Position and Relevant Work for PIM

Yi Zhang
School of Engineering
University of California Santa Cruz
yiz@soe.ucsc.edu

Biography

Yi Zhang is an Assistant Professor at Baskin School of Engineering, University of California Santa Cruz. Her research is related to information retrieval, statistical machine learning, and natural language processing. She has collaborated with startups, large corporations and government agencies on information retrieval research. Dr. Zhang received her Ph.D. and M.S. from Carnegie Mellon University and B.S. from Tsinghua University. She is a committee member of SIGIR and has received the Best Paper Award in ACM SIGIR 2002.

Position and Relevant Work to Personal Information Management

Finding information is the most important problem now and large scale internet search tools, such as Google and Yahoo, have gain wide popularity among normal user. The huge capacity of digital personal storage, the explosion of digital information accessible locally or over the internet, and the ever increasing usage of smart personal devices makes it an important current problem as well as future problems to locate the information easily across the user's various local and global information resources such as the internet, intranet, local PC, laptop, pocket PC, smart phone smart key, smart card, smart wallet.

One major research goal of Dr. Zhang is advancing the fundamental theory, investigating long term and short term practical techniques, and developed efficient tools to help each individual user to locate information distributed across heterogeneous and evolving information resources, including the user's own personal stores as well as unknown information resources on the web. To solve this problem in a principled way, we model the user's information seeking process based on Bayesian decision theory. Under this framework, a personalized information management system adaptively learns its belief about the user (user model) in contexts of uncertainty, and the user model is updated in the light of changing evidence. Expected utility maximization is the basis for the system to make rational decisions on how it should interact with the user in order to avoid certain kinds of undesirable results The utility of a piece of information for a user depends on many factors. We have explicitly modeled the user criteria as a multivariate utility function defined over sub utilities, (such as relevancy, authority and novelty of the information) and demonstrated how to automatically learn user specific utility functions from explicit user feedback.

Recently, we are working on solving a practical concern that any user new to a personalized system must endure poor initial performance because there is insufficient information about that user. To help the system better serve a new user, we use explicit and implicit feedback from the user to build user specific models, and we use Bayesian hierarchical methods to utilize information about existing users to learn the model for a new user. We analyze the adaptive performance of the models using two data sets gathered from user studies where users' interaction with a document, or implicit user feedback, was recorded along with explicit user feedback. We also evaluated the benefits of utilizing implicit feedback in user profiling and find mixed results on different data sets.

Dr. Zhang has much prior work in a related area called adaptive information filtering. Her Ph.D. dissertation was focusing on developing a personal information filtering system monitors an incoming document stream to find the documents that match information needs specified by user profiles. One major challenge in personal adaptive filtering is to develop a system to learn user profiles efficiently and effectively from very limited user supervision.

In order to overcome this challenge, the system needs to do the following: use a robust learning algorithm that can work reasonably well when the amount of training data is small and be more effective with more training data; explore what a user likes while satisfying the user's immediate information need and tradeoff exploration and exploitation; consider many aspects of a document besides relevance, such as novelty, readability and authority; use multiple forms of evidence, such as user context and implicit feedback from the user, while interacting with a user; and handle various scenarios, such as missing data, in an operational environment robustly. We used the Bayesian graphical modeling approach as a unified framework for filtering. We customized the framework to the filtering domain and develop a set of solutions that enable us to build a filtering system with the desired characteristics in a principled way. We evaluated and justified these solutions on a large and diverse set of standard and new adaptive filtering test collections. Firstly, we developed a novel technique to incorporate an IR expert's heuristic algorithm as a Bayesian prior into a machine learning classifier to improve the robustness of a filtering system [1]. Secondly, we derived a novel model quality measure based on the uncertainty of model parameters to trade off exploration and exploitation and do active learning [3]. Thirdly, we carried out a user study with a real web-based personal news filtering system and more than 20 users. With the data collected in the user study, we explored how to use existing graphical modeling algorithms to learn the causal relationships between multiple forms of evidence (including implicit and explicit feedback from the user) and improve the filtering system's performance using this evidence[1]. We have built a longer-term learning environment for the study and demonstrated that we can collect a significant amount of data about a user's interests. Second, we have built user independent and user specific models from the data. We have modeled the information need of a user as a utility function based on a set of broader, realistic and more distinct criteria, such as topical relevance, novelty, readability and authority. There is much prior work talking about modeling users and going beyond topical relevance. However, the prior works usually come down to a weighted bag of words. Explicitly modeling different criteria and learning the importance of each criterion using probabilistic inference algorithm from the data, the filtering system goes beyond the bag of words approach and combines multiple forms of evidence. This broader, sophisticated, realistic, and data-driven user modeling approach may lead to a system that serves the user better.

Because of the possibility of accumulating a large amount of user information over a long period of time, the PIM environment offers many opportunities for research. The interesting and challenging research topics are: how to let systems learn sophisticated user models to better satisfy the user's information needs over time; how to develop a robust learning algorithm that works reasonably well when the amount of training data is small and become more effective with more training data; how to trade off immediate gain vs. long-term gain; how to help the user explore different information sources; how to use multiple forms of noisy evidence, such as user context and implicit feedback from the user, while interacting with a user; and how to handle various problems like missing data in an operational environment robustly. Our prior research in personal adaptive filtering may help to solve some of these problems. We are also interested in new problems arising from the PIM environment, such as how to integrate heterogeneous personal information sources/devices seamlessly, how to model the context of the user, how to model the change of the user as well and the change of the information resources, and how to set up experimental environment for research in PIM. I would be very glad to exchange ideas with researchers attending the PIM workshop.

[1] Y. Zhang, J. Callan, Combine Multiple Forms of Evidence while Filtering, In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing , Vancouver, Canada, 2005 (HLT 2005)

[2] Y. Zhang Using Bayesian Priors to Combine Classifiers for Adaptive Filtering In Proceedings of the 27st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield , United Kingdom, 2004 (SIGIR 2004)

[3] Y. Zhang, W. Xu, J. Callan "Exploration and Exploitation in Adaptive Filtering Based on Bayesian Active Learning International Conference on Machine Learning (ICML 2003)