Web content personalization: a state-of-the-art review

Alex Gain, Cody Licorish and K. Brent Venable

Department of Computer Science, Tulane University, USA
Email: again1@jhmi.edu, clicoris,kvenabl}@tulane.edu

The Consortium for Resilient Gulf Communities (CRGC) [1], funded by the Gulf of Mexico Research Initiative (GoMRI)[2], has been established with the purpose of assessing and addressing the public health, social and economic impacts of the 2010 Deepwater Horizon oil spill in the Gulf of Mexico. The consortium focuses on understanding and promoting communities' resilience to adverse future events. A high priority for the consortium is to develop an effective strategy to match potential users with relevant findings. In this context, the CRGC risk-communication team is pursuing the design of a website capable of presenting content in a personalized way to users of different types. Indeed, the consortium's findings and products will be of interest to different kinds of users, such as, for example, political decision makers or members of prominent gulf industries like fisheries, tourism and ship building. Content personalization aims at getting the right information to the right people and involves identifying items of significant interest to a given user among the products of the consortium and highlighting them when he or she accesses the website.

This review outlines possible approaches to content personalization, highlighting the strengths and weaknesses of each to show how the consortium decided upon an approach for their website. Our review is summarized below and follows, top-down, the topics as depicted in Figure 1.

Web personalization. Artificial Intelligence plays a fundamental role in improving web functionality, in particular in the context of web personalization [26]. Web personalization is becoming a necessity rather than a luxury in several areas including e-business and customer relationship management. Our work is aimed at demonstrating that similar approaches can improve the effectiveness of information not for lucrative purposes but for building community resilience.

Web personalization has three elements: (1) personalized web searches [37, 14], (2) personalized advertising delivery, and (2) personalized content [26]. Most of the literature on *personalized web searches* is in some way connected to the so-called Semantic Web [32], where semantically-enhanced information is added to Web documents through ontologies, that is structures for knowledge organization capable of storing information about the knowledge itself (meta data). This allows to manage information automatically. An ontology is usually comprised of four components: classes, representing fundamental concepts (e.g. "city" and "state"), instances, representing the ground level of concepts (e.g. "Austin" and "Texas"), relations defined over classes (e.g. "is the capital of" defined between class "city" and class "state" and satisfied by instances "Austin" and "Texas"), and axioms expressing constraints (e.g. "a state can only have one capital"). Artificial intelligence plays a key role by providing techniques which automatically extract ontologies and methods that populate them [26, 12]. Among the

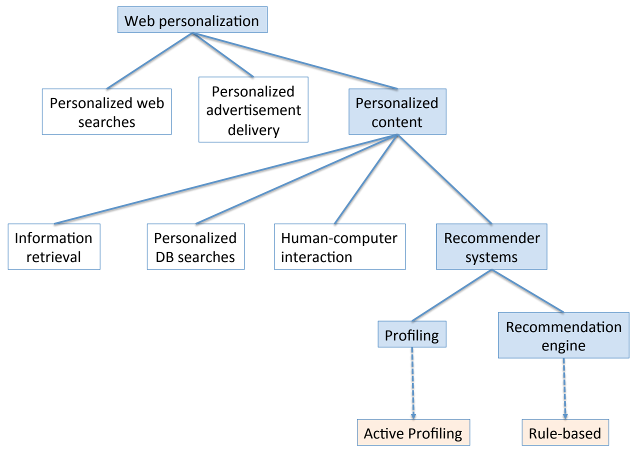

Fig. 1. Organization of topics on web personalization. The boxes highlighted in light blue are the most relevant to the website we are developing. The boxes highlighted in light orange correspond to the design options being pursued.

different kind of ontologies, personalized ontologies [27] are a formal conceptualization of user profiles encapsulating the user's personal model of their information needs. Ontologies and meta data are a necessity for personalized web searches on the World Wide Web scale. However, their sophistication in terms of modeling structures and the computational overhead of the associated algorithmic machinery makes them unsuitable for smaller scale settings such as ours.

Personalized advertising delivery. The second element of web personalization, personalized advertising delivery, is the focus of web or digital marketing [8, 3] techniques and behavioral targeting [7]. The aim of these techniques is to provide the user with the most preferred item out of a larger set but they are unsuitable for our consortium's website. In the marketing domain the driving force is the revenue of the advertiser, which is not relevant to our consortium. In behavioral targeting, the user profile is built almost exclusively by observing his web browsing behavior. Our goal is to present information on resilience in the Gulf coast region, and it seems reasonable to assume that such type of behavioral data is very limited and difficult to obtain. Importantly, while the vast majority of practices in this context advocate for deep personalization mostly based on behavioral data tracked through different kinds of technologies, in [11], the author argues in favor of a "leaner" form of personalization. For instance, user profiles based only on explicit information, as the one collected through registration forms or online questionnaires, combined with manually defined business rules, driving recommendations towards the business needs, can be sufficiently effective in many small to medium size contexts. The leaner approach to personalization is relevant to the consortium because the specificity of the topic and the size of target audience makes behavioral data impractical to obtain.

