Minimizing Information Overload:
The Ranking of Electronic Messages
Journal of Information Science
15 (3) 1989, 179–189.

Robert M. Losee, Jr.
U. of North Carolina
Chapel Hill, NC 27599-3360 U.S.A.
losee@ils.unc.edu

June 28, 1998

Abstract

The decision to examine a message at a particular point in time should be made rationally and economically if the message recipient is to operate efficiently. Electronic message distribution systems, electronic bulletin board systems, and telephone systems capable of leaving digitized voice messages can contribute to "information overload," defined as the economic loss associated with the examination of a number of non- or less-relevant messages. Our model provides a formal method for minimizing expected information overload.

The proposed adaptive model predicts the usefulness of a message based on the available message features and may be used to rank messages by expected importance or economic worth. The assumptions of binary and two Poisson independent probabilistic distributions of message feature frequencies are examined, and methods of incorporating these distributions into the ranking model are examined. Ways to incorporate user supplied relevance feedback are suggested. Analytic performance measures are proposed to predict system quality. Other message handling models, including rule based expert systems, are seen as special cases of the model. The performance is given for a set of UNIX shell programs which rank messages. Problems with the use of this formal model are examined, and areas for future research are suggested.

1 Introduction

With the increase in number and importance of sophisticated telecommunications systems, microcomputers, and computer terminals in offices, we have entered an age in which information overload is an increasingly common occurrence [15]. Information overload has been defined as information received at such a rapid rate that it cannot be assimilated [31]. Sowell suggests that a sufficient inundation of information may lead to information saturation [33]. When saturation occurs, less attention is paid to each received message and thus less information is received. Hiltz and Turoff suggest that in message systems one may suffer from information entropy, "whereby incoming messages are not sufficiently organized by topic or content to be easily recognized as important" [13]. They further note that "structures that will distinguish communications that are probably of interest from those that probably aren't" are desirable.

Controlled experimental studies have suggested that reasoning decreases when large amounts of information are provided [14, 24, 25]. As military systems increase in complexity, performance figures such as weapons accuracy have been shown to decrease [34]. Fighter pilots often disable warning devices in the cockpit to avoid the information overload that they and their compatriots have experienced in past life-threatening situations.

For our purposes, we will define information overload as the receipt of more information than is needed or desired to function effectively and further the goals of an an individual or organization. More formally, receipt of undesirable or non-relevant messages that result in an economic loss for the recipient represents information overload. A system that explicitly minimizes the probability of retrieval of such economically undesirable messages minimizes the expected information overload.

We propose a model which can determine whether a message should be examined, based on all message features available to the recipient. Although our model is media independent, we will be particularly concerned with those oral, written, or keyboarded messages in natural languages capable of being stored in digital form and thus easily analyzed by a computer system [1, 10, 11, 37, 38]. The model, based on economic and statistical decision theory [2, 41] and making use of work done in information retrieval and automatic indexing systems, suggests a method whereby unexamined messages are ranked [29] from those most likely to be of interest, or "relevant" to our needs vis a vis the organization, to those least likely to be relevant. The recipient of more than one message may examine the message with the greatest expected worth, then the message with the next highest expected worth, and so on until a decision is made that the examination of each unexamined message remaining has a negative expected worth [17, 20].

Hiltz and Turoff "believe that no automated routine can simultaneously filter out all useless and irrelevant communication for addresses, and at the same time assure their receipt of all communications that may be of value to them." [13] We agree that such a perfect system may never be developed. We do feel that, given a decision theoretic model's constraints, an optimal system may be developed. Hiltz and Turoff further claim that "there can be no single design that will optimize the tradeoffs between useful information from unanticipated sources and information overload." Our system provides economically optimal performance and is capable of learning to what

extent the system should pay attention to "unusual" messages and, at the same time, avoid information overload.

Hiltz and Turoff make several conclusions about needed functional capabilities of a message handling system. They note, for example, that "to inhibit the flow of 'useless junk' is to risk the loss ... of potentially useful information." The proposed model, through proper parameterization, may allow differing classes of messages to be examined as desired by the user. The system may determine the bandwidth and shape of a message filter, unlike most existing systems. They also note that "Individuals learn to self-organize communication." Many existing message handling systems have the capability to avoid the examination of certain messages, based upon the presence of features with certain values, and may allow examination of messages sorted, for example, by subject or arrival time. Our system has the capability to act as these systems do but, more importantly, a decision theoretic model may learn what is important to the user. Self-organization may be seen as an inherent feature of our model.

