FASiL Adaptive Email Categorization System

Yunqing Xia[1], Angelo Dalli[1], Yorick Wilks[1], Louise Guthrie[1]

[1] NLP Research Group, Department of Computer Science, University of Sheffield
Regent court, 211 Portobello Street, Sheffield, S10 4DP
{y.xia, a.dalli, y.wilks, l.guthrie}@dcs.shef.ac.uk
http://nlp.shef.ac.uk/

Abstract. This paper presents an adaptive email categorization method developed for the Active Information Management component of the EU FASiL project. The categorization strategy seeks to categorize new emails by learning user preferences, with a feature-balancing algorithm that improves the data training effectiveness and with a dynamic scheduling strategy that achieves the system adaptivity. The results of our evaluation with user-centric corpora constructed automatically from email servers are presented, with around 90% precision consistently being achieved after three months of use. Adaptivity of the system is also evaluated by studying system performance within the continuous three months.

1 Introduction

Email is one of the most ubiquitous applications used on a daily basis by millions of people world-wide. Typically, emails are stored in different folders for easy access, imposing some structure on the increasingly unmanageable amount of information received by email. Our work is focused on creating better ways of categorizing email automatically in a way that adapts to the changing needs of a user.

This work has been done as part of the EU FASiL project, which aims to construct a conversationally intelligent Virtual Personal Assistant (VPA) designed to manage the user's personal and business information through a voice-based interface accessible over mobile phones. Mobile phones have achieved high penetration rates in most major EU states and around the world, with 73% of the EU population using a mobile phone, compared to 59% in Japan and 46% in the USA and 16% average world-wide(David, 2003). As the quality of life and productivity of EU citizens is to be maintained and improved in an increasingly information dominated society, people need access to information anywhere and at any time. In order to provide this capability, two fundamental issues must be addressed: information access and information overload. When trying to access email, especially over a voice-based interface, an ideal system should categorize and present timely messages in a prioritized manner.

Empirical studies show that an active email user will create between 10 to 130 subfolders to file their emails, with a value of around 73 sub-folders being typical(Fisher and Moody, 2001). Other studies show that as number of sub-folders increases, users