2005

Persian Email Classification Based on Rocchio and K-Nearest Neighbor Approach

H. Bashiri
Bu-Ali Sina University

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

A. Moeini
University of Tehran

Persian Email Classification Based on Rocchio and K-Nearest Neighbor Approach

Hassan Bashiri, *Department of Computer Engineering, Bu-Ali Sina University, email:* hbashiri@acm.org

Farhad Oroumchian, *Faculty of IT, Wollongong University in Dubai*, *email:* foroumchian@acm.org

Ali Moeini, *Faculty of Science*, *University of Tehran*, *email:* moeini@ut.ac.ir

These days, electronic mail (email) has become an essential form of communication in all aspects of every day life. The main reason for this popularity among other things like the speed of delivery and the low cost is the convenience of managing and handling emails. However this connivance is diminishing by the growth and availability of the emails. Managing emails is becoming more difficult every day. Not only SPAM (unsolicited emails) is flooding our mailboxes but locating important and vital information among the huge number of emails that are finding their ways into our mailboxes has turned into a laborious and time consuming daily activity. Because of this, we are witnessing an increasing interest in research and development of all sorts of different products and technologies that can remedy the email explosion problem.

This research is concentrated on developing a new Persian Email server with the ability of automatically classifying and organizing incoming emails into folders.
Eventually, the user interface (UI) of this system will provide advanced Email management facilities such as classification, prioritizing, topic detection and summarization along with the normal search, storage and retrieval of messages. The folders could be hierarchical, meaning that there would be folders within each folder. Making it easier to browse through folders and locate relevant information.

Because of lack of experimental Persian email test collection, a small collection of emails for University faculties have used in this project. A professor's emails have been divided into four standard categories. Each incoming emails will be placed in one of the below folders if possible. These categories are:

1. *Seminars category:*
   This category collects emails that are about conferences and seminars.
2. *Students category:*
   This category collects emails from students that are about courses, exam dates, grades, assignments, course projects and appointments.
3. *Co-workers category:*
   Include emails from other professors and co-workers and they are about work status.
4. *General category:*
   All other emails are collected in this category.

The emails in the general category are further divided into smaller groups by clustering algorithms.

Two different classification approach including KNN (K-Nearest neighbors) with different parameters of K and Rocchio algorithm [2, 4] have used in the experiment in order to find the useful and efficient classification

Organized by:
Islamic Students' Association in UK
Union of Islamic Students' Association in Europe
Scientific Representation of I. R. Iran in UK

algorithm. Also, adaptation has been studied in order to adjust the folder classifiers to personal preferences of users.

For the purpose of the experiment, all the words in the emails have been stemmed byBON Persian indexer and indexed [3]. The training set for the classifiers of each category contained at least 20 emails. For the evaluation purposes, random emails have been submitted to classifiers and their results have been validated. The test emails also were processed by the same indexer and BON stemmer as the training set.

Precision and Recall are two most used criteria to evaluate information retrieval systems [5]. In classification these modified according to table 1 which have shown blow [1].

Class of Ci		*Expert Judgment*	
		Yes	*NO*
Classifier Function	*Yes*	*TPi*	*FPi*
	No	*FNi*	*TNi*

$$\Pr_i = \frac{TP_i}{TP_i + FP_i}$$

$$\mathrm{Re}_i = \frac{TP_i}{TP_i + FN_i}$$

Table 1: Precision and Recall used to evaluate our email classifier

Our experiment shows that unlike most classifier Rocchio algorithm has better result of K-NN. Results for Rocchio and K-NN algorithm by K = 3, 5, 8 have shown in table 2. Precision-Recall diagram have been drawn in figure 1 too.

	Term weighting is tf.idf			
	Rocchio	*3-NN*	*5-NN*	*8-NN*
Precision	0.857	0.714	0.785	0.8
Recall	0.923	0.714	0.846	0.857

Table 2: Precision and Recall of Persian Email Classifier

Organized by:
Islamic Students' Association in UK
Union of Islamic Students' Association in Europe
Scientific Representation of I. R. Iran in UK

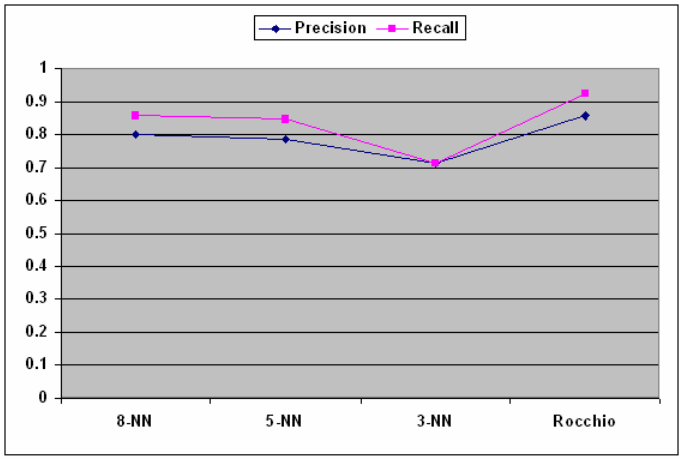

Figure 1: Precision-Recall diagram

Although this paper focuses on emails on faculty mailboxes, we assert these methods can be easily adapted to other domains such as managerial, military and other emails as well.

Key words: Classification, Persian Indexing, Email Classification, Persian Email system.

References:

[1] Sebastiani F., **"A Tutorial on Automated Text Categorization",** Istituto di Elaborazion dell'Informazione, 1999

[2] - Basili R., Moschitti A., **"A Robust Model for Intelligent Text Classification",** Department of Computer Science, 2001

[3] – M. Tashakori, M. Meybodi, F. Oroumchian, "**Bon: The Persian Stemmer."**, In EurAsia-ICT. pp. 487-494. 2002

[4] - Sebastiani F., "**Machine Learning in Text Categorization**", Journal of ACM computing Surveys, 2002

[5] - Baeza-Yates R., Ribeiro-Neto B., "**Modern Information Retrieval**", ACM Press, 1999