It should be noted that the proposed model is designed to economically optimize the information reaching the user, *not* the information handling capabilities of an entire message handling system. We say nothing, for example, about the cost of sending so-called "junk-mail" and the effect it may have on network performance.

2   Message Presentation Systems

We define a message as information made available to an organization or individual who may, in turn, choose to examine, ignore, memorize, or discard the information. Messages may be in any form or transmitted through any medium; the only restriction we will place upon a message is that it have feature values which are available to facilitate a decision by the recipient to examine or not examine the message's full text. For example, the fact that a message did or did not originate at site "cs.unc.edu" must be available and binary if it is to be considered a feature. Models similar to those discussed below and consistent with fuzzy feature values have been discussed by Bookstein and Ng [5].

Message features may be assigned to a message explicitly or implicitly, either by the originator of the message or by the recipient. When formatting an electronic message for transmission, one is often asked for one or more keywords indicative of the subject of the message to be attached to the message header. Other features may not be intentionally assigned by the originator yet would be of significance to the recipient. For example, a message with the explicitly assigned keyword "lunch" would have less significance for me than a message which mentions the name of my favorite Chinese restaurant in the body of the message; the system learns what textual features occur in messages of greatest interest to the recipient. A system based upon the proposed model is capable of learning such personal quirks.

Virtual features assigned values by the recipient may reflect the relationship between the message and the current state of the message presentation system. For example, when reading a bulletin board type of messaging system, such as the USENET news, it may be desirable for the user to be able to examine postings to the network grouped by subject. Messages may thus be assigned the virtual feature ¡*shares feature*

X with message currently being examined¿; the value of this virtual feature will change as different messages are examined.

Existing message-receiver interfaces often have the capability to include or exclude a message based upon one of a small number of features, including ¡*author*¿, ¡*subject*¿, ¡*origination date*¿, etc. [9, 16, 35, 32]. These systems often lack the ability to combine features intelligently when making decisions. Exceptions include the Information Lens system described by Malone et al., which allows rules combining features to be included in the message selection system [26] and a similar system proposed by Chang and Leung [7]. An illustrative rule Malone provides is

IF	From:	Silk, Siegel
THEN	Set Characteristic:	VIP
IF	Message type:	Action request [and]
	Characteristic:	VIP
THEN	Move to:	Urgent.

Messages from Silk or Siegel are labeled as being from a VIP. If an incoming message is from a VIP and requests action, the message should be placed in a location marked "urgent" for immediate examination. This and other recent research have improved the quality of message presentation interfaces, yet formal models for the presentation of messages have received little attention.

3   An Economic Model for Examination Decisions

We now suggest a formal economic rule for deciding whether to examine a message: a message should be selected for examination if the cost of doing so is less than the cost of not doing so. This may be formalized as a decision rule: select a message if and only if $C_s < C_r$, where C_s denotes the cost of selecting a message for examination and C_r the cost of rejecting a message–that is, choosing not to examine a message.

Before deciding to examine a message, these costs must be estimated if the decision is to be rational and economically based. A decision rule based on these expected costs might then be to select a message if and only if the expected cost of selection for examination is less than the expected cost of non-selection for examination, or,

$$\Pr(G|m)C_{S,G} + \Pr(B|m)C_{S,B} < \Pr(B|m)C_{R,B} + \Pr(G|m)C_{R,G},$$

where G and B ("good" and "bad") denote the binary message qualities *relevant message* and *non-relevant message*, S and R represent the decision to select a message for examination or reject a message, $Pr(relevance\ class|features)$ denotes the probability a message with features m is in the given relevance class, and $C_{a,c}$ represents the conceptual cost of performing action a given that condition c, a relevance class, holds. If we expect information overload to occur, the expected costs associated with examining a message must exceed the expected costs of non-examination of a message we have decided to examine. If we obey our rule, we should not expect information overload, although some may occur.

4

Through algebraic manipulation, the following decision rule is obtained: select a message if and only if

$$\frac{\Pr(G|m)}{\Pr(B|m)} > \frac{C_{S,B} - C_{R,B}}{C_{R,G} - C_{S,G}} = C,$$

where C denotes a cost ratio combining our earlier cost figures. Below we denote $\Pr(G|m)/\Pr(B|m)$ as the Message Status Value of a message with set of features m (MSV_m).

