AUTOMATIC EMAIL CLASSIFICATION USING USER PREFERENCE ONTOLOGY

Niladri Chatterjee
Department of Mathematics, Indian Institute of Technology, Delhi, India

Saroj Kaushik
Department of Computer Science and Engineering, Indian Institute of Technology, Delhi, India

Smit Rastogi, Varun Dua
Department of Mathematics, Indian Institute of Technology, Delhi, India

Keywords: Email classification, Ontology, Automatic email categorization, Inference engine, Feature weighing.

Abstract: In this work we have extended and implemented an ontology based approach for email classification based on user characteristics proposed by Kim et al.(2007). The approach focuses on finding relationships between user interests and their responses to emails. Rules and Ontology are created using the data and metadata of user characteristics, their preferences and responses to emails. Rules and ontology are then used to predict the response of a user to a new email. In Kim et al. (2007) approach, labels to emails were provided manually by a human expert. We have endeavored to remove the human intervention by developing an Automated Email Categorizer to provide label to an email based on its contents. We have also proposed a new term weighing method for emails to incorporate prominence of subject terms. Finally, we have integrated and tested the Ontology Based Classifier in conjunction with Email Categorizer where the former effectively uses the label provided by latter to classify an email based on user preferences.

1 INTRODUCTION

Email has emerged as an efficient and popular communication mechanism in recent times. But one of the major problems is the huge number of unsolicited/spam mails that fill in user's inbox regularly. Consequently, email management has gained attention and has become an important problem for individuals and organizations.

One major aspect of email management is classification of emails. In this context the terms email classification and email filtering are often used interchangeably. Filtering may be considered as a subclass of classification with just two classes (Itskevitch, 2001); while one of its applications being spam filtering.

Several Email-filtering techniques have been developed so far. The successful ones are either rule based or statistics based (Zhang and Yao, 2003). Recently some attempts have been made towards ontology based email-filtering that allows machine-understandable semantics of data (Youn and Mcleod, 2006).

(Brewer et al., 2006) have suggested a system that uses ontology to discover relationships between tokens in an email thus utilizing semantics of an email as an additional parameter for classification. A framework for email filtering using ontology, based on frequency count of words in an email, is proposed in (Youn and Mcleod, 2007). The counts of various words occurring in spam/non-spam training emails were used to create a decision tree for classification. The decision tree formed was represented in the form of ontology using RDF. A new email is then classified using this ontology.

Several supervised learning methods based on statistical analysis have been applied to the email filtering task (Zhang et al., 2004). However, the major drawback with all the statistical approaches is that they do not incorporate the interests of a specific

Chatterjee N., Kaushik S., Rastogi S. and Dua V..
AUTOMATIC EMAIL CLASSIFICATION USING USER PREFERENCE ONTOLOGY.
DOI: 10.5220/0003061501650170
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 165-170
ISBN: 978-989-8425-29-4

user towards particular types of mails. Content based classifiers cannot incorporate preferences and contexts to take a particular feature in an email as a parameter for classification. Users behaviours to emails vary according to the different personal preferences. Therefore it is meaningful to provide user-oriented classifier based on the preferences.

An email filtering scheme based on the personalized ontology to classify an email as spam or non-spam, has been proposed in (Youn and Mcloed, 2009)). They created a personalized ontology by combining a user profile ontology and a structured taxonomy. The user profile ontology creates a black-list of contacts and topic words while the structured taxonomy is used as a global ontology filter.

An ontology based email-filter incorporating the preferences of a user has been proposed by (Kim et al., 2007). A user preference ontology is built to formally represent the important concepts and rules derived from a data mining process. Then an inference engine is used that utilizes the knowledge to predict the user's action on new incoming emails. This approach utilizes a label given manually to each email, based on its content. The labels are from a predefined set of categories, and represent the content of the email in the ontology and the data mining process. The major drawback of this approach is that the label has to be provided manually by a human expert. However, to read and provide a label to an email is a very tedious task.

In this paper we have extended the approach suggested by (Kim et al., 2007) by proposing an improvement in terms of automatically providing a label to each email through a text categorizer. We have built an email-categorizer which when given an email as input, outputs a label for that email. The ontology based classifier uses this label along with the user preferences and characteristics to predict the response of the user for the email. For the text categorizer, we have also proposed a new term weighing method specifically applicable to emails which enhances the weights of terms depending on their position in the email. However, our major contribution has been in terms of the categorizer and usage of its result in classification process, along with some improvements in the classification process (Table 2 and section 4.3) as compared to that originally suggested by (Kim et al., 2007).

