

Improved Topic Identification for Similar Document Search on Mobile Devices*

Kristóf Csorba[†] and István Vajk[†]

Abstract

This paper presents a novel, two level classifier ensemble designed to support document topic identification in mobile device environments. The proposed system aims at supporting mobile device users who search for documents located in other mobile devices which have similar topic to the documents on the users own device. Conforming to the environment of mobile devices, the algorithms are designed for slower processor, smaller memory capacity and they maintain small data traffic between the devices in order to keep low the cost of communication. We propose a keyword list based topic comparison, enhanced with a two level classifier ensemble to accelerate the topic identification process. The new technique enables document topic comparison using few communication traffic and it requires few calculations.

Keywords: document classification, topic hierarchization, keyword selection

1 Introduction

Due to the rapid improvement of mobile devices (like mobile phones and PDAs) in storage capacity and processing capabilities, more and more information can be stored on them. People reading e-books on PDA are not unusual. There is an increasing interest in searching techniques designed for such ubiquitous environments, and most of the search engines are still based on keywords specified by the user. The techniques presented in this paper are designed for an automatic searching for documents which might be of the user's interest. The proposed system is analyzing the local documents and notifies the user if there is a document with similar topics available for retrieval. It is running as a background process and requires user interaction only if it finds something. Our work is part of a project aiming at supporting semantic search in mobile devices connected to a peer-to-peer network [1].

*This work has been fund of the Hungarian Academy of Sciences for control research and the Hungarian National Research Fund (grant number T68370).

[†]Budapest University of Technology and Economics, Department of Automation and Applied Informatics, Goldmann Gy. tér 3., 1111 Budapest, HUNGARY.
E-mail: {kristof,vajk}@aut.bme.hu