

Frances McConihe

Information Retrieval, Professor Benoit

Final Paper

December 1, 2005

Harnessing the Utility of Information Production by Individuals on the Web

An approach to personalization with an eye on privacy

Introduction

“Web personalization is the process of customizing the content and structure of a Web site to the specific and individual needs of each user, without requiring them to ask for it explicitly” (Eirinaki and Vazirgiannis, 2003, p. 21). Many approaches to personalizing web content have been suggested, but even as the internet continues to expand exponentially, ways of finding or discovering information that is relevant or of interest to a particular individual remains elusive. The key lies in taking advantage of the information created by each individual user.

Users are generating massive amounts of information every time they access the internet. Some of this information is collected by site servers—page-views, query terms, click-streams, to name just a few forms—but this information is pooled with all other users of the site and provides only general insights. Other information—purchase history, item interest, browsing patterns—is collected, primarily on e-commerce sites, connected with a single user, and associated with other identifying information, such as demographic makeup and financial details (raising privacy concerns). Some of this information is used to personalize the site, usually through recommender systems. But this personalization is limited to that particular site.

Users themselves have little, if any, access to the information collected about them, and gain even less utility from it to help them navigate through the increasingly large

amount of information available on the internet. Their personally derived information is scattered across domains and locked behind proprietary e-commerce walls. And the preferences that they have delineated through bookmarking, saving, printing, etc., are not currently taken into consideration at all.

The following paper proposes a system to gather all information generated by the user, regardless of domain, and place it back in the hands of that user, allowing him to reclaim control and regain the utility of his information. The information previously pooled and generalized on servers becomes a valuable resource in creating a context for systems to personalize and improve recommendations and retrieval. In addition, commercial sites are able to gain a more accurate portrayal of each user, increasing their ability to target each user more directly with products and information he may be interested in. But importantly, this browsing and searching information is uncoupled from the user's identifying information, and the user controls the collection of information and access to his history.

In this paper, I will begin with a review of the literature and history of personalization methods and systems developed over the past decade. I do not intend to be exhaustive, as several broad reviews of the literature already exist (Burke, 2002; Perugini & Gonçalves, 2002; Pierrakos, Paliouras, & Papatheodorou, Spyropoulos, 2003; Eirinaki & Vazirgiannis, 2003). The evolution of these systems will be reviewed and various elements will be examined. Issues and concerns relating to personalization will be discussed. Finally, a description of a simple user-empowering approach to web personalization, developed by the author, will be presented.

Literature Review

To date, there are three models for adapting information on the web for an individual user: recommendation, social network exploitation, and personalization (Perugini & Gonçalves, 2002). The first actively seeks out and presents information that may be of

interest to the user, gathering information deemed relevant based on user-generated information. The second analyzes social structures inherent in web sites and web communities, clustering users together to allow them to find information via their nearest neighbor. The last is similar to the first, but instead of gathering information of possible interest, it instead leverages the user-generated data to filter out information; it remains in the background, shaping the user's interactions without needing explicit user input. This paper will focus on this last model of adapting information to the user, though the distinctions between the three models are not always clear and have not been firmly established as individual entities in the literature, which is itself dispersed across domains from information retrieval to user modeling, cognitive psychology to marketing research.

Personalization emerged from information filtering, in which the user would explicitly delineate a set of rules for incoming information to meet before it would be presented to that user (email push systems are one example of such a system). It has in fact been suggested that information filtering and information retrieval are "two sides of the same coin" (Belkin & Croft, 1992) and even that all retrieval systems are recommender systems (Furner, 2002). The early versions of what we would today call personalized or recommender systems, as defined in the introduction, began with content-based filtering, a mutation of information filtering, where elements of items of previously rated by the user were examined and then used to recommend similar items. But content-based systems have several drawbacks, the greatest of which is the need for the user to provide or generate a large number of ratings to get useful results.

Shortly thereafter, collaborative systems appeared. Instead of looking at the content of the items, collaborative systems created clusters of users with similar taste and then recommended unseen items rated highly by nearby neighbors. (Goldberg, Nichols, & Oki, 1992; Terveen, Hill, & Amento, 1997) But collaborative systems have problems as well; they suffer from the "cold start" problem, where a new item cannot be recommended before being rated by another user. Also, if there are too few users with similar tastes, there are no overlapping items to be recommended.