Personalized content and recommender systems. We now turn our attention to the third and, in our opinion, most relevant element of web personalization for our purposes, that is, personalized content. Content personalization is used in information retrieval systems [23], in database systems [17], in human-computer interaction systems [31] and, most prominently, in so-called recommender systems [28, 13, 36].

To directly address the need to design a consortium website that is relevant specifically to the user [15, 33], we focus on *recommender systems* [28, 13, 36]. Personalization technologies modify the pages that are viewed by users in order to emphasize content which is judged as more relevant to users' interests. The literature agrees on the abstract architecture of web and website content personalization [28, 13, 36]. For example, in [33], the authors outline a very general architecture for web personalization which we reproduce graphically in Figure 2 for clarity. The inputs to the system are the Web logs (describing how pages are accessed by users), the Web site's content, the Web site's structure and the user profiles. Information coming from all of these components is processed by a Web analysis and pattern discovery module which, then, feeds its output to a recommendation model that interacts directly with the user.

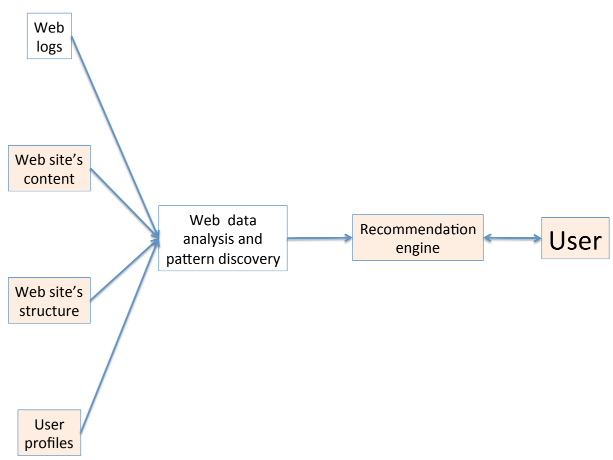

Fig. 2. Web personalization architecture as described in [33]. We highlight in light orange components which are present in our system.

This architecture is applicable in our case with the exception of Web logs and a one-way interaction between the recommendation model and the user. A very similar architecture is described in [19], with the exception that user profiles are the output of an additional component called the profile learner, which takes as input both the feedback from the user as well as the performance of the current user profile representation on some examples.

Building users profiles. As mentioned above, user profiles are an essential and crucial component of any website content personalization architecture. For this reason, we briefly summarize the main state-of-the-art methods for acquiring the information on the user and organizing it into structures called profiles. To personalize a web site [9, 15], information about the users must be gathered and stored. To accomplish this, many web sites create individual visitor or group profiles. In general, the user profiling process consists of two main phases: first, raw information is gathered about the user and then a profile construction phase follows. The next step is to use the profile for recommendation purposes. The most common profiling techniques used are:

- active profiling: the user volunteers information on his interests;
- collaborative filtering [30, 34, 24]: some initial information is obtained (e.g. via active profiling) on some users. Then, users are mapped into types corresponding to "like-minded" users;
- behavior based profiling: as defined above in the context of behavioral targeting, users are observed in terms of their browsing activities and patterns;
- short term interests monitoring: a leaner version of behavior-based profiling, mostly based on analyzing the words a user types into search engines.

In [33] these profiling approaches are distinguished into: *interviewing*, that is, relying on manually acquired data (e.g. active profiling); *semi-interviewing*, where acquired information is on categories and groups of items rather than on single items; and *non-interviewing*, data on users in acquired by observation and without the user's direct involvement or knowledge (e.g. collaborative filtering and short-term profiling).