Most message recipients will find it difficult to estimate the cost constant C for a given message, making our model appear difficult to use. We may still use our model by ranking messages by their *MSV*. The system presents for user examination the message with the highest *MSV*, then the message with the second highest *MSV*, and so on. The selection of messages ceases when the selector feels that an appropriate cutoff point in the message examination process has been reached. Note that the model does not explicitly claim that the first messages examined are "better" than those examined later in the process, although this can be shown through use of different constraints [4]. The ranking is performed solely because the cost constant C is unknown. Those messages at the bottom of the ranked list are the messages with *MSV* values below C whose retrieval would likely result in information overload if given to the user for examination.

Our ranking model requires us to assume that (1) messages are either relevant or non-relevant and (2) costs are identical for examining each message involved in the ranking. Both assumptions are only approximations to the true state of messages, but they allow the development of a relatively simple model. A decision theoretic model suggesting means to avoid making assumptions of binary quality or identical costs for all messages is suggested by Bookstein [4].

Using Bayesian methods, our *MSV* may be re-expressed as

$$MSV_m = \frac{\Pr(m|G)}{\Pr(m|B)} \frac{\Pr(G)}{\Pr(B)}.$$

Because the prior odds of message relevance, $\Pr(G)/\Pr(B)$, remain constant for all messages, the expression $\Pr(G)/\Pr(B)$ may be removed from our definition if the *MSV* is used only to rank messages for selection. The ranking function may now be written as

$$MSV_m = \frac{\Pr(m|G)}{\Pr(m|B)}$$

If the simplifying assumption of feature independence is made, our function may be further modified. The independence assumption states that, if the feature set m is explicitly represented in terms of its feature components, $m_1, m_2, m_3 \ldots, m_i$, then

$$\Pr(m|G) = \Pr(m_1|G)\Pr(m_2|G)\Pr(m_3|G)\cdots\Pr(m_i|G),$$

with a similar expression for non-relevant messages. Use of this simplifying assumption will allow us to develop a more easily understood model but will result in some

performance degradation of a system based on the model [22, 42]. Similar but computationally more expensive models formally incorporating dependence information are discussed below.

Assuming feature independence, the ranking function becomes

$$\prod_{i=1}^{j} \frac{\Pr(m_i|G)}{\Pr(m_i|B)}$$

assuming the availability of j features concerning the message.

If the f features with non-zero values are rearranged so they are first and numbered 1 through f, our function becomes

$$\prod_{i=1}^{f} \frac{\dfrac{\Pr(m_i|G)}{\Pr(0_i|G)}}{\dfrac{\Pr(m_i|B)}{\Pr(0_i|B)}} \prod_{k=1}^{j} \frac{\Pr(0_i|G)}{\Pr(0_i|B)}$$

The right hand portion does not vary among messages, and may be ignored if the function is only used for ranking, suggesting

$$MSV_m = \prod_{i=1}^{f} \frac{\dfrac{\Pr(m_i|G)}{\Pr(0_i|G)}}{\dfrac{\Pr(m_i|B)}{\Pr(0_i|B)}}$$

4   Binary Independent Features

To apply this decision theoretic message ranking procedure, it is necessary to compute probabilities of message features. Those features of a message which take values for m equal to 0 or 1 and are statistically independent of other features are said to be binary independent features. Typical binary information might include whether a message is from one's superior, whether everyone in the organization is scheduled to receive a copy of a message, or whether a message is about a specified topic. It may also be useful to code non-binary data to a binary form. For example, we may categorize all messages received in the past 5 business hours as "recent," and all other messages as "old."

A binary feature's distribution in relevant messages may be described by

$$\Pr(m_i|G,p) = p^{m_i}(1-p)^{1-m_i},$$

with a similar expression for non-relevant messages using the parameter q. Parameter p denotes the probability a feature will have a value of 1 given that a message will be relevant, while q denotes the probability that a feature will have a value 1 given that a message will be non-relevant. Note that subscripts are omitted for notational simplicity where it is obvious that an individual feature is being referred to.

Our criterion function under the binary independence assumption is now

$$\prod_{i=1}^{f} \left(\frac{p_i/(1-p_i)}{q_i/(1-q_i)} \right)^{m_i}$$

For simplification, this may be transformed to

$$\sum_{i=1}^{f} m_i \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)},$$

if the expression is used only for ranking messages. Features occurring only in relevant messages have a positive MSV, indicating their ability to discriminate between relevant and non-relevant messages, while those occurring with similar frequencies in both classes of messages have a discrimination value near 0, and those features more likely to occur in non-relevant messages have a negative discrimination value.

We may generalize from this result by noting that if statistical independence of features is assumed under several feature distribution models, we may rank messages for selection by

$$MSV = \sum_i m_i \delta_i$$

where the δ_i component measures the degree to which feature i distinguishes between messages to be examined and messages to be rejected. In the binary independence model,

$$\delta = \log \frac{p/(1-p)}{q/(1-q)}$$

represents the discrimination value of a binary independently distributed feature.

