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 email 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. Copyright 2002 ACM 1-58113-459-2/02/0001...\$5.00. 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 emails 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

nartlook			
ML ML4UM Acadé	mique 🛃 🗙 Options -		
S Boîte de réception - Microsoft O	utlook		
Fichier Edition Affichage Favoris O	utils Action <u>s ?</u>		
🥸 🗢 • 🔿 🔁 🔳 🔂 י	🗠 🍛 Messages 🔹 💽 🚰 🏹 🗸		
🔁 Nouyeau 🗸 🗙 🛃 🚱 Régond	re 🕵 Répondre à tous 🦃 Transférer 🛛 📑 Envoyer/recevoir 🖓 Organiser 📴	• *	
Boîte de réception		6	
Liste des dossiers ×	t D 🕅 🕼 De Objet Reçu ⊽	^	
🖃 🧐 Outlook Aujourd'hui - [Boîte aux 🔺	DAMHET, Xavier RE: my tailor is really rich jeu. 28/06/01 11:26		
Boîte de réception	😑 Machine Learning Machine Learning List: Vol. 13, No. 3 lun. 25/06/01 02:06		
- 🎯 Boîte d'envoi	BESANCON, Etienne Lettre e-business du Groupe Bouygues N°1 ven. 22/06/01 17:33		
- 🚯 Brouillons	LINZ, Arnaud Silver Skater mer. 20/06/01 09:24		
- 🙆 Calendrier	iat01@kis.maebas ml4um: Final CFP: IAT-2001 jeu. 08/03/01 13:25		
Contacts		7	
Carter de la construcción de	De: Machine Learning mailing list À: ml@isle.org		
	Objet : Machine Learning List: Vol. 13, No. 3 Cc :		
Notes		-	
Tâches	Machine Learning List: Vol. 13, No. 3 Sunday, June 24, 2001		
∃ - 🧐 Dossiers personnels			
Gadémique			
Citations			
- 1 Eléments envoyés	Contents		
- 🔞 Éléments supprimés	Calls for Papers and Other Meeting Announcements		
- 🔞 En attente	European Workshop on Reinforcement Learning		
🗄 🍘 Informatique	CFPs: NF 2002 (Cuba) and ICAIS 2002 (Australia)		
🖻 🍘 Listes	Special Issue on Learning Automata		
⊕-@2 ML	IEEE Data Mining 2001: Call for Tutorials		
🕀 í ML4UM	Special Issue of JASS journal		
🛨 🏠 Local 📃	STACS 2002 Call for papers Call for Papers ICA2001		
		•	

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.

KL + Witten-Bell			
lguess	3 guesses	6 guesses	
76.02±.11	$83.43 \pm .07$	85.16±.12	
$76.29 \pm .10$	$83.37 \pm .09$	$85.02 \pm .10$	
76.03±.11	$83.83 \pm .12$	85.71±.12	
$72.03 \pm .14$	79.97±.12	82.75±.13	
$62.22 \pm .20$	73.48±.21	77.39±.19	
$79.04 \pm .26$	88.71±.20	90.91±.19	
$59.00 \pm .26$	$72.09 \pm .29$	$76.75 \pm .30$	
69.06±.33	79.57±.30	$82.54 \pm .29$	
$85.86 \pm .82$	93.19±.43	$94.02 {\pm}.39$	
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 1. Mail archivesused in the experiments.

Messages

6621

5044

4090

2001

1628

764

655

639

643

602

429

312

Folders

362

146

102

60

93

23

34

34

12

20

24

30

User

1

2

3

4

5

6

7

8

9

10

11

12

Table 2. Smartlook's first releaseactual prediction accuracy and userestimate of this accuracy (for 3guesses).

Naïve bayes

3 guesses

77.35±.13

 $65.86 \pm .17$

 $65.74 \pm .12$

66.71±.17

 $58.29 \pm .22$

 $82.09 \pm .20$

 $67.78 \pm .30$

 $68.66 \pm .38$

90.13±.20

 $90.17 \pm .16$

 $89.55 \pm .19$

 $72.24 \pm .32$

6 guesses

81.47±.21

 $69.54 \pm .18$

69.77±.13

73.25±.14

 $64.07 \pm .21$

86.11±.22

 $73.80 \pm .25$

 $73.72 \pm .34$

 $94.89 \pm .15$

 $92.82 \pm .16$

90.26±.21

 $78.60 \pm .34$

1 guess

66.31±.13

 $57.36 \pm .18$

 $52.30 \pm .12$

 $57.92 \pm .19$

 $47.45 \pm .24$

71.82±.23

 $50.48 \pm .36$

 $55.25\pm.43$

 $84.25 \pm .24$

 $84.48 \pm .15$

86.33±.18

 $61.27 \pm .35$

Table 3. Smartlook's lastrelease actual predictionaccuracy.

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

User

40

80

95

80

90

80

50

90

85

100

- 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.