Do Users Tolerate Errors From Their Assistant? Experiments with an E-mail Classifier

Jean-David Ruvini **Jean-Marc Gabriel**
E-Lab Bouygues
1, avenue Engène Freyssinet
78061 St Quentin en Yvelines – France
jdruvini@bouygues.com jmgabriel@bouygues.com

Abstract
Smartlook, an e-mail classifier assistant, helps users filing their e-mails into folders. For a given message, it predicts the six most likely folders for that message and provides shortcut buttons that facilitate filing into one of the predicted folders. In this paper, we report results from user tests that show that although Smartlook does not achieve 100% prediction accuracy, a small percentage of errors does not hurt since users tolerate some errors from such an assistant.

Keywords
E-mail classification, personal assistant, usability.

INTRODUCTION
As e-mail is becoming increasingly important in every day life activity, mail reader users spend more and more time organizing and classifying the e-mails they receive into folders. Smartlook is an assistant for Microsoft Outlook 2000 that aims at decreasing the workload of hierarchically organizing and filing messages into folders. For a given message, it uses a text classifier to predict the six most likely folders for that message and provides shortcut buttons that facilitate filing into one of the predicted folders.

However, the goal of this paper is not to propose another e-mail classifier assistant. Although there have been a lot of work in the area of personal assistants, there is still no evidence that end-user are willing to use them [3]. The goal of this paper is to clarify this issue through experiments with an e-mail classifier assistant.

OVERVIEW OF SMARTLOOK
As most today's mail readers, Outlook 2000 allows to store messages in hierarchically organized folders. To file messages in folders, the user can move them manually, which can be tedious and error prone if there are many folders, or can write rules to automatically file messages into folders. These user defined rules are very powerful but are generally tedious to write and do not evolve with user filing habits.

IUI'02, January 13-16, 2002, San Francisco, California, USA.

Figure 1 shows how Smartlook facilitates the task of filing messages. When the user clicks on a message, it predicts the three folders where the message is most likely to be filed and offers shortcuts to file it into one of these folders. If one of the predicted folders is correct, the user just has to mouse-click on the corresponding button to quickly store the message in that folder. Of course, the user is free to ignore Smartlook's suggestions and to manually file the message. If the user clicks again on the same message, Smartlook deletes the suggestions and displays three more suggestions (the 3 next most likely folders).

Smartlook is an Outlook re-implementation of SwiftFile [2]. Like in SwiftFile, suggestion buttons are ordered from left to right, the leftmost button displaying Smartlook's best guess, the middle button the second best guess and so on. However, unlike SwiftFile, it is able to display 6 suggestions (2 sets of 3 suggestions).

Smartlook uses machine learning techniques to classify e-mails into folders. Smartlook's learning engine is the Rainbow text classifier [1]. In our context the training documents are the user's e-mails pre-classified into the user defined folders, represented by a bag-of-words after eliminating stop words.

EXPERIMENTAL RESULTS
The motivation of the experiments we have conducted was to evaluate Smartlook user satisfaction by comparing its actual prediction accuracy with the users' estimation of this accuracy.

We have evaluated actual prediction accuracy through a classical cross validation approach on the mail archives of 12 users (see Table 1) for two releases of Smartlook, respectively based on the Naïve bayes classifier (Table 2) and a Kuback-Leiber (KL) divergence based method using Witten-Bell smoothing (Table 3). Column "1 guess" presents prediction accuracy when Smartlook suggests only one folder, columns "3 guesses" and "6 guesses" present prediction accuracy over 3 and 6 suggestions respectively. Of course we have tested many other algorithms but we found no significant differences between the performances of Naïve bayes, TFIDF and KL.

Users' estimate of Smartlook first release prediction accuracy (for 3 guesses, see Table 2) has been evaluated through a user test by asking directly to users to give their estimate after two months of real use of the Smartlook. All our users are research engineers in our laboratory, between 25 and 36 years old.

