Web Personalization and Cohort Information Services for Natural Resource Managers

Crystal E. Redman
Colorado State University
Computer Science Department
Fort Collins, Colorado 80523
http://www.cs.colostate.edu/∼ redman/

Introduction

Bing and *Google* are finely tuned to quickly serve the frequent and popular information needs of the masses. Topic specificity, customizability, and automatically pursuing the long term unique information needs of individual users are not among the strengths of current main stream search engines (Jansen, Spink, and Saracevic 2000) (Teevan, Dumais, and Horvitz 2005). This gap has inspired web personalization and collaborative information seeking tools such as *Google Alerts* and has encouraged topic-specific blogs and podcasts.

Web personalization tools lie on a spectrum between individualized and fully coordinated searches, between short and long term information interests and between public and private information needs. Much research attention has recently focused on web personalization (Castellano and Torsello 2009) (Smyth et al. 2009) (Chu and Park 2009) (Memari, Amer, and Gmez 2010) (Stamou and Ntoulas 2009) (Lacomme, Demazeau, and Camps 2010) (Amin and Nayak 2010).

We aim to address yet unanswered questions in web personalization by providing an integrated view of documents found through different tools, considering user confidentiality, emphasizing the benefits of client-side web personalization, and highlighting the power of social webpage recommendation when traditional information retrieval and collaborative filtering methods are used in conjunction. These principles define a framework which leverages groupings of users with overlapping interests called user cohorts.

A Prototype Web Personalization Tool for Natural Resource Managers

Matilda [1], a prototype system we are building for the United States Geological Survey (USGS), helps ecologists, natural resource managers and wildlife scientists collect and share documents and data pertaining to climate change. Matilda users are not climatologists, but they need to make long term resource management decisions while accounting for the impact of climate change in their areas. The cohort has similar, but not identical information needs as they individually focus on different species, habitats, geographic areas, time scales and resources. Their information needs are long term and highly dynamic - nearly everything about this topic is in flux. For these users, information search can be made more effective with knowledge about the field and about the types of documents being retrieved. Because the resource management decisions require judgment about the materials collected, the users require confidentiality and must trust the sources.

Matilda is designed to 1) tailor information collection for a particular group of users who share overlapping interest in climate change impact and 2) support their sharing and organization of the information that is found. Matilda suggests pages, the users give feedback on whether or not the pages are relevant and Matilda maintains a repository of relevant documents and refines its model of users' interests to fuel the search for new content on the subject. Thus, Matilda mediates between the user and a search engine and functions as a tailored Web crawler. Matilda also offers users the ability to find and leverage information gathered by other users. Users can elect to receive colleague recommendations from Matilda to form relationships based on similarity of interests. Also, users can receive topic recommendations based on the interests of their colleagues.

Currently, we are focused on collecting and implementing improved document feature extraction by incorporating automatic topic discovery into Matilda. We are leveraging data we have already collected from alpha testers and from crawling from *Wikipedia* and *Bing* using terms already deemed relevant from web pages collected in alpha testing. Automatic query reformulation for topics and for users is also among our current priorities. This allows us to continually seed web crawls to provide users with fresh information relevant to the topics they are following.

Research Trajectory

Data to Inform Search

Knowledge gained from a variety of sources is the key to tailoring search to a specific topic. Web documents, topics, relevance judgments, and similarity among users are all important sources to inform search.

Documents from the web collected using a web crawler

[1] http://beethoven.cs.colostate.edu/Matilda/

and by allowing users to contribute pointers directly are maintained in a repository. Using these web documents, we calculate features such as a variant of TFIDF, type of document, and important phrases. In the future, we will incorporate features such as prose sophistication, readability, associations with reputable authors, and citation trails. Though we currently focus just on HTML documents, but intend to include other types of documents like *pdf* files to provide users with documents from peer reviewed sources.

Topics are an important way to organize web personalization tools. In Matilda, users initially describe the topic and provide relevance feedback for documents suggested for the topic. Currently, any user who indicates an interest in the topic will be receive recommendations for all pages other users following that topic have recommended highly. In the future, machine learning algorithms will be used to analyze the feedback to create user specific and cohort specific topics. A current open question is whether initializing with expert domain knowledge and automatically updating the subtopics of a specific domain like *climate change* will improve topic specific search. Another question is when and how topics should be coalesced or separated.