The more interesting systems began to appear in the late 1990s. Some combined the two approaches in an attempt to neutralize the weaknesses of the collaborative and content-based systems on their own. (Balabanovic, Shoham, & Yun, 1997) Some brought other information into play, including bookmarks (Rucker & Polanco, 1997), human created knowledge bases (Burke, 1999), link analysis (Kautz, Selman, & Shah, 1997), and others (Burke, 2002).

Much of the research showed that users were unwilling to put much effort into providing the system with ratings or other information on which to base its recommendations, spurring the desire for methods of collecting user data implicitly. So another approach soon emerged from the marketing and e-commerce fields: web mining. The detail of information collected on each user provided by web server logs enabled recommender systems to move closer to personalized systems, setting the stage for a true tailoring of content for each user. What people were accessing, how they were getting there, and what search terms they used to find information were all captured in the web server logs. Buchner & Mulvenna (1999) suggested a shift from the traditional work-intensive marketing knowledge, generated by marketing experts, toward the copious, and significantly cheaper, web data provided by server logs to tailor information to each user. By 2000, frameworks for using the server logs were developed (Mobasher, Cooley, & Srivastava, 2000; Mobasher, Dai, Luo, Sun, & Zhu, 2000) and several systems were using the data to personalize each user's web experience. (Wasfi, 1999; Srivastana, Cooley, Deshpande, & Tan, 2000)

Villa & Chalmers (2001) argued that user behavior captured in server logs is a sort of "action language" linking explicit user behavior (mouse clicks, printing, bookmarking) with implicit motives (interest, relevancy); an essential connection in making sense and use of the web mining data. Ruthven, Lalmas, & van Rijsbergen (2003) correlated user search behavior and interaction with search results to user relevance judgments, suggesting that information extracted from user behavior can aid query reformulation and improve

subsequent results. Claypool, Le, Wased, & Brown (2001) and Kelly & Teevan (2003) correlated implicit feedback, (e.g. reading time, printing, etc.), with user interest.

Until recently, the user's information collected by various site servers—browsing, navigation, query, and other data—and over sessions remained within that a single site or session, becoming operational to the user only within that limited framework. But increasingly, client-side systems have been proposed. Clifford Lynch (2001) suggested a system that resides on the client's computer and passes along the user's history and preferences to the site being visited, enabling the content to be personalized. This approach has three advantages: the user gains greater control over their information, information can be carried across sessions, and each site gains previously isolated information about the user's interests generated on other sites. Subsequently, such systems have been developed. (Cassel & Wolz, 2001; González, Lòpez, & de la Rosa, 2005) Speretta & Gauch (2004) achieve a similar result by "wrapping" a search site with a user's query and search result interaction history, though this approach does not address the privacy and cross-domain issues that client-side systems solve. The data that they collect, the user's query history and interaction with the results, is quite useful in calculating user-specific interests and document relevancy.

The next logical step after moving the user's information to their local machine is to make that information portable. Chan (2005) proposes a mobile smart card for storing cookie information, allowing the user to maintain session independent information regardless of the machine he is using. Unfortunately, this is a step backward in privacy and security: such a card would contain a vast amount of sensitive personal information and make the information as portable to the user as to someone with malicious intent.

Pitkow, Schutze, & Cass (2002) showed significant success with *Outride*, a personalization system which resides close to the user, allowing cross-domain and cross-session user-specific relevance. *Outride* augments the user's search queries using contextual information from the recent sites or history and filtering results through the user's model, formed with information from submitted bookmarks and the user's own Open Directory

Project site weighting. Similarly, Teevan, Dumais, & Horvitz (2005) used user query and browsing history, as well as local machine information (documents and email analysis) to provide relevance feedback and aid in search result ranking.

But as recommender systems became more common, especially on e-commerce sites, issues of trust and privacy, as well as user perception of effectiveness of the systems, arose. (Swearingen & Sinha, 2001) Users were often concerned that information gathered about them, especially with data that is collected implicitly and hidden from the user, will be used for purposes that they did not intend or will portray them inaccurately. (Wærn, 2004) Cranor (2003) discusses methods to alleviate privacy concerns; she suggests storing the user's data on the client-side, giving the user control over the collection of information, in addition to others that do not apply to this study. A study of user attitudes done by Alpert, Karat, Karat, Brodie, & Vergo (2003) found that users insisted on having control of their data and be given access for understanding how it is used to feel comfortable and trust in the resulting personalization. The W3C's P3P (W3C, 2005) proposal is another solution for allowing the user to establish acceptable privacy levels and allowing access to their information only when the site agrees to his preferences, though the details of this proposal are beyond the scope of this paper.

