A Combination Approach to Web User Profiling

JIE TANG
Tsinghua University
LIMIN YAO
University of Massachusetts Amherst
DUO ZHANG
University of Illinois at Urbana-Champaign
and
JING ZHANG
Tsinghua University

In this article, we study the problem of Web user profiling, which is aimed at finding, extracting, and fusing the "semantic"-based user profile from the Web. Previously, Web user profiling was often undertaken by creating a list of keywords for the user, which is (sometimes even highly) insufficient for main applications. This article formalizes the profiling problem as several subtasks: profile extraction, profile integration, and user interest discovery. We propose a combination approach to deal with the profiling tasks. Specifically, we employ a classification model to identify relevant documents for a user from the Web and propose a Tree-Structured Conditional Random Fields (TCRF) to extract the profile information from the identified documents; we propose a unified probabilistic model to deal with the name ambiguity problem (several users with the same name) when integrating the profile information extracted from different sources; finally, we use a probabilistic topic model to model the extracted user profiles, and construct the user interest model. Experimental results on an online system show that the combination approach to different profiling tasks clearly outperforms several baseline methods. The extracted profiles have been applied to expert finding, an important application on the Web. Experiments show that

This work was done when L. Yao and D. Zhang were studying in Tsinghua University.
This work is supported by the Natural Science Foundation of China (No. 60703059, No. 60973102), Chinese National Key Foundation Research (No. 60933013), National High-Tech R&D Program (No. 2009AA01Z138), and Chinese Young Faculty Research Fund (No. 20070003093).
Authors' addresses: J. Tang (corresponding author) and J. Zhang, Room 1-308, FIT Building, Tsinghua University, Beijing, 100084, China; email: jietang@tsinghua.edu.cn, zhangjing@keg.cs.tsinghua.edu.cn; D. Zhang, 1125 Siebel Center for Computer Science, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL 61801; email: dzhang22@illinois.edu; L. Yao, Department of Computer Science, University of Massachusetts Amherst, 140 Governors Drive, Amherst, MA 01003; email: lmyao@cs.umass.edu.