-triggered* and *world-triggered* factors.

User-triggered factors are related to whether a user interest shifts in time, and how it shifts. World-triggered factors are independent of shifts in user's interest. They are directly related to the nature of the interest with respect to the real world. The world may produce considerably different but still relevant documents.

3.1 User-triggered Shifts in Interest

A user who sticks to her initial request has a *stable interest*. However, the user interest can also deviate over time. For instance, as the user reads more and more documents about the initial request, she wants to know more specific or general information, or slowly becomes interested in a similar subject which is referred to in the documents already retrieved. In this case, the user has an *drifting interest*. [9] has demonstrated that such drifts can be handled readily by phasing out old context.

A *multimodal* or *vagrant interest* usually arises when the user does not exactly know what she is looking for, consequently the interest is vaguely formulated. She will probably find different kinds of documents relevant, in the search of her real interest. The interest may switch between closely related — specific or relatively random — domains.

We will assume here a rational user who does not abruptly change her mind. An abrupt shift should be considered as a different interest and be treated as a new filtering process. Thus, white noise behaviour can not arise for user interests in filtering; it rather corresponds to user interests in traditional retrieval tasks.

3.2 World-triggered Shifts in Document Content

Consider filtering an interest about *HIV treatments*. Over the years, treatments have changed; new and more effective ones have been slowly developed, while the less effective ones have been fading out. In such cases, where the contents of relevant documents slowly change in time, there is *content drift*.

Document contents can show multimodal or vagrant characteristics. Multimodality arises when the interest is such that it combines two or more stable but relatively distant interests, for example, *operating systems* AND *computer architecture*. The contents of relevant documents will switch between the two different subjects at irregular intervals.

A special kind of vagrancy arises in what we call *event-driven interests*. As example consider the interest *terrorism*. Such an interest is driven by real world *events* which can be relatively different and unexpected, for example, *NYC subway bombing* or *flight TR-304 hijacking*. An important event is usually associated with bursts of relevant documents for some period of time. Then, documents about the subject tend to disappear completely from a news stream, while some other (relatively random) terrorist event may happen.

3.3 Relevance

User interest shifts and *document content shifts* are related in the sense that the idea of relevance changes. Whether a shift comes from the user or the world side is not of importance. What is important is that future relevant documents will be different than the ones of the past. Consequently, we will talk about *relevance shifts*, irrespective of who or what causes them.

The user and the world can both be viewed as sources of introducing disturbances in relevance. In this respect, the source with the highest *entropy* defines the class of the topic. For example, a stable user interest but vagrant contents in relevant documents results in a vagrant topic.

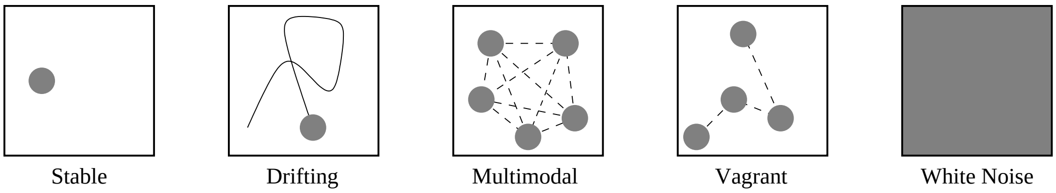


Figure 2: A relevance classification of topics.

In summary, an interest is what a user has in mind. An interest may be satisfied by a number of (finite or infinite) subjects. A document contains a few subjects. The same subject may spread across documents. A topic is the substream of all documents containing subjects that satisfy the interest at the time of their arrival. With respect to how relevance changes, i.e., how the contents of relevant documents change in time, topics may be classified as shown in Figure 2. The Figure shows possible trajectories of relevance in the document space.

The classification of topics we have just considered is related to the types of adaptivity we will introduce in Section 5. In Section 7 we will discuss this relationship. First we will attempt another classification of topics.

4 A Temporal Classification of Topics

The classification of topics considered in the previous Section is purely based on relevance aspects. We have considered how relevance changes in the ordered set (stream) of a topic's relevant documents. In this Section, we consider the actual times of arrival of relevant documents.