5   Poisson Independently Distributed Features

The distribution of occurrences of features having non-binary integer values may be modeled in some circumstances by the Poisson distribution. It is possible to think of such features as having been produced by a conceptual feature generator that produces features at a fixed average rate with the interval between feature occurrences a random exponential variable. The Poisson model has been suggested as providing a satisfactory description of the occurrence frequencies of natural language terms [12, 23, 36]. Use of this distribution may be helpful if advantage is taken of content analytic methods for counting the number of times particular positive or negative terms or expressions occur in a message [6, 18]. Because few short messages have more than a single occurrence of each content-bearing word, the Poisson distribution is most effectively used for lengthy messages.

The Poisson distribution suggests that a feature occurs with frequency m with probability

$$\Pr(m|G, \lambda) = exp(-\lambda)\lambda^m/m!,$$

with a similar expression using parameter $\bar{\lambda}$ for non-relevant messages. Parameters λ and $\bar{\lambda}$ represent the average frequency of features in relevant and non-relevant messages, respectively.

A *MSV* based on this distribution is

$$\sum_{i=1}^{f} m_i \log \frac{\lambda_i}{\bar{\lambda}_i},$$

where $\delta_i = \log(\lambda_i/\bar{\lambda}_i)$ represents the discrimination value of feature i.

Feature weights of different distributional types may be summed together if statistical independence of features is assumed. For example, given two features, binary distributed feature 1 and Poisson distributed feature 2, the *MSV* would be calculated as

$$m_1 \log \frac{p_1/(1-p_1)}{q_1/(1-q_1)} + m_2 \log \frac{\lambda_2}{\bar{\lambda}_2}$$

6 Binary Dependence Models

Message features are usually statistically dependent, not independent as we have assumed. Several different techniques have been proposed to incorporate dependencies into decision theoretic models similar to the one proposed [19, 39]. Such models have met with limited success in other similar applications, due in part to the increased difficulty in estimating the greater number of parameters found when dependence information is incorporated into the model.

The quality of a parameter estimate varies with the number of items used to make the estimate. Thus, in realistic situations, there is a trade-off between the system performance obtained by using a more accurate (e.g. dependence) model with less accurate parameter estimates, and performance achieved by providing good estimates of parameters based on large sample sizes with an imprecise model. Dependence models incorporate combinations of features; such combinations occur with less frequency than do the features taken individually, resulting in less accurate estimates of dependent term frequencies. Thus, performance may increase if an unrealistic statistical independence assumption is made (as above) and a small number of parameters are accurately estimated, rather than if a more realistic dependence assumption is made, producing poor estimates for a larger number of parameters [22, 28].

Yu et al. [42] provide a summary of the computational techniques useful in models incorporating arbitrary amounts of dependence information. Experimental results suggest that such dependence models may be useful and provide accurate rankings, given accurate parameter estimates. Tests in the information retrieval field [22, 42] suggest that the probability text will be correctly classified is lower for dependence based systems than for independence based systems when learning has not taken place and crude initial estimates are used; this order is reversed when significant amounts of learning have taken place. System tests will be necessary to determine the relative worth of incorporating these computationally expensive dependence assumptions, compared to the independence assumption, into our message ranking model, given the difficulty of estimating parameters accurately.

An alternative to the use of dependence involves the judicious selection of virtual features representing groupings of similar types of dependent information. These virtual features have the capability to organize such information. As an example, consider the fire detection system in a five story building, where each floor has ten detectors spread evenly along its corridors. If a fire were to spread through a majority of the third floor corridor, a message stating that there was a large fire on the third floor would be more beneficial, in most cases, than numerous messages that individual sensors had detected fire. These organizing virtual features have a discrimination value greater than the discrimination values for the individual components used in making up the virtual feature, thus assuring that the organizing message with the virtual feature is received before the sensor messages.

7   Relevance Feedback

For our message ranking system to improve its performance over time, it must obtain more accurate estimates of parameter values. This is most easily done through the use of relevance feedback, in which the user supplies the system with a binary relevance judgment about each message [3, 21]. The system then revises its estimates of parameter values, based on the Bayesian combination of prior and new knowledge, and will thus provide more accurate rankings for the user in the future.