The paper is organized as follows. In Section 2 we describe the overall system design. Section 3 explains the implementation of the scheme including essential sub-components, such as processing of emails, data collection, rule generation, ontology construction and final classification procedure. Section 4 provides analysis of the results and comparison with the original system. Finally, Section 5 concludes this paper.

2 OVERALL SYSTEM DESIGN

We have assumed that broadly users might have two responses to an email: Interested or Not-interested. The proposed architecture for the system is shown in Figure 1.

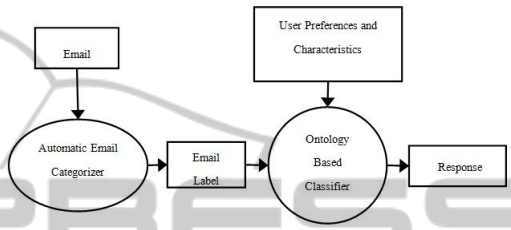

Figure 1: Architecture of the proposed system.

The system has two main components: Automatic Email Categorizer and Ontology Based Classifier. To predict the response of a user for any email, first the label for that email is generated by the Categorizer based on email's contents. The Classifier takes that label along with the preferences of the user as input to predict the response of that user to that email. The Categorizer has been trained and tested independently for various parameters, the details of which are presented in section 3.1. The architecture of these two subsystems is explained in the following sub-sections.

2.1 Automatics Email Categorizer

Architecture of the Automatic Email Categorization (AEC) is shown in Fig. 2. AEC assigns label to a new email based on its contents. Each component of the classifier is briefly explained below.

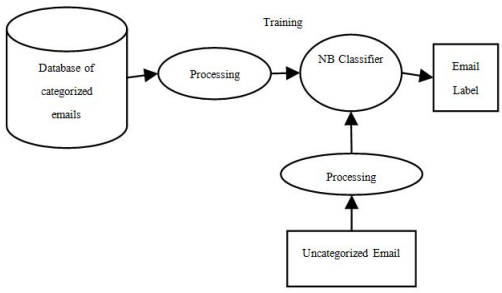

Figure 2: Architecture of Automatic Email Categorization System.

Database consists of Ken Langs 20newsgroups dataset (Rennie, 2010) which contains approximately

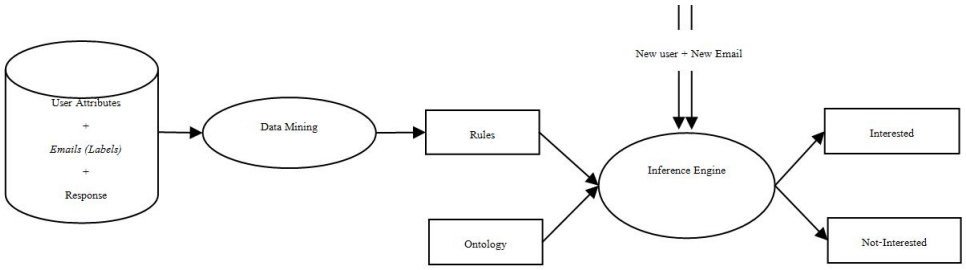

Figure 3: Architecture of the Ontology based Classification System.

20000 messages taken from 20 different netnews newsgroups partitioned (nearly) evenly across 20 categories.

Initially, the AEC system is trained using Nave Bayesian Classifier with labeled emails which are represented by feature vectors containing important terms. Once the system is trained, a label to a new unlabeled email is predicted.

2.2 Ontology based Email Classifier

The ontology based classifier uses the label of new email provided by AEC system along with user preferences and finally classify whether the user is interested in this mail or not. The architecture of the ontology based classifier is shown in Fig. 3. The ontology is constructed (using OWL-DL) to represent the metadata involving user characteristics, email category and response of a user towards an email.

A database of user characteristics (gender, age etc), their preferences and their responses to sample emails (interested or not-interested) is created through a survey from different users. In Fig. 3, User Attributes represent the characteristics and preferences of various users. Email corresponds to the email for which a Response from the user has been sought for training the system. The details are given in Section 3. After building the database, we mine it using the WEKA tool to find correlations between the preferences of users and their responses to emails and if-then Rules are generate. The steps involved in rule generation are explained in detail in section 3.2.