The qualitative classification we consider has the following classes:

- **simply periodic:** Single relevant documents arrive at approximately constant time intervals.

- **random** or **uniform:** Relevant documents arrive at irregular intervals.

- **periodically clustered:** Some relevant documents arrive at regular time intervals.

- **aperiodically clustered:** Bursts of relevant documents arrive at irregular time intervals.

Figure 3 depicts the above occurrence patterns. This classification of topics is rather orthogonal to the relevance classification considered in Section 3.

Next we will see how *uniformity* may be quantified. The measure we will introduce enables the temporal classification of topics as discussed above. Then we will briefly discuss the implications that such a classification has for filtering effectiveness.

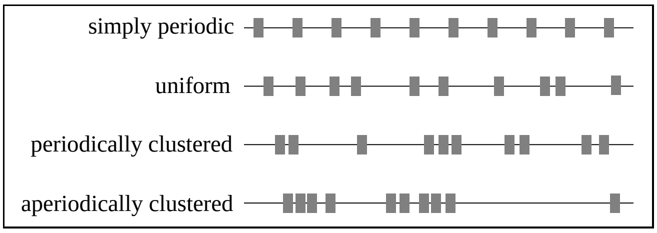


Figure 3: A temporal classification of topics.

4.1 A Measure for Uniformity

Let us consider a *normalized time-line* $[0, 1]$, where the initiation of a filtering task is located at 0 and the present time is at 1. Each document occurrence can now be represented by a point in that interval, and the occurrence pattern of a topic of length n by a list of points $x_1, \ldots, x_n$. Measures of (non)uniformity of point-lists are called *discrepancies*. 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 n occurrence points can be converted to an unbiased estimator $S_n(x)$ of the *cumulative* distribution function of the probability distribution function from which it was drawn: $S_n(x)$ is the 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 [10], which is the sum of the maximum distances of $S_n(x)$ above and below $P_U(x)$:

$$V_n = D_+ + D_- = \max_{0<x<1}[S_n(x) - P_U(x)] + \max_{0<x<1}[P_U(x) - S_n(x)] \; . \tag{1}$$

The method is demonstrated in Figure 4a.

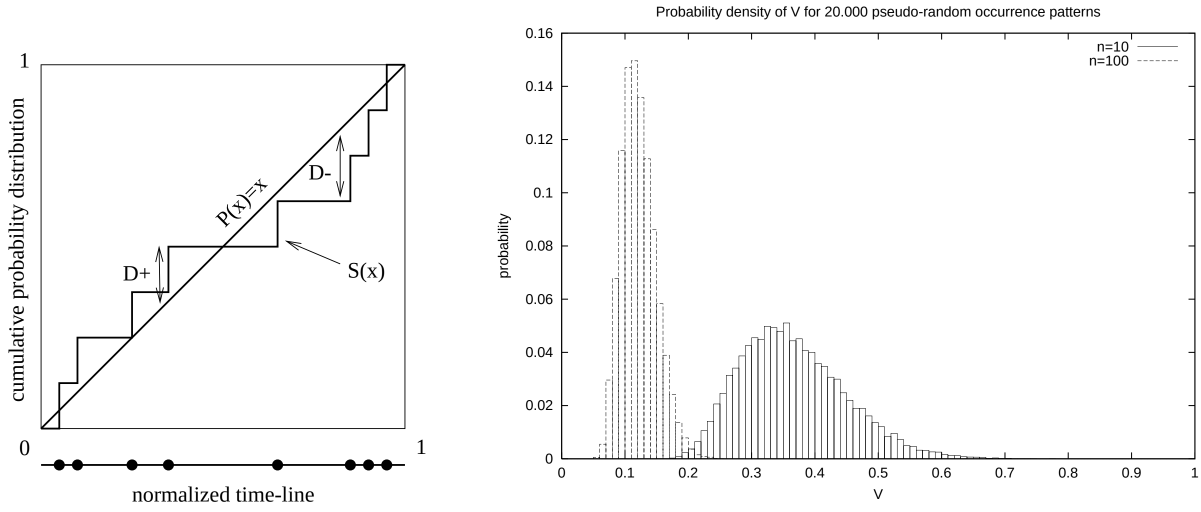
This statistic guarantees equal sensitivities at all values of x, in contrast to the original K-S test which tends to be more sensitive around the median value where $P_U(x) = 0.5$ and less sensitive where $P_U(x)$ is near 0 or 1. It is also invariant under re-parameterizations of x and shifts on the circle created by gluing points zero and one of the time-line. K-S-like statistics have a computational complexity linear to n. More details on how to compute them can be found in [11].