Relevance judgments can be used to refine document relevance. We allow the user to determine whether a suggested document is relevant to one or more topics. To leverage a variety of collaborative filtering algorithms, we will use a five valued relevance feedback mechanism. We believe that combining collaborative filtering and information retrieval algorithms will strengthen the recommendations the user receives especially for topics that are not widely followed by the typical user.

Cohort membership, whether intentional or inferred, is used to provide personalized weightings for recommendations. Significant overlap in the topics a group of users are following can define a collaborative topic model. In the future, Matilda will consider together the topics a user is following, colleague relationships, and user trust in making recommendations. For example, does a specific user consistently contribute recommendations to other members of the cohort which are well received? Privacy will also becomes an important factor as cohort membership functionality is embellished.

Automatic Topic Discovery and Organization

From alpha testing we have completed with a cadre of natural resource manager users, we have collected many webpages the users have judged relevant. These pages along with well known government web sites, such as the USGS web site, can be crawled for terms that are related to *climate change*. These terms can then be used to seed crawls from Wikipedia, Bing, or Google searches. Using the results from these crawls, we can get a better idea of the sub-topics that define *climate change* and create a better, somewhat automatic listing of topics to present to users.

Clustering Algorithms for Profiling

Unsupervised and semi-supervised clustering algorithms are useful in personalized and cohort document search because user feedback can be introduced into the clustering process and metrics can be adapted given the small subset of data that is actually labeled with topic specificity or relevance feedback. We will be further investigating clustering algorithms to eliminate terms which do not contribute to the model, allow for missing relevance feedback, determine the topic subsets, and preprocess document sets and relevance feedback data using dimension reduction techniques.

Acknowledgement

This research is supported by a grant from the USGS to Colorado State University.

References

Amin, M., and Nayak, R. 2010. Theoretical model of user acceptance: In the view of measuring success in web personalization. In Forbrig, P.; Patern, F.; and Mark Pejtersen, A., eds., *Human-Computer Interaction*, volume 332 of *IFIP Advances in Information and Communication Technology*. Springer Boston. 255–264.

Castellano, G., and Torsello, M. 2009. How to derive fuzzy user categories for web personalization. In Castellano, G.; Jain, L.; and Fanelli, A., eds., *Web Personalization in Intelligent Environments*, volume 229 of *Studies in Computational Intelligence*. Springer Berlin / Heidelberg. 65–79.

Chu, W., and Park, S.-T. 2009. Personalized recommendation on dynamic content using predictive bilinear models. In *Proceedings of the 18th international conference on World wide web*, WWW '09, 691–700. New York, NY, USA: ACM.

Jansen, B. J.; Spink, A.; and Saracevic, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.* 36:207–227.

Lacomme, L.; Demazeau, Y.; and Camps, V. 2010. How to integrate personalization and trust in an agent network. In Filipe, J.; Fred, A.; and Sharp, B., eds., *Agents and Artificial Intelligence*, volume 67 of *Communications in Computer and Information Science*. Springer Berlin Heidelberg. 247–259.

Memari, A.; Amer, M.; and Gmez, J. M. 2010. A beehive-like multi-agent solution to enhance findability of semantic web services and facilitate personalization within a p2p network. In Davcev, D., and Gmez, J. M., eds., *ICT Innovations 2009*. Springer Berlin Heidelberg. 227–236.

Smyth, B.; Briggs, P.; Coyle, M.; and OMahony, M. 2009. Google shared. a case-study in social search. In Houben, G.-J.; McCalla, G.; Pianesi, F.; and Zancanaro, M., eds., *User Modeling, Adaptation, and Personalization*, volume 5535 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 283–294.

Stamou, S., and Ntoulas, A. 2009. Search personalization through query and page topical analysis. *User Modeling and User-Adapted Interaction* 19:5–33.

Teevan, J.; Dumais, S. T.; and Horvitz, E. 2005. Beyond the commons: Investigating the value of personalizing web search. In *In Proceedings of the Workshop on New Technologies for Personalized Information Access*, 84–92.