In the binary independence model, the parameter p for a given feature may be estimated as

$$p = \frac{S + s}{R + r}$$

where S and R are arbitrarily chosen values such that $S < R$, s denotes the number of relevant examined messages containing the feature, and r denotes the number of relevant messages examined. S and R represent knowledge available prior to examining any relevant messages and s and r represent posterior, or learned knowledge obtained from user-supplied relevance feedback judgments. The denominator of our definition may be denoted as R'; it represents our confidence in our current estimate. The effect of examining any additional relevant messages on the estimation of parameters is determined by the relative values of R and the number of messages in the learning sample set. The numerator, denoted as S', may be seen as our best guess of the expected value of p multiplied by the confidence factor R'. If we arbitrarily establish S and R at 1 and 2, respectively, and no relevant messages have been examined, thus making s and r both equal to 0, p would be computed as $(1 + 0)/(2 + 0) = 1/2$. If 17 messages were then examined and found relevant, with 5 having occurrences of the feature in question, our estimate of p would be revised to incorporate the new information: $p = (1 + 5)/(2 + 17) = 6/19$. A similar procedure may be used for estimating the parameter q for non-relevant messages, with parameters $\bar{S}$, $\bar{R}$, $\bar{s}$, and $\bar{r}$.

Under the two Poisson independence model, we may also incorporate learning into our estimation of the parameter values λ and $\bar{\lambda}$. We define $\lambda = (S + s)/(R + r)$, where S and R are initial values, as with the binary independence model; s denotes the number of times a feature has occurred in the relevant messages examined and r

denotes the number of relevant messages examined. A similar definition exists for $\bar{\lambda}$, with parameters $\bar{S}$, $\bar{R}$, $\bar{s}$, and $\bar{r}$.

Accurate estimates of the values of $\bar{S}$ and $\bar{R}$, representing our initial knowledge of the characteristics of non-relevant messages, may be obtained by making a reasonable approximation under either binary or two Poisson independent models [8, 21]. We begin by setting $\bar{R}$, representing our degree of confidence in the estimate, to 1. Variable $\bar{S}$ is then estimated as the number of occurrences of a feature in a large sample of messages divided by the number of messages in the sample; we make the assumption that if most messages in an organization are not relevant to a given user, we may estimate the characteristics of the messages a user will find non-relevant as being approximated by the characteristics of all messages.

8 An Analytic Performance Model

The expected performance of the proposed ranking model may be computed, given certain parameter values [20]. We may estimate the probability a message with set of features m is relevant as

$$\Pr(G|m) = \frac{\Pr(m|G)\Pr(G)}{\Pr(m)}.$$

Given feature independence, this may be modified to

$$\Pr(G|m) = \Pr(G)\prod_{i=1}^{f}\frac{\Pr(m_i|G)}{\Pr(m_i)}$$

where f represents the number of system features. Given historical data, we may directly calculate $\Pr(G)$ and $\Pr(m_i)$, as well as the parameters necessary to compute the feature probabilities $\Pr(m_i|G)$.

One use of this model is to estimate the number of relevant messages available. If we want to know the expected number of relevant messages, $E(G)$, from among the first n ranked messages, we calculate

$$E(G) = \sum_{i=1}^{n}\Pr(G|m_i),$$

where $\Pr(G|m_i)$ represents the probability of relevance for message i.

A second use for our performance model is to aid the user in determining when to cease examining messages [17, 20]. Figure 1 provides a graph of the expected performance of the ranking system. At any stage in the message examination process, it is possible to produce such a graph, showing the expected quality of both examined and unexamined messages, given the current estimates of parameters. The display of such a graph in a window with an indication of the position of the most recently examined message allows the user to examine the expected quality of the messages yet unexamined. Such information assists the user in accurately determining when it is appropriate

to stop examining messages. Inaccuracies in the graph are due to inaccuracies in parameter estimations. These errors should decrease as more messages are examined and more messages for each relevance class are available for the computation of parameter estimates.

9 A Comparison with Existing Models

We have developed a model which enables us to predict whether a message should be examined, given a set of costs. If the costs are unavailable, the messages may be ranked and the message predicted to be the most useful examined first.

Several traditional message ranking procedures may be described by the decision theoretic model developed above. One traditional means of handling messages is first-come, first-served, where messages are ordered in a First In, First Out (FIFO) queue for examination.

The FIFO model may be incorporated into our model by assigning an ascending message number to each message as it arrives. The discrimination value for the feature ¡*message number*¿ is made a negative value. All other features have zero discrimination value. The weight for message $n-1$ is thus greater than the weight for message n, which it also predates. Thus message $n-1$ would be examined before message n, resulting in FIFO examination of messages.