V_n takes values in $[\frac{1}{n}, 1]$. Figure 4b shows the empirical probability densities of V_{10} and V_{100} for 20.000 pseudo-random occurrence patterns. Values close to $1/n$ are obtained for simply periodic occurrences. Truly random patterns get slightly larger values; how much larger is determined by the number of occurrences n. Values of V_n close to 1 correspond to serious clustering of the occurrences in the timeline.

Using Kuiper's statistic, topics may be quantitatively classified into the classes defined at the beginning of this Section. We simply split the range of values $[\frac{1}{n}, 1]$ of V_n into four intervals. These intervals are determined by three cut-values x_1, x_2, and x_3. For a certain n, we define x_1 and x_2 through the Equations

$$P(V_n < x_1) = p \quad \text{and} \quad P(V_n > x_2) = p \; , \tag{2}$$

for some small p, e.g., $p = 1\%$. For a certain p, x_1 and x_2 may be obtained from standard tables with the confidence levels of the statistic, e.g., from [12]. Moreover, we define x_3 as some number between x_2 and 1; its exact value is a rather subjective matter and it should be justified empirically.



(a) $V_n = D_+ + D_-$.

(b) Probability density of V_n for uniform occurrence patterns.

Figure 4: Kuipers' discrepancy test V_n.

4.2 Using Temporal Information

We will outline how information about the occurrence pattern of a topic in time may be used for filtering. Let us consider again a stream of N documents and a topic T of length n. The *density* of relevant documents in the stream for T is

$$\rho = \frac{n}{N} \ . \tag{3}$$

If the topic occurs randomly in the stream, then ρ may be interpreted as the probability that the next arriving document will be relevant. However, high topic uniformity is not the case in general. Periodic and clustering characteristics introduce uncertainty into the interpretation of density as probability. The uncertainty decreases with the topic uniformity.

V_n, ρ, and periodicity information may provide means for filtering irrespective of document content. ρ can be seen as the *expected value* of the *a priori* probability of relevance $P(\text{rel})$, i.e., $\mathsf{E}(P(\text{rel})) = \rho$. The variance of the distribution of $P(\text{rel})$ in time is some increasing function f_n of V_n, i.e., $\mathsf{V}(P(\text{rel})) = f_n(V_n)$. Periodicity information may give an estimate of $P(\text{rel})$ that corresponds to a certain time-point t:

$$P(\text{rel}|t) = \mathsf{E}(P(\text{rel})) + g(\mathsf{V}(P(\text{rel})), t) \ , \tag{4}$$

where g is some function that accounts for periodicity and/or temporal clustering.

In principle, one could blindly retrieve documents by sampling the document stream with probability $P(\text{rel}|t)$. $P(\text{rel}|t)$ is usually small, since it depends on ρ which is small because there are usually many more relevant than non-relevant documents in a stream. However, depending on the temporal nature of the topic, $P(\text{rel}|t)$ may peak at usable values. In any case, $P(\text{rel}|t)$ may be seen as additional evidence that together with $P(\text{rel}|D)$ (the probability of relevance estimate based on document content) contributes to the decision of whether to select a document or not.

