University of Macau

Abstract

CLUSTERING USERS FROM USER PROFILES IN THE DIGITAL LIBRARY

by Ao Ieong U

Thesis Supervisor: Prof. Gong ZhiGuo
Software Engineering

With the explosive information available over the Internet, users need to take more and more burdens in finding their interesting information. Of all the difficulties, there are two main factors that need to be investigated carefully. One of them is the great amount of information sources; information users have to deal with so many heterogeneous sources. The other is due to the high rate of information update; users should access the sources frequently in order to get the new information. One effective solution to the first problem is probably the Mediator/Wrapper architecture that can provide a unique global interface to the users. A typical solution to the second problem may be to equip the systems with automatic dissemination functions. The system will automatically send the new coming documents to the corresponding users by comparing the documents with the users' profiles. But in the Internet environment, there are too many users to each of the source systems. Therefore, it will degrade the system performances drastically if the system trys to match the new documents with each profile.

User profiles can be examined to determine the clusters of users that show similar needs of information. Based on which category an individual user falls into, the system can dynamically suggest the corresponding documents to the user in some active ways, say, by E-Mail. In order to raise the matching performance of the system, it is necessary to cluster users into different classifications. In this way, when the new document comes up it is not possible to search each user profile to ensure which document he wants. Clustering is the appropriate method to realize our target. We will compare the key words of the new document with each vector of user groups, by using the Clustering Algorithm, to find out the nearest group of users. Key words of the target user group are similar with the ones of the new document. As the result, we will send the new document information to those users by e-mail.

In our work, we will: represent user query information with keyword vectors, apply clustering techniques to classify users into different groups, match documents with cluster of users, provide new documents to users by e-mail.

Keywords: Clustering, Category, K-Means.