Properties of Electronic Texts for Classification Purposes as Suggested by Users

Fredrik Kilander
IntFilter
Department of Computer and Systems Sciences
Stockholm University
Electrum 230, S-164 40 Kista, Sweden
fk@dsv.su.se

January 15, 1996

1 Introduction

Students at the Department of Computer and Systems Sciencs (DSV) were given the task of writing a short essay on the problem of using machine learning to classify electronic texts and messages such as electronic mail (email), Usenet News postings or World Wide Web (WWW) pages. For the more than one-hundred students the essay was a small part of a wider examination. Fifty-four of them elected to adress the issue.

This paper summarizes the student's suggestions and opinions as real users of email, News and WWW. The purpose is to collect information from real users on what they perceive to be important properties in electronic messages. This is in turn motivated by the long-term goal of the IntFilter project at DSV to classify electronic texts guided by user-supplied examples.

2 The Authors

The authors of the essays were all students at DSV, at the time halfway through their second year of a four year curriculum. With all likelyhood, all of them used WWW and email at DSV, some of them also Usenet News. A non-neglible fraction indicated in their essays experiences from working in other environments where email played an important role. The majority of the authors were 20–30 years old. Female authors provided nine of the essays.

3 The Essays

Table 1 shows a summary of the properties suggested by the students. Most of them had approached the task with email in mind and listed the sender

Keywords	40
Sender	37
Subject	30
Length	21
Date and time	21
Categories of senders	13
Recipients	11
Relations to other texts	9
Semantic text analysis	6
Readability analysis	6
Keyword synonyms	5
Attachments	5
Language	4
Priority	3
Newsgroup	3
Layout	2
Negative keywords	2
Author's keywords	1
Nof readers	1
History of sender	1

Table 1: Summary of suggested properties

first, followed by the subject. This likely reflects the way most people peruse their mail. Many also noted keywords in the body of the message as an important property, but as an afterthought or after having spent time thinking further about the problem. The appreciation of length appeared to be a very individual issue: some authors apparently used length in conjunction with the sender to hypothesize the contents of the message (again, mostly for mail). Time was also an important factor for many; essentially the discussion revolved around deadlines, events and the time a message was allowed to be stored but not read.

3.1 Sender (37)

Although more students listed keywords as an important property, most of those who noted the origin of a message did so first. This indicates that this property was foremost in their minds when they set about to tackle the problem. The concept of the sender was developed in several ways:

- Senders can be persons, organizations or mechanisms.
- Senders have properties.
- Senders can be categorized.

Senders that are persons are usually known to the user. If not known personally the social and organizational relation between the sender and the user is often already established: supervisor, co-worker, employee, client,

customer etc. If the sending person is not known in this manner, certain information can be imputed from the organization behind the sender's adress. The domain adress symbolizes the company or university and lends credibility, suspicion or anonymity to the sender. In various degree this raises expectations about the contents and importance of the message.

The contents from senders that are organizations or mechanisms are usually more predictable: advertisements, regular subscriptions or mailing-lists has a known level of commitment for the user. They appear to carry the same distinctions in email as the morning paper and the monthly bill does for paper mail.

3.1.1 Categories of senders (13)

Several authors indicated that senders should be categorized for easier processing. Some recurring examples of such categories included *friends, clients, co-workers, providers of vital information* and *bosses*. Some authors recognized that the categories could overlap (your boss can also be your friend) in which case other mechanisms would have to be employed in order to determine which *role* the sender is having in the message.

3.1.2 History of sender (1)

One author mentioned that the *history* of a sender is an important property. This is quite natural considering that one learns to expect a certain amount of quality, quantity and content from each persistent source.

3.2 Recipients (11)

A number of authors felt that knowing the group of recipients of a message was important. For email, the group was defined as the immediate receivers and those who had received carbon-copies or were included in a chain of forwardings. Knowing that others has recieved a message influences several factors. Each user may be a member of several groups (social or organizational). As a member of a particular group each user can infer or hypothesize about the sender's expectations, ie the *role* and level of commitment expected of the receiver. In practical terms one is usually more inclined to answer a question recieved only by oneself, than one put to a group. In the group anyone can answer the question and some members may be uniqely qualified.

Mailing-lists and news define their groups explicitly by topic. In the case of news, messages crossposted to other groups was taken as an indication of content that spanned several interestgroups.

One author noted that the number of readers of a message was important. This is used on the WWW to estimate the popularity of certain pages. It is an insecure measurement, however, in that it doesn't reveal how many of the reader's that liked what they got. Also, good pages does not get read automatically (except by web-spiders), they must be fetched on demand.

3.3 Subject (37)

Many authors felt that the subject line was very important and eschewed that it be searched for keywords or phrases. Some authors recognized that subject lines can be useful, but also devastating if their import is overly emphasized. Writing a good subject or title is difficult and sometimes willfully abused.

Some author remarked that captions and subheadings were important properties of the overall text. This is recognized on the WWW where such entities can be easily examined through the HTML markup.