What we have just described is rather crude, and we do not claim that this is the best way to deal with the temporal aspects of filtering. Summarizing the problem, the questions are: How can $P(\text{rel}|t)$ be estimated for the history? How can one extrapolate $P(\text{rel}|t)$ for the future? What is the appropriate way to combine the two pieces of evidence $P(\text{rel}|t)$ and $P(\text{rel}|D)$? In fact, the tools are already there; here are a few keywords: *Fourier analysis*, *time-series analysis*, or more contemporary and geometrically, *phase space reconstruction* and *Poincaré sections*.

5 A Temporal Classification of Adaptivity

Disregarding the actual techniques used for creating or altering a filtering model, filtering systems may be classified according to the *temporal location* from which they obtain the information for doing that. To reach such a classification we will follow an approach similar to the one in [3].

Let us consider a system that is initiated at time 0 and the current time is t; thus the system has a history of length t. The importance of an *event* that happened at time x within this history can be modeled by a *history weight function* $H(x,t)$ with the following property:

$$\int_0^t H(x,t)\,dx = 1\;, \quad \forall t > 0\;, \tag{5}$$

that is, the area below the $H(x,t)$ curve amounts always to 1 for all t. For instance, a history weight function that weighs equally all history is:

$$H(x,t) = 1/t\;. \tag{6}$$

Irrespective of its form, the $H(x,t)$ curve is characterized by its mean value, which is mathematically defined as:

$$\bar{H}(t) = \int_0^t H(x,t)x\,dx\;. \tag{7}$$

It will be useful for this analysis to define the distances $a(t)$ and $b(t)$ of this mean from the beginning and the end of the history respectively:

$$a(t) = \bar{H}(t)\;, \quad b(t) = t - \bar{H}(t)\;. \tag{8}$$

Figure 5 visualizes all the above so far.

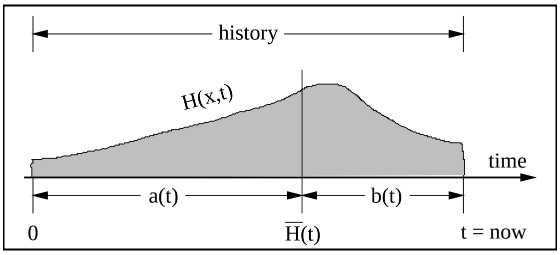


Figure 5: Example of a history weight function.

Adaptive systems may be classified according to the behaviour of $\bar{H}(t)$ as the history grows, that is $t \to \infty$. We distinguish between the following classes of adaptivity:

- **non-adaptive:**
$$a(t) = 0\;, \quad b(t) \to \infty\;.$$
- **locally adaptive:**
$$a(t) \to \infty\;, \quad b(t) < c\;,$$
where $b(t) < c$ means that $b(t)$ is bounded by a constant c as $t \to \infty$.

- **asymptotically adaptive:**

$$a(t) \to \infty \ , \quad b(t) \to \infty \ .$$

Non-adaptive systems do not use the history whatsoever. Asymptotically adaptive systems spread the emphasis over the whole time-line in such a way that the mean $\bar{H}$ in not bounded. An example of an asymptotically adaptive system is a system which weighs all events of the past equally, like one with a history weight function of Equation 6.

Locally adaptive systems rely most heavily on data collected in the recent past, degrading the value of the early past as the history grows. A minimum amount of *emphasis* is always given to a bounded length of the recent history, and the rest of the emphasis is spread over the rest of the history. A special case of local adaptivity shows up in *windowed locally adaptive* systems which consider only the recent history within a fixed time window. A typical history weight function of this form is:

$$H(x,t) = \begin{cases} 1/W \ , & \text{if} \quad t - W \leq x \leq t \ . \\ 0 \ , & \text{if} \quad 0 \leq x < t - W \ . \end{cases} \tag{9}$$

where W is the window size.

In Section 7, we will discuss the effectiveness of the aforementioned types of adaptivity in relation to the nature of user interests. First, turning the theory into practice, we will discuss some practical issues in implementing adaptivity.

6 Adaptivity and Incrementality

Our discussion so far has assumed that the whole history and an unlimited amount of memory and computational power are available at any point in time. However, practical models in order to be feasible should satisfy the following requirements:

- use a fixed finite amount of memory.
- process the available history in a fixed finite number of computations.

