A Hybrid User Model in Text Categorization

Sanghee Kim, Wendy Hall
Intelligence, Agents, Multimedia Group
Department of Electronics and Computer Science
University of Southampton, U.K.
44-(0)23-80-59-3256
{sk98r, wh}@ecs.soton.ac.uk

Andy Keane
Computational Engineering Group
Mechanical Engineering
University of Southampton, U.K.
44-(0)23-80-59-2944
ajk@soton.ac.uk

ABSTRACT
A user model that specifies user preferences on message handling is an essential component of an e-mail message categorizer. We present an approach that combines two learning algorithms, i.e. the Naïve Bayesian Classifier (NBC) and Progol, to model implicitly and explicitly reflected user preferences that may not be modeled by using either the algorithms alone. An experiment demonstrates the improvement of categorization performance compared to that of using the two algorithms independently.

Keywords
Text categorization, user modeling, symbolic learning, statistical learning, e-mail communication

1. INTRODUCTION
Adaptive text categorization draws from user modeling and text filtering studies. A user model provides a knowledge base upon which a system can perform categorization of texts. Text filtering actively monitors uncategorized texts, and evaluates them into proper categories.

Reviews of current techniques show that user models are often constructed from statistical descriptions of user defined categories. While these methods perform reasonable categorization, however, the performance can be improved by taking into account user's cognitive attitudes, such as the user's preferred method of text organization. The user's strategies for text organization function as a knowledge base describing how each user employs domain-specific characteristics in organising relevant texts to his/her convenience. Knowing how each user perceives the same texts differently is important in the design of an effective categorizer, since this can improve categorization accuracy.

In this paper, we present an approach that combines two learning algorithms, i.e. the Naïve Bayesian Classifier (NBC) and Progol, as an extension of a user model in the e-mail message categorizer. In our presumed domain, the e-mail message is composed of a set of attributes, e.g. the author's name, the primary recipient's name(s), the secondary recipient's name(s), subject, and content, and each of these contributes to the categorization task. The strategies are converted into the preferential value of each attribute, which explains how the user relies on the specific attributes to categorize the e-mail messages.

2. E-MAIL MESSAGE CATEGORIZER
An e-mail management system is a good information source to experiment with adaptive text categorization. E-mail messages are organised by an individual to suit his/her convenience. A user might prefer to organise the messages by the names of senders, or by similar topics in contents. More knowledgeable users may employ quite complex or context-sensitive criteria, for instance two messages that are sent by the same author are filed into separate categories, depending on the subject matter discussed. The number of categories is diverse among users: a socially active user might maintain more than 100 categories. Message contents can be anything: greetings from a friend or a notification of new changes concerning an ongoing project. Many users are keen to manage their e-mail messages well, so that they can concentrate on useful and important messages while filtering out less- or not- relevant ones. [3] saw this problem as the cause of economic loss associated with the examination of non- or less-relevant information, and thus makes the users have less time to concentrate on more important information.

3. A HYBRID USER MODEL
Our categorizer is a hybrid user model, where the initial step extracts the user preferences by deriving significant attribute values of our presumed domain. These derived values are fed into statistical models that define informative features (i.e. keywords or e-mail addresses) of each attribute. Progol, as a symbolic learner, performs well on qualitative reasoning, and can discover hidden relations between attributes and a given category [6]. NBC is a statistical learner and has the capability to describe a given category as a collection of features associated with weighted importance. The hybrid user model extends NBC to accept prior probabilities gathered by Progol.

Specific to Progol, a coverage number defines the number of messages correctly satisfied (covered) by a generated rule, and can be used for describing the strength of an attribute [5]. In our system, Progol generates the common patterns of user preferences by comparing normalised coverage values of each attribute.