3.4 Keywords (40)

Almost all authors suggested that the classification of the message should be aided by a keyword search. Several speculated that each user could have a keyword list for each specific topic. Two writers noted that keywords could be negative, ie not allowed in a matching text. The lists were to be built manually, interactively from user-supplied examples or completely automatically in the style of a learning apprentice system.

One writer held that keywords supplied by the author (sender) should be treated with greater emphasis.

3.5 Keyword synonyms (5)

Some authors noted that the search for relevant keywords could be enhanced by using a dictionary of synonyms. It should be noted that the benefit of such a dictionary is greatest when there is a number of disassociated queries to answer, eg in an information retrieval system. If individual users create their own lists of keywords for specialized topics information in the dictionary is transferred into each list. The list then becomes, if you like, a micro-dictionary for each topic. A centralized dictionary is even more of value if it is capable of providing matches on word stems as well as synonyms.

3.5.1 Language (4)

Four authors recognized the distinction created by different languages and the possible confusion that could arise when keyword search did not take the language of the text into consideration. Further, the user is usually limited to a few languages and may take special interest in messages written in a particular language.

3.5.2 Readability analysis (6)

Six authors felt that a surface analysis of the readability of the message would be of value. Such analysis is usually based on the number of long words, the average number of words per sentence and perhaps the presence of certain idiom and phrases. The analysis results in some index indicating the required reading effort.

3.5.3 Semantic text analysis (6)

A few authors felt that the *meaning* of the text could not be disregarded and suggested that natural language methods be employed.

3.6 Length (21)

Length was quoted by many authors as important for categorization but the opinions on its interpretation varied, apparently from their personal experiences. The basic import of a message's length was that a longer message would take more time to read. Some authors argued that short messages usually were the result of errors or test and could therefore mostly be ignored, while others felt that the sender and subject determined the importance of short messages.

One interesting argument was that the length of the text could not be used alone to estimate the time needed to read the message. Length should be used in conjunction with readability analysis in that a short, difficult message may cause the user substantial more effort than a longer, more easily assimilated message.

One author used the personal heuristic that very long messages from particular senders consisted not text at all but of embedded binaries. The willingness to accept long messages could also be dependent on practical circumstance, for instance when using a dial-up communication line and a slow modem.

None of the authors suggested that the proportion between cited and original text was an important property, but neither did any one specify explicitly that length was begin counted only as original content.

3.7 Date and time (21)

Information about dates and times were considered to be important by many authors. Several suggested that important time information could be embedded within the message and should be identified and extracted. Surprisingly enough, some were apparently ready to discard messages about events past without reading them.

At least one author suggested that unread email should be subject to aging, ie receive higher priority with time. This implies that a sorting mechanism other than arrival time is in operation.

3.8 Relations to other texts (10)

A message sent in reply to a question was regarded as important. The impact on the desire to read the message was of course dependent on the context: most users are presumably interested in answers to their own questions, and less impressed by information they already have. As a side note, this is one of the reasons for the development of Frequently Asked Questions (FAQ) in mailinglists and news; experienced readers tire of seeing the same questions being raised and answered over and over again.

A few authors noted that the subject line sometimes acts as an indication of relationship between news postings; the string *Re:* is usually prepended to a followup and may in some cases be the only indication that the text was prompted by another. The Internet message header has a field for a *References:* relation but this is occasionally empty due to faulty software or inexperienced users.

Some authors also noted that tails of long threads were usually less interesting as the debate usually diverted from the original subject or degraded into nit-picking between factions.

3.9 Attachments (5)

Five authors remarked that the type and quantity of attachments to a MIME-type message could help in the classification of the message's importance.

3.10 Priority (3)

Some authors had evidently been in contact with other mail-systems where explicit priority could be assigned by the sender. This appears to be an obvious source of information but just like many other properties affixed by the sender it is not safe from abuse.

3.11 Layout (2)

Two authors placed emphasis on the layout of the message. The layout gives information about the nature of the message; if its written in haste of after careful editing. The authors may have had WWW-pages in mind, and while not a layout language, HTML markup gives a lot of information about the expected appearance of the page.

4 Final comments

A number of authors suggested that a combination of techniques and properties should be used for classification. Considering that many of the authors wrote with their personal experience in mind, it should be safe to say that people use whatever information they have at hand and form heuristics around that. For example, most email clients display the sender, the subject and the length of the message. It is therefore not surprising that these properties are mentioned so frequeuntly in the essays. Similarly, no-one appears to have used keyword search to sort their email, but it is a convenient way to find information in news archives and on the WWW.

On a more general level, the attention given to email by the authors appears guided by two things, the commitment expected of them and crucial time information, ie deadlines. Commitment is not easily quantified, but many cite a letter from their "boss" as an example of something which is given the highest priority. It may be possible to model the expected commitment using categories of senders and receivers.

In order to pursue computer-aided classification of email, one could allow the user to formalize his or hers personal heuristics in the email client, in the form of simple rules. Like one author suggested, the rules would not have to be internally consistent, and each rule is allowed to vote on each message. The sum of the votes is the final classification.