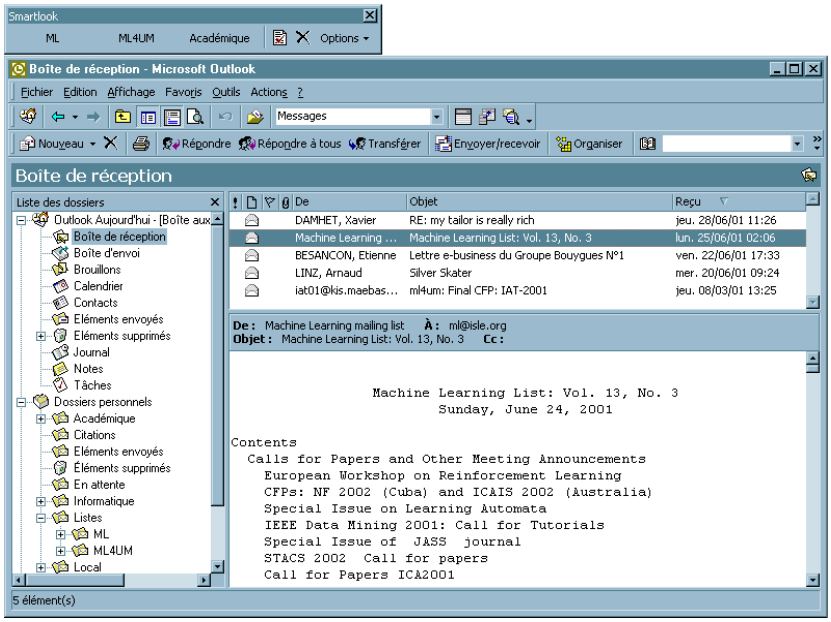

Figure 1. The user has selected a message in the inbox folder and Smartlook suggests to file it in the "ML" folder where machine learning related messages are usually stored. The user just has to click on the corresponding button to file the message in one of the predicted folder.

User	Messages	Folders
1	6621	362
2	5044	146
3	4090	102
4	2001	60
5	1628	93
6	764	23
7	655	34
8	639	34
9	643	12
10	602	20
11	429	24
12	312	30

Table 1. Mail archives used in the experiments.

Naïve bayes			User
1 guess	3 guesses	6 guesses	
66.31±.13	77.35±.13	81.47±.21	-
57.36±.18	65.86±.17	69.54±.18	40
52.30±.12	65.74±.12	69.77±.13	80
57.92±.19	66.71±.17	73.25±.14	95
47.45±.24	58.29±.22	64.07±.21	80
71.82±.23	82.09±.20	86.11±.22	90
50.48±.36	67.78±.30	73.80±.25	80
55.25±.43	68.66±.38	73.72±.34	50
84.25±.24	90.13±.20	94.89±.15	90
84.48±.15	90.17±.16	92.82±.16	85
86.33±.18	89.55±.19	90.26±.21	100
61.27±.35	72.24±.32	78.60±.34	-

Table 2. Smartlook's first release actual prediction accuracy and user estimate of this accuracy (for 3 guesses).

KL + Witten-Bell		
1guess	3 guesses	6 guesses
76.02±.11	83.43±.07	85.16±.12
76.29±.10	83.37±.09	85.02±.10
76.03±.11	83.83±.12	85.71±.12
72.03±.14	79.97±.12	82.75±.13
62.22±.20	73.48±.21	77.39±.19
79.04±.26	88.71±.20	90.91±.19
59.00±.26	72.09±.29	76.75±.30
69.06±.33	79.57±.30	82.54±.29
85.86±.82	93.19±.43	94.02±.39
90.48±.18	94.61±.15	95.23±.16
90.00±.10	94.22±.14	95.47±.13
71.95±.36	79.72±.35	85.59±.33

Table 3. Smartlook's last release actual prediction accuracy.

Except user 10 who is the designer of Smartlook, none of our users has skills or knowledge in the area of machine learning.

These experiments show that: (1) Smartlook last release prediction accuracy is above 80% for most users (even for user 1 who has a lot of folders); (2) using a large number of guesses if preferable and (3) users have over-estimated Smartlook's performance. Users also stated that using Smartlook reduces by 25% the time they spend every day in managing their e-mails.

DISCUSSION

The fact that users over-estimate Smartlook's performance is quite surprising and satisfying. It suggests that as far as an assistant achieves reasonable performances, users tolerate errors from it. This is an encouraging finding for predictive interface designers since building a predictive model of users that makes no error (100% accuracy) is rarely possible.

REFERENCES

[1] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996. www.cs.cmu.edu/~mccallum/bow.

[2] R. B. Segal and J. O. Kephart. Incremental Learning in SwiftFile. In Proc. of the 17th International Conference on Machine Learning, pages 863-870, Morgan Kaufmann, San Francisco, CA, June 2000.

[3] B. Shneiderman. Looking for the Bright Side of User Interface Agents. Interactions 2(1), pages 13-15, ACM Press, 1995.