A Proposed System

The author proposes a system that allows cross-domain and cross-session personalization, maintains privacy, and enables portability. In this system, all queries on all sites performed by a single user, as well as every word on every page that user visits, and any accessible metadata associated with items the user interacts with while browsing are collected (excluding demographic or personally identifying information). The information is gathered into a "bag of words" created and maintained without direct user input (unless user chooses to interact with the collected data in prescribed ways); the system requires no explicit input, obviating any extra effort by either the user or the system designer.

This information is then processed to calculate a weighting of the collected terms. The elements that determine the weighting include the time spent on a page—the longer the user spends on a page, the higher the weight of the terms found on that page (though a minimal level of interaction will be necessary to ensure that the user is in fact interacting with the page). In addition, the greater the number of occurrences of a term within the user's "bag," the greater the weight for that term, with a moving window which lowers the weights for terms outside a certain time frame. Items for which special interactions are logged (printing, saving, browsing, purchasing) will also be given a higher weight. As the terms build up, co-occurring words will be clustered, with greater co-occurrence bringing terms closer.

The bag of words then becomes a context in which to orient the user's current query, producing results with increased relevance to user's interests. All of this information is kept on a small, portable device which can be plugged into a computer for greater personalization or remain unplugged for privacy. The user, not the domain, keeps this information, returning the control of the collection of and access to the information to the user, and additionally allowing the information to traverse domains and sessions.

The user is also given the ability to manipulate the data recorded through an interface within the device. He can examine the information by query or domain and turn information from a certain domain or from a certain cluster "on" or "off" so that it is not submitted to a site for inclusion in the recommendation computation. And, for greater transparency, though it is unlikely to occur (Wærn, 2004), the user could directly access information recorded on the device and edit it.

An general example:

The user obtains a small device (i.e. a thumb drive-like device or smart card) that has a unique identity. The device is plugged into a computer while the user interacts with the internet and is removed and transported with the user when they are finished. This device stores only usage patterns, no identifying or security-sensitive information, and the device

itself only records, weights, and clusters the information and then serves this information to the recommender engine housed on sites visited. (The device is not in itself a recommender system.) There is no way to identify the user from the device, and one user can have multiple devices: one for work related interactions, one for personal, one for anonymous, etc.

While the device is plugged in, the user enters a search and the query terms are recorded. The user's interaction with the results set is recorded: links that are clicked in the results list, time spent on each site clicked from the results. (A URL entered directly into the address bar or a new query submission, in which the terms do not cluster near previous search, can be used to identify topic shifts.) Then when a user visits a site that has a system for personalizing its content, this information is made available to the site's recommender/relevance system, giving it a greater body of information to create a context for the query.

Implications and possible applications

The proposed system is a personalization system, or rather an aid to personalization systems, as it tailors the content of the web to an individual user without explicitly interacting with the user. The system addresses many of the issues and concerns raised in previous research. It asks little of the user to make the system work and does not require maintenance unless the user desires it. User behavior captured in the process of browsing and searching intuits the user's interests and uses that information to improve relevancy and recommendations, as web mining has most often proved the most efficient and accurate of all the approaches used by systems to date. The data is only minimally transformed, allowing for greater transparency and easier interaction with the information by the user, and the provided interface for turning data on and off, or removing the device entirely, gives the user greater control over what information is being used or accessed. The system provides the collected information to sites visited to aid in modeling user interests and improving relevancy in retrieval. And finally, storing the information on the client-side

and making it portable allows for personalization to carry across domains, sessions, and machines, and provides increased privacy.

Limitations

This paper has several limitations. As my background is not technical, most of the architecture and framework research on recommender and personalization systems was beyond my reach, and as a result, the findings and proposals within this paper are largely theoretical. Additionally, I did not take current computing limitations or scalability into account.

Limited access to materials was also a problem; several seemingly pertinent articles were published in IEEE publications, which I could not gain access to in the short time period in which this paper was written. The limited time frame also prevented a thorough review of the literature in related disciplines (social psychology, marketing/consumer psychology, behavioral science, etc.)