The rules and the ontology are loaded in JENA inference engine (a Java framework for building Semantic Web applications) and the system predicts response to a new email for the user. An accuracy of the system is calculated by comparing the predicted results of test emails with the actual responses. Section 4 has the details.

3 IMPLEMENTATION DETAILS

In this section we describe briefly the implementation of both the components namely, Automatic Email Categorizer and Ontology Based Classifier with Rule generation.

3.1 Automatics Email Categorizer

We have used supervised learning approach for building proposed categorizer where we have utilized the structure of emails to propose a new method for term weighing which works visibly better for emails as compared to other traditional term weighing methods. As a result our proposed method outperforms the traditional term weighing methods (TF, TF-IDF).

3.1.1 Feature Selection

In our system we used only the contents of subject and body of the documents from 20newsgroup. These contain all the necessary informative terms. Non-contextual/optional words are removed using a stop-list. Stemming has been performed using Porter's algorithm.

To further reduce the size of feature space, we used several feature selection methods and analyzed their performances. We used the term frequency (TF) and the document frequency (DF) feature selection methods both for retaining the most important features. We have also used Mutual Information (MI) and Chi-square feature selection methods. MI and Chi-square methods, although computationally expensive, yield significantly better performances. As observed from Table 1, the top five terms selected by Chi-square and MI methods are more relevant to windows category than the terms selected by DF and TF.

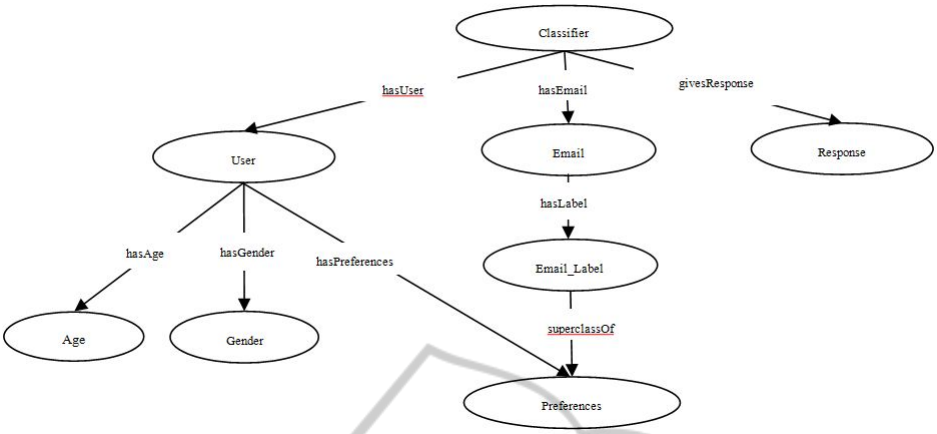

Figure 4: Structure of the ontology.

3.1.2 Weights of Features

Weight of a term t in document d is calculated by the following formula (Eq. (1)).

$$\omega_{td} = (n_b + An_s) \log (D/df_t \times (tf_{td}/T_d)/(tf_t/T)) \quad (1)$$

Here, n_b is the number of occurrences of the term in body, n_s is the number of occurrences of the term in the subject, D denotes the number of emails in the training set, df_t is the document frequency of the term t. tf_{td} is the number of occurrences of term t in document d and T_d is the total count of all the terms in document d. Similarly tf_t is the number of occurrences of term t in the entire training set and T is the total count of all the terms in training set. 'A denotes the adjustment value according to the importance of the subject terms in the email.

We estimate the adjustment value by experiments. Experiments show that the proposed term weighing method significantly enhances the performance of the categorizer.

Table 1: Top 5 features of different categories using different feature selection methods in category comp.os.ms-windows.misc.

Chi-sq	MI	DF	TF
nt	directori	m	window
app	load	microsoft	File
setup	product	support	Do
mous	print	univers	Re
desktop	font	read	Write

3.2 Ontology based Email Classification

We merged some of the categories, in the 20newsgroup dataset, that were similar to each other and reduced 20 categories to 8 categories namely, *Sports, Automobiles, Religion, Advertisements, Computers, Politics, Science, Medicine*. We selected 50 emails belonging to these categories. User attributes and preferences, along with their responses to these 50 selected emails from 30 users were also collected. We created a user profile format to represent the user interests and their responses to emails in binary. Fig. 4 shows the structure of ontology that captures the metadata of user preferences and their responses.