A second traditional model for ranking incoming messages is by category, where all the messages in one category are examined before any messages in a second category are examined, and so on. The categories are each represented by a single feature, such as ¡*message from boss*¿, ¡*message from spouse*¿, or ¡*message about widget production*¿. Messages are assigned to only one category. We may obtain ranking from our economic model consistent with the more traditional model if we renumber our features, for notational simplicity, so that a low numbered feature i, representing category i, is considered to be more important than any higher numbered feature j, $j > i$. Given the Importance of message i, $I(m_i)$, we see that

$$I(m_1) > I(m_2) > \ldots > I(m_{n-1}) > I(m_n)$$

for n features. We now assign ad hoc discrimination values δ for each feature in such a way that $\delta_1 > \delta_2 > \ldots > \delta_{n-1} > \delta_n$. Given binary independent values for message features and all other features being set to zero, our probabilistic model will result in message rankings identical to those suggested by the more traditional model.

The rules incorporated into the Information Lens system represent a sophisticated set of capabilities than these traditional message manipulation functions, and can emulate sophisticated human procedures for coping with messages. To obtain rankings equivalent to those provided by the Information Lens system, our decision theoretic model must incorporate dependence information or make use of virtual features. In an earlier example, we noted a rule stating that if a message arrived from either Silk or Siegel and was an "action request," it was to be treated as urgent. We incorporate this rule into a decision theoretic ranking system by setting appropriate prior parameters indicating, for example, that we have a great deal of confidence that if a message from

Silk is an action request and is received, then it is highly probable that the message is relevant.

10 A Hypothetical Application

Let us hypothesize that we are the director of research and development for a corporation planning to bring widgets onto the market and wish to use this decision theoretic model to assist in message ranking decisions. We have only two message features available to use in making decisions: (1) whether the message was produced over 5 hours ago, which we assume to be binary independently distributed and (2) the number of times widget occurs in a message, which we assume to be Poisson distributed.

Let us also assume that parameters have been computed for both features. For the binary feature, it was found that in the past, 50 percent of the messages 5 or less hours old were found useful, while only 5 percent of the messages over 5 hours old were found useful, thus p=.5 and q=.05. For the Poisson distributed feature, it was computed that the term widget occurred an average of two times in useful messages but only an average of 2/3 times in rejected messages, thus $\lambda = 2$ and $\bar{\lambda} = .666$.

Sets of features are given in Table 1 ordered by their *MSV* values. The computation of the *MSV* for the highest ranked message in the table is

$$4\ln\frac{2}{.66} + 1\ln\frac{.5/(1-.5)}{.05/(1-.05)} = 7.34.$$

Messages would be examined by starting at the highest ranked message and selecting individual messages, moving down the ranked list, until the recipient makes the decision to stop.

11 Experimental Separation of Messages

Several experiments have been performed to provide initial evaluations of the ability of a computer program consistent with the proposed model to separate relevant from non-relevant messages. Messages were obtained from USENET newsgroups. Software was developed using UNIX shell programs.

Parameters were estimated using an even-odd technique [30]. Even numbered messages were used to establish parameter values, then the odd numbered messages were ranked based on these parameter values. Next, the odd numbered messages were used to establish parameter values, followed by a ranking of the even numbered messages. This effectively doubles the number of rankings performed.

Performance results are given in terms of precision, the percent of messages presented to the user that are relevant, for each recall decile, that is, when 10%, 20%, ..., 80%, 90% of the relevant messages have been presented. To the left of each set of precision measures is the experiment number and a letter, with B representing the assumption of binary feature distributions and P representing the assumption of Poisson feature distributions. To the right is the average precision, a single valued measure combining the precision at 25%, 50%, and 75% recall. These results are provided not

to indicate what a user might expect in a production system but to provide a comparison indicating relative worth of different assumptions and techniques.

The first two experiments used one hundred messages from each of five newsgroups (rec.audio, rec.ham-radio, rec.aviation, sci.math, and comp.arch). Each of the newsgroups was compared with each other newsgroup, for a total of 10 comparisons. Messages in one of the two groups was arbitrarily chosen to be "relevant." The procedure separates different messages and groups like messages with the following results:

| 1B | 99.1 | 98.1 | 97.0 | 95.7 | 95.1 | 94.7 | 93.9 | 93.3 | 92.7 | 95.4 |
| 1P | 98.8 | 98.1 | 97.8 | 97.2 | 95.5 | 94.4 | 93.6 | 93.0 | 91.7 | 95.6 |

USENET messages contain headers, typically containing sender, subject, date sent, the name of the newsgroup, as well as keywords indicating the subject of the message. Can our decision theoretic procedure separate messages if there is no header and only the message body is included? The same set of comparisons were performed without the headers yielded the following results:

| 2B | 98.2 | 95.2 | 93.8 | 92.8 | 92.2 | 92.0 | 91.7 | 89.9 | 87.5 | 92.5 |
| 2P | 92.8 | 92.1 | 92.6 | 92.5 | 91.0 | 90.9 | 90.1 | 89.3 | 86.7 | 91.0 |

suggesting that the use of headers does provide some increase in system separation capabilities.