These requirements imply that only a finite portion of the history should be retained, and that models should be implemented *incrementally*.

Let us assume a filtering model that records frequencies of certain features occurring in relevant documents, in order to make predictions of relevance in the future. Incremental asymptotic adaptivity in such a simple model can be achieved by accumulating the values of the occurring features in an array of registers; one register per feature. Of course, there is another minor concession we make here, that is to allow registers of infinite width. Double precision arithmetic approximates this assumption well; in any case, all accumulators can be divided by a constant, whenever a value approaches the maximum width of the registers, without invalidating the model.

A locally adaptive system may be implemented in a similar manner by additionally maintaining a document buffer of some length W. Every time a new document arrives, registers accumulate the values of the occurring features, but they are also decremented by the values of features which occurred in the oldest document in the buffer. This approach is incremental, but it has two disadvantages: it uses more memory because of the document buffer, and it discards all information beyond what is in the buffer at any time.

An alternative approach, which uses all information but weighs it appropriately, is to perform a *decay* operation. We define the *half life* h of a document as the age that a document must be before it is half as influential as a fresh one. If a document D_i has arrived at time t_i and the current time is t_n, the history weight of the document is:

$$l_i = \exp\left(\frac{\ln 0.5}{h}(t_n - t_i)\right) \ , \tag{10}$$

where t_n, t_i, and h are expressed in the same units, e.g., months. Figure 6 demonstrates the decay operation.

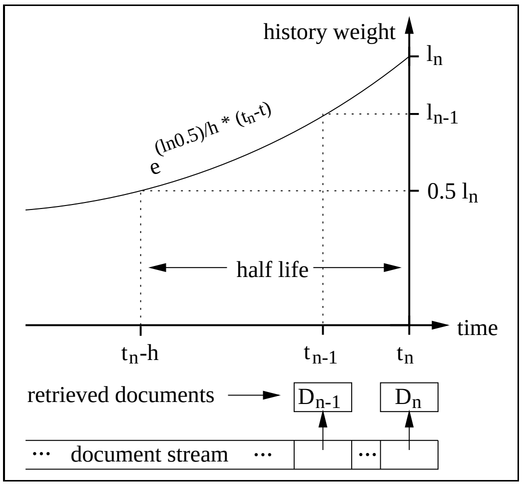


Figure 6: Decay and half life.

The decay operation can be performed incrementally, and it does not require any document buffers. It is easy to show that when D_n arrives, all accumulators have only to be multiplied by l_{n-1} before the new values of the features occurring in D_n are added.

7 A Comparison of Adaptivity

Non-adaptive systems will perform poorly, unless the initial topic representation is complete and precise and the topic is stable. However, the initial representation is bound to be incomplete and imprecise, due to two factors:

- the incapability of users to verbalize their precise interest,
- the weaknesses of the representation scheme itself.

Consequently, locally and asymptotically adaptive systems present more interesting features.

Locally adaptive systems use more recent information and they are capable of responding to relevance shifts quickly. Therefore, they can *track* a drifting topic. However, the disadvantage of them is that they will never converge to a stable topic. Asymptotically adaptive systems have the ability to *converge* to a stable topic. The choice between a locally or asymptotically adaptive system should be made on whether *responsiveness* or *convergence* is more important.

Implicit in the idea of tracking a topic using the history is that the history gives an indication of where the topic may currently be located. A fundamental trade-off exists in tracking topics. While it is advantageous to use as many instances of the history as possible to estimate accurately a topic's position, it is disadvantageous to use outdated instances. Relevant instances of the far past indicate the position of the topic at the time they occurred and they do not reflect the topic's current position. Thus they are less informative than recently occurred instances. A practical solution to this problem is to estimate the speed of a drifting topic and use this estimation to choose an appropriate window size W or half life h.

In the TREC-9 filtering task, the user requests were given as being stable, suggesting that an asymptotic behaviour would be more proper. However, the test stream (OHSUMED) consists of documents collected in a period of five years and it is likely that there are document content drifts. As an example, think of new treatments developed for the same sickness. Indeed, our experiments have shown that the average effectiveness (as this is measured by T9U) peaks for a half life value of around 4 years [5]. Analysis per topic, however, has revealed that effectiveness is optimal at a considerably different half life value per topic.