Email_Label is the label consists of Interested or Not-Interested. The rest of the attributes describe the interests of a user towards a particular category. Nodes in the ontology are self explanatory.

A decision tree was trained to discover the association rules between various groups of users and their responses by the sample email data. We used WEKA API for the implementation of C4.5 decision tree. The Rules generated were of following form: *If Email_Label = 'sports & sports = 'False' & gender = 'Male & computers = 'True then 'Interested.*

The developed ontology and the rules are used to predict the response of a user for a new email.

4 EXPERIMENTS, RESULTS AND ANALYSIS

Experiments were performed to test the accuracy of both: 1) Email Categorizer which provides a label to emails 2) Ontology Based Classifier which predicts the response of a user towards an email based on user characteristics, preferences and the label of test email. In this section, first we present the results and analysis of the Categorizer along with the performance of the proposed term weighing method. Finally we present the results of the Ontology Based Classifier and its comparison with original scheme.

Table 2: Comparison of the proposed system with Kim's scheme.

S. No.	Aspect of the Classifier	Original Scheme by Kim et al. (2007)	Proposed Scheme
1	Emails used	Korean emails collected	English emails from 20news-group
2	Email Label	Manually provided	Automatically generated
3	Data collection for Ontology based Classifier	40 emails-90 students	50 emails-30 students
4	Responses sought from users	Inbox/Reply/Delete/Spam	Interested/Not-Interested
5	Decision tree used	ID3	C4.5
6	Ontology language	Web-PDDL	OWL-DL
7	Reasoner used for response prediction	OntoEngine	Jenas Rule based Inference Engine
8	No. of rules generated	89	42
9	Overall Accuracy	60.1%	67.6%

4.1 Analysis of Categorization System

Fig. 5 shows the performance of several methods used for feature space reduction using subject-based term weighing method. The number of features for each method has been set to 20000. As observed from the graph, the Chi-square method outperforms all other methods in dataset used. We have compared the

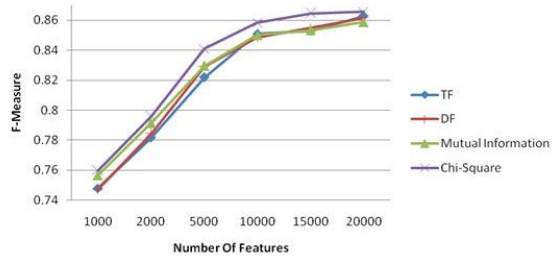

Figure 5: Performance of various feature selection methods.

performance of proposed Subject-based term weighing method with traditional term weighing methods (TF and TF-IDF) over MI and Chi-square feature selection methods. The proposed weighing method considerably outperforms TF and TF-IDF weighing methods over all the feature selection methods experimented with. Fig.6 and Fig.7 display the results of the experiments conducted with MI feature selection and Chi-square feature selection methods respectively. The adjustment value A in our proposed term weighing method (Eq 1) was taken as 5 while comparing the term weighing methods. The effect of variation in value of A over the categorizer performance was also analyzed.

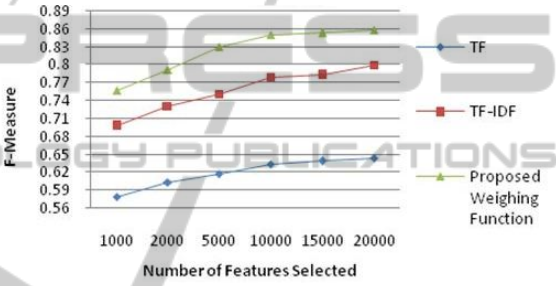

Figure 6: Comparison of term weighing methods (using Mutual Information feature selection method).

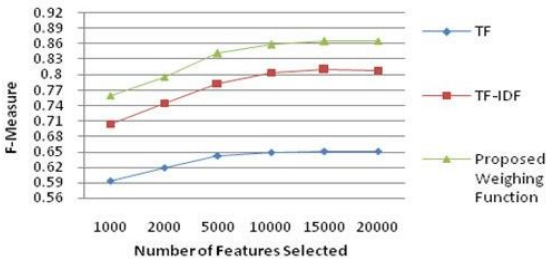

Figure 7: Comparison of term weighing methods (using Chi-square feature selection method).