It is apparent that these procedures may be used to separate messages on significantly different subjects. A second set of experiments separated messages in single newsgroups based upon subjective judgments of relevance or non-relevance. Judgments of message relevance or non-relevance were obtained by analysis of so-called "kill" files produced by the UNIX "rn" (read news) program, which contain message subject lines that the reader wishes not to examine. Pre-existing kill files were used to determine which messages the user would claim were non-relevant, while other messages were marked as relevant.

Because of the limited number of newsgroups available with existing kill files and over 80 messages in the newsgroup, only 4 newsgroups were used for this set of experiments. With the use of even-odd parameter estimation, there are effectively 8 rankings used in each set of performance figures. As before, we begin with the ranking of messages based on the text and headers.

| 3B | 89.5 | 90.3 | 88.4 | 90.0 | 88.7 | 88.5 | 87.4 | 87.1 | 85.3 | 88.4 |
| 3P | 82.3 | 89.2 | 88.4 | 88.0 | 89.3 | 86.7 | 86.4 | 85.0 | 83.9 | 87.9 |

The second pair of evaluations used the message text without the headers:

| 4B | 81.9 | 86.9 | 88.4 | 90.0 | 85.5 | 85.5 | 85.4 | 85.8 | 86.7 | 86.3 |
| 4P | 82.3 | 86.9 | 88.6 | 88.0 | 85.3 | 85.1 | 84.5 | 85.3 | 84.0 | 86.0 |

We can conclude from these initial tests that (1) the proposed procedure is effective in separating relevant and non-relevant messages, although not a perfect separator; (2) the use of message headers does improve system performance very slightly; and (3) systems consistent with the binary independence model slightly outperform those consistent with the two Poisson independence model, as suggested in [22].

13

12 Summary

We have developed a formal model of the decision process used in deciding whether to examine a message in an organizational setting. The model may be used to rank for examination incoming electronic mail, digitized telephone messages, or electronic publications. Messages may be examined in order of expected worth to the user or in an order suggested by more traditional models, such as FIFO or ranking by membership in one of a number of categories.

A system which takes advantage of this cost minimizing model may be formally seen as minimizing the expected information overload, if information overload is defined as the examination of messages with an associated expected loss to the individual or organization, and taking into account model constraints. Future researchers may wish to further examine those assumptions made explicit by our model. Independence assumptions of the model are unrealistic, yet experimental evidence suggests that they may be satisfactory for many purposes. Systems based on independence assumptions may be superior to those assuming statistical dependence because of the difficulty of estimating parameters for large numbers of features. Future use of such models by high volume systems will allow for the determination of the relative merits of independence assumptions. The Poisson distribution has been shown to be a tractable and reasonable approximation of term distributions in a desired class of messages. Answers to these statistical questions, as well as the development of non-binary cost models of organizational messages, may allow for more accurate and thus more beneficial models of the decision to examine information in organizational settings.

References

[1] David Barcomb. *Office Automation*. Digital Press, 1981.

[2] J. O. Berger. *Statistical Decision Theory*. Springer-Verlag, 1980.

[3] Abraham Bookstein. Information retrieval: A sequential learning process. *Journal of the American Society for Information Science*, 34(4):331–342, September 1983.

[4] Abraham Bookstein. Outline of a general probabilistic retrieval model. *Journal of Documentation*, 39:63–72, 1983.

[5] Abraham Bookstein and K. K. Ng. A parametric fuzzy set prediction model. *Fuzzy Sets and Systems*, 17:131–141, 1985.

[6] D. S. Champion and M. F. Morris. A content analysis of book reviews in the AJS, ASR, and Social Forces. *American Journal of Sociology*, 78:1256–1265, 1973.

[7] Shi Kuo Chang and L. Leung. A knowledge based message management system. *ACM Transactions on Office Information Systems*, 5(3):213–236, July 1987.

[8] W. Bruce Croft and D.J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, December 1979.