As a first step in optimizing h per topic, we define the *effective relevance velocity* v of a topic as:

$$v = \frac{d(D_1, D_n)}{n-1} \ , \ \ v \in [0, \epsilon] \ . \tag{11}$$

Note that the definition considers only the initial and the last position of relevance, and discards the trajectory in between. Moreover, the velocity is defined with respect to the number of steps taken, rather than the actual time. Obviously, h and v are related in an inverse way, however, their more precise relation should be established experimentally.

The types of adaptivity we have defined are capable of dealing with stable and drifting topics. The question of how multimodal or vagrant topics should be treated still remains. A solution would be to model their subtopics separately. In the multimodal case, all subtopics may be assumed stable and be dealt with by asymptotic adaptivity. However, it may be more effective for the vagrant case to assume that subtopics are drifting. We should remind the reader that the *poles* (the gray circles in Figure 2) of a vagrant topic may not be revisited by relevance in the future. Thus, a locally adaptive system would eliminate such old outdated context. Next, we will discuss an alternative way of dealing with multimodality and vagrancy.

8 Stabilizing Multimodality or Vagrancy

The solution of modeling subtopics separately is not practical, although it may be effective. A more practical solution is to, first, re-construct the document space so as to bring the different poles as close together as possible, and then assume a larger fuzziness parameter so that the topic may be considered as stable or drifting. The idea of re-constructing the document space is widely known as *feature selection*. It is depicted in Figure 7.

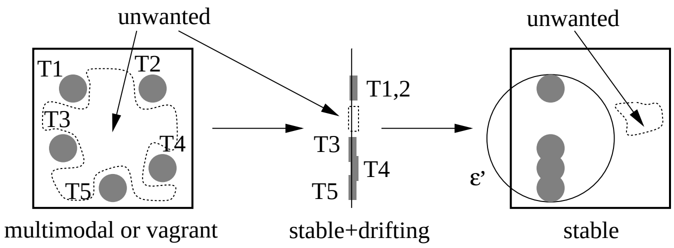


Figure 7: Stabilizing multimodality or vagrancy by re-constructing the document space.

The first transformation shows how poles may be brought together by eliminating a dimension (feature). The second transformation shows how unwanted areas of the document space may be moved away from the topic area by adding a dimension (a different dimension than the one eliminated before). By selecting an optimal set of features in this way and by increasing the fuzziness constant (e.g. to ε') if necessary, multimodal or vagrant topics may be treated as stable.

Traditional feature selection schemes usually favour features which occur frequently in relevant documents but infrequently in the rest. In order to eliminate multimodalities or vagrancies, however, it is also important that a feature occurs across poles; these features bring the poles together. High frequency in relevant documents implies that a feature may also occur across poles, but not necessarily.

The uniformity measure we have introduced in Section 4.1 may be recruited once more. Based on the hypothesis that features which occur uniformly in time are more valuable than others, we have introduced in [4] a novel feature selection method, namely the *term occurrence uniformity* (TOU). A small experiment has neither proved nor disproved the hypothesis. The results, however, have been promising, since the method seemed as effective as other powerful term selection methods such as document frequency thresholding[2].

[2]Document frequency thresholding has proven to be more than just an *ad hoc* approach for feature selection, and quite powerful in text categorization environments [13].

The approach taken has been a brute-force one; candidate features were ranked simply according to their uniformity. A wise integration of a TOU method and some other powerful time-disregarding term selection method may combine the benefits of both approaches.

A fundamental difference between adaptive filtering and classification (non-adaptive) systems is that in filtering the document space may be reconstructed several times in order to optimize effectiveness and efficiency. *On-the-fly* feature selection schemes should be applied with respect to possible relevance shifts. Moderate cutoffs will be more appropriate. Due to the fixed-memory model required for practical systems, every time a cutoff is applied, some low-frequency features will be irretrievably lost. Relevance drifts are associated with frequency increments of previously low-frequency features. Therefore, applying repeatedly aggressive cutoffs will not allow for the tracking of relevance drifts.