In our case, the goal is to distinguish users belonging to different categories in terms of their interests on the impact of the oil spill, community resilience and community action planning. Given the level of detail of these topics, tracking data from which to passively extract users' interests is unavailable and impractical to obtain. Thus, the technique which appears to be more suited for us is *active profiling*, as other consortium members will be carrying out in-depth interviews with community members which will inform, together with the information acquired at registration time, the user profiles. In active profiling, users are asked to complete online registration forms that request basic personal information and details about special interests. This approach is limited when users provide incorrect information or refuse to provide any information [9, 15]. Nonetheless, they are at the base of popular Web portal personalization tools such as MyYahoo! (my.yahoo.com). Categories of features which have been identified as useful in the context of active profiling are:

1. Geographic
2. Cultural and ethnic
3. Economic conditions and income
4. Level of decision making and title
5. Size of company
6. Age

7. Values, attitudes and beliefs
8. Knowledge and awareness
9. Lifestyle
10. Buying patterns
11. Media used

We predict only a subset of these categories will be relevant given the content and target of our website. Another issue user profiling must deal with is user identification. Common methods of identification include software agents, logins, enhanced proxy servers, cookies, and session ids [9]. Given that in our case we will not track the user, logins seem to be the optimal choice.

The second phase of user profiling can be broken down into choosing a user profile representation and then populating its instances. The most widely used profile representations are weighted keywords, semantic networks, weighted concepts and association rules [9]. Sets of keywords can be generated both by active or passive profiling data extraction techniques. Weights, that is numbers reflecting the user's interest in a topic can be added to enhance the model. Weighted keywords were one of the first approaches to be investigated [22] and have been used in several systems ranging from personalized online newspapers [16] to browsing assistants [18]. Semantic network profiles go one ste further and represent users with a set of weighetd keywords structured in a newtork[10, 20]. More precisely, each user is associated with a network where nodes represent concepts, modeled as weighted keywords, and links between nodes represent associations between concepts. In this approach, both concepts and links between concepts are used to describe the user and his preferences. Concept-based profiles are similar to semantic networks where nodes are now populated by abstract topics of interest to the user rather than keywords or sets of keywords. The concepts can be organized hierarchically [4], and in a static [35] or dynamic [6] fashion. All approaches except the weighted keywords approach, rely heavily on machine learning techniques for profile construction from data [9] and are thus outside of our scope. We adopt a from of weighted keywords and rely on manual construction and refinement of user type profiles.

Matching content to profiles. The other fundamental component of a recommender system is the recommender engine, which matches the users' profiles to items to be recommended. Algorithmically, recommender paradigms can be classified into three main categories: rule- based systems, content-based filtering systems and collaborative filtering systems [21]. Content-based filtering algorithms base their recommendations on what the user has liked in the past [25, 19], while collaborative filtering algorithms recommend items chosen by "like-minded" users [30, 34, 24]. Since we do not have information on what users have liked in the past, neither of the latter methods are applicable. A more traditional, rule-based approach system [21] better fits our needs. Such systems rely on manually or automatically generated decision rules which match users to recommended content. Such a paradigm, allows system designers to specify rules based on the user profiles. The rules affect the content which is presented to users when their profile satisfies the rule's conditions. Note that a key role is played by the knowledge engineering used to define rules in accordance to the specific characteristics of the domain. This is particularly appealing to us, since such type of knowledge will be gathered through in-depth interviews with community members. Among the possibilities for defining such rules we focus on Brafman's approach of relational rules for control [5]. This approach is well suited to our case due to the presence of rule-based preference specification language which extends several AI-based preference definition models and, thus, allows the use of state of the art and finely tuned preference reasoning engines [29].

Conclusions. Summarizing, we have presented a survey of the state-of-the-art in web content personalization, in light of the design of a website aimed at providing content recommendation to users where the content is mostly of scientific nature and with a narrow scope. Much of the literature relies on data obtained by tracking online behavior of users. The unavailability of this type of data in our setting makes these approaches impractical from our point of view and limits our options to active profiling and rule-based recommendation. On the other hand, the size of the problem we tackle in terms of number of different types of users and number of available items is limited. Furthermore, we can leverage an unusually large amount of data obtained via in depth surveys and interviews carried out by other teams of the consortium. We foresee this will significantly mitigate the well-known drawbacks of the approaches which we will pursue.

1 Acknowledgements

The authors would like to thank Melissa Finucane for her insightful and detailed comments.

This research was made possible by a grant from The Gulf of Mexico Research Initiative. Data are publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org.