NBC is a simplified Bayeisan theorem which assumes that features are independent given a category [4]. The prior probability of each category calculated by NBC provides an initial value for a promising hypothesis, e.g. the most appropriate category for a new message. Instead of estimating the prior probabilities based on the proportion of category c_j (i.e. the number of the messages in c_j divided by the total number of the messages across categories), in the hybrid model the prior probabilities are measured by Progol as the preferential values of each attribute in category c_j that are derived from the coverage values.

4. EXPERIMENTAL RESULTS

Table 1 shows the details of data sets that were collected from four users. 70% of the messages from each category were chosen randomly to create a classifier (training set), and the remaining 30% (testing set) was used for testing. This was done five times to create five training and testing sets from the original data sets. The value of each data set is averaged over these five testing sets.

Table 1: Data sets

Data set	#Folders	#Messages
data set1 (user1)	21	3316
data set2 (user2)	11	884
data set3 (user3)	10	1320
data set4 (user4)	5	2050

Table 2 shows the performance results of the hybrid user model measured by precision, recall, breakeven point, and error rate [7].

Table 2: Performance results

Data set	Precision	Recall	Breakeven point	Error rate
data set1	0.95	0.83	0.86	0.1524
data set2	0.96	0.88	0.90	0.1030
data set3	0.94	0.75	0.85	0.0822
data set4	0.96	0.90	0.92	0.0387

When compared to the results on e-mail filtering by [2], the error rate in Table 2 is not better than that of the RIPPER'S performance (i.e. approximately 0.0641), but lower than that of the TFIDF algorithm. [1] experimented with nearest neighbor and neural network algorithms on categorizing e-mail messages. The results of precision and recall in Table 2 are slightly lower than those of the nearest neighbor and neural network algorithms, which are close to 98% accuracy. However, as the two algorithms are either slow to categorize or slow to train, these may be inappropriate for practical approaches. We have implemented a pilot study which investigates whether our categorizer could be used in a real environment. This study was carried out for 10 days, and all subjects have used our categorizer for sorting new incoming e-mail messages. It is observed that the categorizer recognized the right categories within a few seconds, and it needed only a small number of messages (i.e. 5-10) for training.

In order to compare to the performance of the existing algorithms, we run NBC and Progol on the same data sets. NBC was tested without the prior probability of each category. The hybrid method obtained slightly better performance, on average approximately 0.05 precision and 0.03 recall compared to the NBC. Progol obtained higher and comparable precision than that by the hybrid and by the NBC respectively. This high precision can be seen from Progol which generates features that are most commonly found in category c_j but not in other categories, so this makes the outcomes to be highly accurate.

5. CONCLUSION

The proposed approach has achieved slightly better performance compared to the NBC and Progol. In practice, combing the two algorithms does not slow down run-time performances, since the two algorithms are fast. This also allows the categorizer to pay attention to the more useful attributes while ignoring non-relevant attributes, thus making the classifier efficient. In addition, Progol is set up in advance, and is re-tested only on a revision process, instead of being run at every prediction time.

6. ACKNOWLEDGMENTS

The work presented here has been supported by UTP (University Technology Partnership in Design) sponsored by Rolls Royce Plc and BAE Systems.

REFERENCES

[1] G. Boone, “Concept Features in Re: Agent, and Intelligent Email Agent”, *In Proceedings of the Second International Conference on Autonomous Agents* ,141-148, (1998)
[2] W. Cohen, “Learning Rules that Classify E-Mail", *In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, (1996)
[3] R. Losee, "Minimizing Information Overload: The Ranking of Electronic Messages", *Journal of Information Science*, 15(3), 179-189, (1989)
[4] T. Mitchell, “Machine Learning”, McGraw-Hill International Editions, (1997)
[5] S. Muggleton, “Inverse entailment and Progol”, *New Generation Computing*, 13, 245-286, (1995)
[6] C. Sammut, “Using Background Knowledge to Build Multistrategy Learners”, *Machine Learning*, 27(3), 241-257, (1997)
[7] Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization”, *Journal of Information Retrieval*, 1(1/2), 67-88, (1999)