9 Selectivity Traps

The output of traditional retrieval systems is usually a ranked list of documents in their decreasing scores (given by the probability of relevance or some other similarity measure) with respect to a query. In binary classification tasks, like document filtering, a decision should be made for every document as to whether it belongs to a given class or not. Thus, decisions such as where to "cut" a ranked list have to be made automatically. In some cases, decisions are required to be made as soon as a document arrives, therefore ranked lists are not even possible.

These considerations suggest the *thresholding* of document scores. We will not expand on thresholding here; in [14], we have elaborated on the score-distributional threshold optimization method. Thresholding has proven to be critical for classification effectiveness and has revealed the twin danger — unique to such environments — of *selectivity traps*: setting a threshold too high retrieves nothing at all, while setting it too low retrieves far too many documents [15]. We will call these traps *overselectivity* and *underselectivity*, respectively.

Bad thresholding, however, is not the only cause of falling into selectivity traps. Another cause may be training. Usually, a system is trained on its history, i.e. it is trained to do past tasks, and then it is applied to future tasks. Consequently, the success of training depends on whether the lessons learnt from the past apply to the future. The most obvious reason why this might not hold is that a topic is drifting faster than a system is capable of tracking. We will call this trap *intractability*.

Another danger of training is what is widely known as *overfitting* a topic profile on history data. For example, putting too much effort into finding the perfect profile for the history may discover and emphasize accidental characteristics (e.g. typographical errors in relevant documents) that do not generalize into the future. Overfitting usually manifests itself as overselectivity. At the other end of the spectrum lies *underfitting*, which leads to underselectivity. Available training data may not be sufficient for training, subsequently the topic profile is far from convergence describing a bit too much of the document space. Table 9 summarizes the possible traps, their causes, and their criticalities for adaptive filtering.

	underselectivity	**overselectivity**	**intractability**
causes	– underfitting – too low threshold	– overfitting – too high threshold	– too fast drift
criticality	not too dangerous	dangerous but recoverable	unrecoverable

Table 1: Selectivity traps.

In adaptive filtering, overselectivity is a more dangerous trap than underselectivity. Adaptivity in filtering counts on the system to keep retrieving documents so it can continuously refine the filtering model. In this respect, adaptivity can pull the system out of an underselectivity trap by improving the topic profile and increasing the threshold. On the other hand, if at some point in time the system is led to an overselectivity trap, it will not retrieve any documents on which it can refine the topic profile and threshold, which leads to "silent" profiles. However, such a situation may be recoverable by the use of special mechanisms; the questions are how one can detect and fix an overfitted topic profile, and how can one be sure that the threshold is too high (as opposed to there just being no relevant documents to be retrieved).

In the long term, the intractability trap has essentially the same effect as overselectivity. Even if a profile still retrieves non-relevant documents when it has lost the relevant document area, these non-relevant documents only give an indication of the area that the estimation of relevance should move away from, without specifying an alternative direction. The profile will eventually fall "silent", because of adaptivity responding by increasing the threshold. We prefer, however, to view intractability as a separate trap from overselectivity, since its cause is different and the situation is rather hopeless and unrecoverable.

One should keep in mind that adaptive filtering is an especially sensitive task. What makes it so sensitive is that the system is provided with absolutely no relevance feedback for non-retrieved documents. Any relevance statistics collected in this way are bound to be *partial* in the sense that they do not represent a sample of the whole document space, but a sample of the *retrieved* space, therefore they may be highly misleading. Compare this situation to other adaptive tasks such as adaptive data compression, where the current frequencies of *all* symbols in a channel are known.

10 Outlook

This theoretical study summarizes our experiences in viewing filtering as an adaptive and temporally-dependent process. All models and ideas we have described are the result of our experimental work in the context of the TREC-9 filtering task [5], [4], and of previously unpublished empirical investigations, and they result in a coherent view on relevance feedback environments involving temporally dependent data.