[9] Peter Denning. Electronic junk. *Communications of the ACM*, 23:163–165, 1982.

[10] C. Ellis and G. Nutt. Computer science and office information systems. *ACM Computing Surveys*, 12:27–60, 1980.

[11] N. B. Finn. *The Electronic Office*. Prentice Hall, 1983.

[12] Stephen P. Harter. Probabilistic approaches to automatic keyword indexing: Part I. *Journal of the American Society for Information Science*, 26(4):197–206, 1975.

[13] S. R. Hiltz and M. Turoff. Structuring computer-mediated communications systems to avoid information overload. *Communications of the ACM*, 28(7):680–689, July 1985.

[14] Jacob Jacoby, Donald E. Speller, and Carol A. Kohn. Brand choice behavior as a function of information load. *Journal of Marketing Research*, 11(1):63–69, February 1975.

[15] R. Johansen, J. Vallee, and K. Spangler. *Electronic meetings: Technological Alternatives and Social Choices*. Addison Wesley, Reading, Mass., 1979.

[16] E. B. Kerr and S. R. Hiltz. *Computer Mediated Communication Systems*. Academic Press, New York, 1982.

[17] Donald H. Kraft and T. Lee. Stopping rules and their effect on expected search length. *Information Processing and Management*, 15(1):47–58, 1979.

[18] K. Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, 1980.

[19] K. Lam and C. T. Yu. A clustered search algorithm incorporating arbitrary term dependencies. *ACM Transactions on Database Systems*, 7:500–508, 1982.

[20] Robert M. Losee. Predicting document retrieval system performance using an expected precision measure. *Information Processing and Management*, 23(6):529–537, 1987.

[21] Robert M. Losee. Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science*, 39(1):8–16, January 1988.

[22] Robert M. Losee, Abraham Bookstein, and Clement T. Yu. Probabilistic models for document retrieval: A comparison of performance on experimental and synthetic databases. In *ACM Annual Conference on Research and Development in Information Retrieval*, pages 258–264, 1986.

[23] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[24] Naresh K. Malhotra. Information load and consumer decision making. *Journal of Consumer Research*, 9:419–430, March 1982.

[25] Naresh K. Malhotra, Arun K. Jain, and Stephen W. Lagakos. The information overload controversy: An alternative viewpoint. *Journal of Marketing*, 46(2):27–37, Spring 1982.

[26] T. W. Malone, K. R. Grant, F. A. Turbak, S. A. Brobst, and M. D. Cohen. Intelligent information sharing systems. *Communications of the ACM*, 30:390–402, 1987.

[27] Murray S. Mazer and Frederick H. Lochovsky. Logical routing specification in office information systems. *ACM Transactions on Office Information Systems*, 2(4):303–330, October 1984.

[28] S. E. Robertson, C. L. Thompson, M. J. Macaskill, and J. D. Bovey. Weighting, ranking and relevance feedback in a front-end system. *Journal of Information Science*, 12(1/2):71–75, 1986.

[29] Stephen E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[30] Stephen E. Robertson, C. J. Van Rijsbergen, and M.F. Porter. Probabilistic models of indexing and searching. In Robert Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*, pages 35–56, London, 1981. Butterworths.

[31] T. B. Sheridan and W. R. Ferrell. *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*. MIT Press, Cambridge, Mass., 1974.

[32] H. T. Smith, editor. *Computer Based Message Systems*. North-Holland, Amsterdam, 1984.

[33] Thomas Sowell. *Knowledge and Decisions*. Basic Books, New York, 1980.

[34] Spectrum Staff. Too much, too soon: Information overload. *IEEE Spectrum*, pages 51–55, June 1987.

[35] M. Stefik, G. Foster, D. Bobrow, K. Kahn, S. Lanning, and L. Suchman. Beyond the chalkboard: Computer support for collaboration and problem solving in meetings. *Communications of the ACM*, 30:32–47, 1987.

[36] D.C. Stone and M. Rubinoff. Statistical generation of a technical vocabulary. *American Documentation*, 19(4):411–412, 1968.

[37] D. Tsichritzis. Form management. *Communications of the ACM*, 25:453–478, 1982.

[38] D. Tsichritzis, editor. *Office Automation*. Springer-Verlag, 1985.

[39] C.J. Van Rijsbergen. A theoretical basis for use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, June 1977.

[40] C.J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.

[41] Robert L. Winkler. *Introduction to Bayesian Inference and Decisions*. HRW, 1972.

[42] Clement T. Yu, Chris Buckley, K. Lam, and Gerard Salton. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2(4):129–154, 1983.