References

1. Consortium for Resilient Gulf communities. http://www.rand.org/gulf-states/resilient-communities.html.
2. Gulf of Mexico Research Initiative Consortia. http://gulfresearchinitiative.org/consortia/.
3. Charlesworth Alan. *Digital marketing: a practical approach*. Routledge, 2014.
4. Eric Bloedorn and Inderjeet Mani. Using NLP for machine learning of user profiles. *Intell. Data Anal.*, 2(1-4):3–18, 1998.
5. Ronen I. Brafman. Relational preference rules for control. *Artif. Intell.*, 175(7-8):1180–1193, 2011.
6. Chien Chin Chen, Meng Chang Chen, and Yeali S. Sun. PVA: A self-adaptive personal view agent. *J. Intell. Inf. Syst.*, 18(2-3):173–194, 2002.
7. Ye Chen, Dmitry Pavlov, and John F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 209–218, New York, NY, USA, 2009. ACM.
8. Chaffey Dave and Ellis-Chadwick Fiona. *Digital marketing: strategy, implementation and practice*. Pearson, 2012.

9. Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. User profiles for personalized information access. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, pages 54–89. Springer-Verlag, 2007.
10. Gianluigi Gentili, Alessandro Micarelli, and Filippo Sciarrone. Infoweb: An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17(8-9):715–744, 2003.
11. Jon Gibs. Lean personalization. www.hegeinc.com/ideas/report/lean-personalization, 2014.
12. Jim Hendler and Tim Berners-Lee. From the semantic web to social machines: A research challenge for AI on the world wide web. *Artif. Intell.*, 174(2):156–161, 2010.
13. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.
14. Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003*, pages 271–279. ACM, 2003.
15. R. Wilson K. Garvie Brown and M. Shamos. Web content personalization overview. Technical report, 2009.
16. Tomonari Kamba, Hidekazu Sakagami, and Yoshiyuki Koseki. ANATAGONOMY: a personalized newspaper on the world wide web. *Int. J. Hum.-Comput. Stud.*, 46(6):789–803, 1997.
17. Georgia Koutrika, Evaggelia Pitoura, and Kostas Stefanidis. Preference-based query personalization. In *Advanced Query Processing, Volume 1: Issues and Trends*, volume 36 of *Intelligent Systems Reference Library*, pages 57–81. Springer, 2013.
18. Henry Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, 2 Volumes*, pages 924–929. Morgan Kaufmann, 1995.
19. Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, pages 73–105. Springer, 2011.
20. Alessandro Micarelli and Filippo Sciarrone. Anatomy and empirical evaluation of an adaptive web-based information filtering system. *User Model. User-Adapt. Interact.*, 14(2-3):159–200, 2004.
21. Bamshad Mobasher. Data mining for web personalization. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, pages 90–135. Springer-Verlag, 2007.
22. Alexandros Moukas and Pattie Maes. Amalthaea: An evolving multi-agent information filtering and discovery system for the WWW. *Autonomous Agents and Multi-Agent Systems*, 1(1):59–88, 1998.
23. Phivos Mylonas, David Vallet, Pablo Castells, Miriam Fernández, and Yannis S. Avrithis. Personalized information retrieval based on context and ontological knowledge. *Knowledge Eng. Review*, 23(1):73–100, 2008.
24. Xia Ning and George Karypis. Recent advances in recommender systems and future directions. In *Pattern Recognition and Machine Intelligence - 6th International Conference, PReMI 2015, Warsaw, Poland, June 30 - July 3, 2015, Proceedings*, volume 9124 of *Lecture Notes in Computer Science*, pages 3–9. Springer, 2015.
25. Michael J. Pazzani and Daniel Billsus. Content-based recommendation systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, pages 325–341. Springer-Verlag, 2007.
26. N. Ichalkaranje R. Nayak and Lakhmi C. Jain. *Evolution of the Web in Artificial Intelligence*. Springer, 2008.
27. B. Ravinshankar and D. Dipa. Web personalization using ontology: A survey. *IOSR Journal of computer Engineering IOSRJCE*, 1(3):37–45, 2012.

28. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.
29. Francesca Rossi, Kristen Brent Venable, and Toby Walsh. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.
30. J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, pages 291–324. Springer-Verlag, 2007.
31. Andrew Sears and Julie A. Jacko. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications (Human Factors and Ergonomics Series)*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2007.
32. Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.
33. S. Sridevi and R. Umarani. Web personalization approaches. *International Journal of Advanced Research in Computer and Communication Engineering*, 2:680–684, 2013.
34. Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.
35. Joana Trajkova and Susan Gauch. Improving ontology-based user profiles. In *In proceedings of the Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2004, 7th International Conference*, pages 380–390. CID, 2004.
36. Gulden Uchyigit and Matthew Y. Ma, editors. *Personalization Techniques and Recommender Systems*, volume 70 of *Series in Machine Perception and Artificial Intelligence*. WorldScientific, 2008.
37. David Vallet and Pablo Castells. On diversifying and personalizing web search. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, pages 1157–1158. ACM, 2011.