We have presented a collection of ideas: a definition of the filtering task, a definition of the topic, two orthogonal classifications of topics (one based on relevance and the other on temporal aspects), a classification of adaptivity, and ways of using temporal information for selecting documents and for feature selection. Moreover, we have discussed potential dangers such as selectivity traps, and paid attention to practical issues such as incrementality. Our analysis has been rough, and we rather pose more questions than provide answers.

The classical view of the filtering task as a special case of the traditional information retrieval task, we believe, is not appropriate. In the last few years, there has been increasing evidence that viewing filtering as an adaptive and temporally-dependent task is beneficial for effectiveness. We are convinced that Information Retrieval in general could benefit by taking into account the effect of adaptation and time. Our work so far is fully described in [16] and it has concentrated in working out these issues.

Acknowledgments

I would like to thank Kees Koster (Computer Science, KUN), Petros Draggiotis (Theoretical Physics, KUN), and Rachael Rafter (Computer Science, UCD), for proofreading this work, and the anonymous reviewers for their useful comments.

References

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

[2] Douglas W. Oard and G. Marchionini. A Conceptual Framework for Information Filtering. Technical Report CS-TR-3643, University of Maryland, Computer Science Department, 1996.

[3] Ross N. Williams. *Adaptive Data Compression.* Kluwer Academic Publishers, Boston/Dordrecht/London, 1991.

[4] Avi Arampatzis, Th.P. van der Weide, C.H.A. Koster, and P. van Bommel. Term Selection for Filtering based on Distribution of Terms over Time. In *Proceedings of RIAO'2000 Content-Based Multimedia Information Access*, pages 1221–1237, Collège de France, Paris, France, April 12–14 2000.
Also available from `http://www.cs.kun.nl/~avgerino`.

[5] Avi Arampatzis, Jean Beney, C. H. A. Koster, and Th. P. van der Weide. Incrementality, Half-Life, and Threshold Optimization, for Adaptive Document Filtering. In Ellen M. Voorhees and Donna K. Harman, editors, *The Nineth Text REtrieval Conference (TREC-9)*, Gaithersburg, Maryland, November 13–16 2000. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication.
Also available from `http://www.cs.kun.nl/~avgerino`.

[6] Peter J. Denning. Electronic Junk. *Communications of the ACM*, 25(3):163–165, March 1982.

[7] Joep Simons, Avi Arampatzis, Bernd C.M. Wondergem, Lambert R.B. Schomaker, Theo P. van der Weide, and C.H.A. Koster. Profile — A Multi-Disciplinary Approach to Information Discovery. In *Proceedings of the Second International Bi-Conference Workshop on Agent-Oriented Information Systems (AOIS-2000)*, 2000.
Also available from `http://www.cs.kun.nl/~avgerino`.

[8] Thomas W. Malone, Kenneth R. Grant, Franklyn A. Turbak, Steven A. Brobst, and Michael D. Cohen. Intelligent Information Sharing Systems. *Communications of the ACM*, 30(5):390–402, May 1987.

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

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

[11] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing, 2nd ed.* Cambridge University Press, Cambridge, UK, 1992.

[12] D. E. Knuth. *The Art of Computer Programming — Volume 2: Seminumerical Algorithms*. Addison Wesley, 1981.

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

[14] Avi Arampatzis and André van Hameren. The Score-Distributional Threshold Optimization for Adaptive Binary Classification Tasks. Technical Report CSI-R0105, University of Nijmegen, January 2001. Available from `http://www.cs.kun.nl/~avgerino`.

[15] Stephen Robertson and Stephen Walker. Threshold Setting in Adaptive Filtering. *Journal of Documentation*, 56:312–331, 2000.

[16] Avi Arampatzis. *Adaptive and Temporally-dependent Document Filtering — Experiments with threshold optimization, local adaptivity, time distributions, linguistically motivated indexing, and incrementality issues, in relevance feedback environments* . PhD thesis, University of Nijmegen, Nijmegen, The Netherlands, 2001. To appear.