4.2 Analysis of Ontology based Classifier

Table 2 presents a comparison of our system with the classification system developed by (Kim et al., 2007). We outlined the differences in the methods, implementations and results, an outright comparison in terms of accuracy being not possible since the datasets used are different. Kims system used Korean emails for experimentation while our system works for English emails. The overall ontology based scheme we have proposed is on similar lines as (Kim et al., 2007) but with a few modifications and a few marked dif-

Table 3: Axioms derived by C 4.5 Decision Tree mining and their performance.

S. No.	Rules	Accuracy(%)
1	Email_Label = 'sports' & sports = 'False' & gender = 'Male' & computers = 'True' then Interested	100
2	Email_Label = 'religion' & religion = 'False' & medicine = 'True' & age = 'Junior' then NotInterested	83.3
3	Email_Label = 'politics' & age = 'Junior' & religion = 'False' & computers = 'True' & gender = 'Male' then Interested	83.33
4	Email_Label = 'sports' & sports = 'False' & gender = 'Female' then NotInterested	77.78
5	Email_Label = 'religion' & religion = 'False' & medicine = 'True' & age = 'Senior' & gender = 'Female' then NotInterested	66.67

ferences in implementation (as outlined in Table 2), the major improvement being the automatic labeling of emails by a categorizer.

As proposed by (Kim et al., 2007), Rule Accuracy or Rule Confidence can be used to calculate correctness of each rule in the ontology. For our classification scheme, this score can be a metric based on confidence of the Rule because more is the confidence of Rule, more is the probability that the Response predicted is correct. Some of the rules along with their accuracy are listed in Table 3. The overall accuracy of our system was found to be 67.6%.

5 CONCLUSIONS

We have extended, and implemented a method for email classification originally proposed by (Kim et al., 2007). The salient feature of the approach is that it incorporates user preferences towards the final classification decision. Another important feature is utilization of a user preference ontology to classify emails. We have removed human intervention from the task of providing label to each email manually, by integrating the ontology based classification system with an automated email categorization system. The limited amount of data collected might have been a constraint towards getting clear correlation between user preferences and their responses since the collected data might have not captured all correlations.

This work also establishes the better performance of the proposed term weighing method over the conventional TF and TF-IDF methods when tested on emails. Various feature selection methods were also compared using the naive Bayesian classifier among which Chi-square method gave best results. However, since emails behave little differently than normal documents during categorization, incorporation of more heuristics and parameters might improve the accuracy of the categorizer and how the effect of any heuristic varies with change in categorization algorithms would be an interesting thing to observe.

REFERENCES

Brewer, D., Thirumalai, S., Gomadam, K., Li, K., 2006. *Towards an Ontology Driven Spam Filter*. In Proceedings of 22nd International Conference on Data Engineering Workshops.

Itskevitch, J. 2001. *Automatic hierarchical e-mail classification using association rules*. MS Thesis, Simon Fraser University.

Kim, J., Dou, D., Liu, H., Kwak, D., 2007. *Constructing A User Preference Ontology for Anti-spam Mail Systems*. In Proceedings of the 20th Conference of the Canadian Society For Computational Studies of intelligence on Advances in Artificial intelligence. Montreal, Canada.

Rennie, J. Ken Lang 2010. *20newsgroup dataset* http://people.csail.mit.edu/jrennie/20Newsgroups

Youn S., McLeod D., 2006. *Ontology Development Tools for Ontology-Based Knowledge Management*. Encyclopedia of E-Commerce, E-government and Mobile Commerce, Idea Group Inc.

Youn, S., Mcleod, D., 2007. *Spam Email Classification using an Adaptive Ontology*. In Proceedings of 4th International Conference on Information Technology: New Generations (ITNG). Las Vegas, NV.

Youn S., Mcleod D., 2009. *Spam Decisions on Gray E-mail using Personalized Ontologies*. In Proceedings of the 2009 ACM symposium on Applied Computing SAC09, Honolulu, Hawaii, U.S.A.

Zhang, L. and Yao, T., 2003. *Filtering Junk Email with a Maximum Entropy Model.* In ICCPOL03, Shen yang, China.

Zhang, L., Zhu, J., Yao, T., 2004. *An Evaluation of Spam Filtering Techniques*. In ACM transactions on Asian Language Information Processing, Vol.3 No.4