Assessing Classification Accuracy in the Revision Stage of a CBR Spam Filtering System

José Ramón Méndez[1], Carlos González[2], Daniel Glez-Peña[1], Florentino Fdez-Riverola[1], Fernando Díaz[3], and Juan Manuel Corchado[4]

[1] Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
{moncho.mendez,dgpena,riverola}@uvigo.es
[2] GFI Informatique, C/ Salvatierra 5, 28034, Madrid, Spain
cgperez@gfi.es
[3] Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
fdiaz@infor.uva.es
[4] Dept. Informática y Automática, University of Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain
corchado@usal.es

Abstract. In this paper we introduce a quality metric for characterizing the solutions generated by a successful CBR spam filtering system called SPAMHUNTING. The proposal is denoted as *relevant information amount rate* and it is based on combining estimations about relevance and amount of information recovered during the retrieve stage of a CBR system. The results obtained from experimentation show how this measure can successfully be used as a suitable complement for the classifications computed by our SPAMHUNTING system. In order to evaluate the performance of the quality estimation index, we have designed a formal benchmark procedure that can be used to evaluate any accuracy metric. Finally, following the designed test procedure, we show the behaviour of the proposed measure using two well-known publicly available corpus.

1 Introduction and Motivation

The greatest steps forward in the field of CIT (*Communications & Information Technologies*) were the Internet and mobile phone introduction. These have allowed the development of modern communication infrastructures which give the final user a wide range of ways to communicate, as well as the freedom of doing it everywhere. Moreover, some relationships have been established between these technologies and nowadays, users can have Internet access through their mobile phones [1] and talk or send SMS (*Short Message Service*) messages by using VoIP (*Voice over IP*) techniques [2].

Unfortunately, these technologies share the same problem: the massive dissemination of spam contents. Spam is not only present on the delivered e-mails. From a

R.O. Weber and M.M. Richter (Eds.): ICCBR 2007, LNAI 4626, pp. 374–388, 2007.

practical and broad perspective, spam can be viewed as a set of irritating techniques used for distributing information taking advantage of the newest communication technologies. Moreover, spam is generally unsolicited by those targeted. Spam communications can also be found in SMS messages, blogs (through commentaries of the posts), newsgroups, search engines, and of course, postal mail and e-mail messages. In this work, we are mainly concerned with the oldest form of spam: e-mails sent across the Internet.

Due to the exponentially increasing amount of spam messages transferred through Internet, several techniques have been introduced for fighting the delivery of spam messages. Although anti-spam filtering software is often classified as *collaborative* or *content-based* [3], most of the successful approaches are classical machine learning techniques with little adjustments for detecting and filtering spam e-mails [4].

In the context of content-based techniques, two innovative anti-spam filtering CBR models have been introduced during the last years. First of all, we highlight the relevance and accuracy of the results achieved by some well-known researchers from the Dublin Institute of Technology [5]. They have started a revolution on the spam filtering domain by introducing a successful CBR system called ECUE [6]. Recently, we have introduced a new way of filtering spam by using an innovative feature selection technique applied in the retrieval stage of our SPAMHUNTING CBR system [7]. All these previous successful results evidence how and why case-based reasoning is particularly suitable for classifying and filtering spam messages.

Among other deserving properties, one of the most relevant characteristics of CBR in the spam filtering domain is the possibility of generating a *null* solution. This situation is identified by CBR systems when not enough cases are recovered during their retrieval stage. Although in many domains a null solution is not suitable, in spam filtering domain it is equivalent to assert that the incoming e-mail is legitimate. These topics are supporting the quality of previous research works on CBR for spam filtering by Delany *et al* [5, 6] and Méndez *et al* [7, 8, 9, 10].

One of the most interesting issues for current techniques is the ability of evaluate the quality of each classification made by the system [11]. The most important question for computing this quality rate is related to measuring the quantity of relevant information available for the classification of a given e-mail. The definition of the above mentioned quality rate could be very useful for the final user and it can be used for reducing the amount of false positive errors (legitimate messages classified as spam e-mails).

Based on our previous work, we provide a study of different successful approaches for quality estimation in spam filtering domain. In this context, we introduce a novel method for computing the reliability of a given e-mail classification that outperforms the precision reached by other well-known techniques. Our proposal is integrated in the revision stage of our previous successful SPAMHUNTING CBR system.

The rest of the paper is structured as follows: section 2 summarizes the relevant findings of previous research works on spam filtering and classification accuracy. Section 3 presents our proposal for computing the quality of the final generated e-mail classification while section 4 introduces the design and configuration of the experiments carried out. Section 5 focuses on showing the experimentation results, discussing the preliminary findings. Finally, section 6 exposes the main conclusions reached as well as the future lines of our research work.