I also chose to focus solely on the personalization of site content for the average web searcher, though there are several elements of a site that can be personalized (i.e. the structure of the information, the structure used to navigate the site, the presentation layout, and the content).

Conclusion

In this paper, the evolution of today's recommender and personalization systems was surveyed and issues arising from the literature were discussed. In light of those issues, a method for improving personalization while maintaining privacy and control over user-specific information was proposed. Currently, the proposal is purely theoretical, but the issues raised and addressed are vital to advancing personalization on the web.

There are many advantages in altering the current server-side, site-specific personalization approaches: increased privacy, improved personalization, and greater trust from the user. Privacy and control of user data is an increasingly important issue and is bound to become only more so, in light of the recent court case using google records. (Cohen, 2005) The development of client-side, user controlled systems and information has great potential and is the next logical step to allow user to take advantage of the ever-increasing amount of information available on the web, while maintaining privacy.

References

- Alpert, S. R., Karat, J., Karat, C., Brodie, C., & Vergo, J. G. (2003). User attitudes regarding a user-adaptive eCommerce web site . *User Modeling and User-Adapted Interaction*, 13(4), 373-396. Retrieved November 3, 2005, from the ABI/INFORM Global database.
- Balabanovic, M., Shoham, Y., & Yun, Y. (1997). An adaptive agent for automated web browsing. *Proceedings of the 1st international conference on autonomic agents*, Marina Del Rey, Ca, Retrieved November 2, 2005, from <http://hugo.csie.ntu.edu.tw/~yjhsu/courses/u1760/Online/papers/stanford.ps.gz>
- Belkin, N.J. and Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12), 29-38.
- Buchner, A., & Mulvenna, M. D. (1999). Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record*, 4(27) Retrieved November 16, 2005, from <http://www.infi.ulst.ac.uk/~cbqv24/PDF/SIGMOD98.pdf>
- Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12, 331-370. Retrieved October 25, 2005, from the ABI/INFORM Global database.
- Burke, R. (1999). Integrating knowledge-based and collaborative-filtering Recommender systems. *Proceedings of the workshop on artificial intelligence for Electronic commerce*, 69-72. Retrieved October 28, 2005, from <http://josquin.cti.depaul.edu/~rburke/pubs/burke-aiec99.pdf>
- Cassel, L., & Wolz, U. (2001). Client side personalization. *Proceedings of the joint DELOS-NSF workshop on personalization and recommender systems in digital libraries*, Dublin, Ireland, 8-12. Retrieved November 18, 2005, from <http://www.ercim.org/publication/ws-proceedings/DelNoe02/CasselWolz.pdf>
- Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. *IUI '01: Proceedings of the 6th international conference on intelligent user interfaces*, Santa Fe, New Mexico, United States, 33-40. Retrieved November 5, 2005, from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/1096000.1096002>
- Chan, A. T. S. (2005). Mobile cookies management on a smart card. *Communications of the ACM*, 48(11), 38-43. Retrieved November 5, 2005, from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/1096000.1096002>
- Cohen, A. (2005, November 28). What google should roll out next: A privacy upgrade. *New York Times*. Retrieved November 28, 2005, from the *New York Times* website, <http://www.nytimes.com/2005/11/28/opinion/28mon4.html?emc=eta1>
- Cranor, L. F. (2003). 'I didn't buy it for myself' privacy and ecommerce personalization. *WPES '03: Proceedings of the 2003 ACM workshop on privacy in the electronic society*, Washington, DC, 111-117. Retrieved November 3, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/1005140.1005158>

- Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Trans. Inter. Tech.*, 3(1), 1-27. Retrieved November 3, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/643477.643478>
- Furner, J. (2002). On recommending. *Journal of the American Society for Information Science & Technology*, 53(9), 747-763. Retrieved November 3, 2005 from the ASIS&T Digital Library, <http://search.epnet.com/login.aspx?direct=true&db=izh&an=7135357>
- Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61-70. Retrieved November 3, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/138859.138867>
- González, G., Lòpez, B., & de la Rosa, Josep Lluís. (2005). A multi-agent smart user model for cross-domain recommender systems. *International conference on intelligent user interfaces 2005, Beyond personalization*, San Diego, California, USA. Retrieved November 18, 2005, from <http://www.grouplens.org/beyond2005/bp2005.pdf>
- Karat, C., Blom, J., & Karat, J. (2003). Designing personalized user experiences for eCommerce: Theory, methods, and research. *CHI '03: CHI '03 extended abstracts on human factors in computing systems*, Ft. Lauderdale, Florida, USA, 1040-1041. Retrieved November 5, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/765891.766138>
- Kautz, H., Selman, B., & Shah, M. (1997). Referral web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3), 63-65. Retrieved November 3, 2005 from the ACM Digital Library, <http://doi.acm.org/10.1145/245108.245123>
- Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2), 18-28. Retrieved November 5, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/959258.959260>
- Lynch, C. (2001). Personalization and recommender systems in the larger context: New directions and research questions (keynote speech). *Proceedings of the DELOS workshop: Personalisation and recommender systems in digital libraries*, Dublin, Ireland, 84-88. Retrieved November 18, 2005, from <http://www.ercim.org/publication/ws-proceedings/DelNoe02/CliffordLynchAbstract.pdf>
- Mobasher, B., Dai, H., Luo, T., Sun, Y., & Zhu, J. (2000) Combining web usage and content mining for more effective personalization. *Proceedings of the international conference on ECommerce and web technologies (ECWeb)*, Retrieved November 13, 2005, <http://www.datamining.org.tw/paperauto/web%20mining/Web-use-content-personalized2000.pdf>
- Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8), 142-151. Retrieved November 5, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/345124.345169>

- Perugini, S., & Goncalves, M. A. (2002). *Recommendation and personalization: A survey*. Unpublished manuscript. Retrieved October 25, 2005, from <http://homepages.udayton.edu/~perugisa/papers/access/RSsurvey/RSsurvey-sperugin.pdf>
- Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulo, C. (2003). Web usage mining as a tool for personalization: A survey; Nov 2003; 13, 4; pg. 311. *User Modeling and User-Adapted Interaction*, 13(4), 311-372. Retrieved November 2, 2005, from the ABI/INFORM Global database.
- Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., & Edmonds, A. et al. (2002). Personalized search. *Communications of the ACM*, 45(9), 50-55. Retrieved November 3, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/567498.567526>
- Rucker, J., & Polanco, M. J. (1997). Sitseer: Personalized navigation for the web. *Communications of the ACM*, 40(3), 73-76. Retrieved November 3, 2005 from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/245108.245125>
- Ruthven, I., Lalmas, M., & van Rijsbergen, K. (2003). Incorporating user search behavior into relevance feedback. *Journal of the American Society for Information Science & Technology*, 54(6), 529-549. Retrieved November 18, 2005 from the ASIS&T Digital Library, <http://search.epnet.com/login.aspx?direct=true&db=izh&an=9505682>
- Speretta, M. and Gauch, S. (2004). Personalizing search based on user search history. *Submitted to CIKM '04*, Retrieved November 13, 2005, from <http://www.ittc.ku.edu/keyconcept/>
- Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorer Newsletter*, 1(2), 12-23. Retrieved November 5, 2005, from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/846183.846188>
- Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. *ACM SIGIR 2001 workshop on recommender systems*, New Orleans, LA, Retrieved November 6, 2005, from <http://www.sims.berkeley.edu/~sinha/papers/BeyondAlgorithms.pdf>
- Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, Salvador, Brazil, 449-456. Retrieved November 5, 2005, from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/1076034.1076111>
- Terveen, L., Hill, W., Amento, B., McDonald, D., & Creter, J. (1997). PHOAKS: A system for sharing recommendations. *Communications of the ACM*, 40(3), 59-62. Retrieved November 3, 2005, from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/245108.245122>

- Villa, R., & Chalmers, M. (2001). A framework for implicitly tracking data. *Proceedings of the joint DELOS-NSF workshop on personalisation and recommender systems in digital libraries*, Dublin, Ireland, 89-94. Retrieved November 18, 2005, <http://www.ercim.org/publication/ws-proceedings/DelNoe02/RobertVilla.pdf>
- W3C, & Marchiori, M. (2002). *The platform for privacy preferences 1.0 (P3P1.0) specification* (Technical Report W3C). Retrieved November 23, 2005, from <http://www.w3.org/TR/P3P/>
- Wærn, A. (2004). User involvement in automatic filtering: An experimental study. *User Modeling and User-Adapted Interaction*, 14(2-3), 201-237. Retrieved November 3, 2005, from the ABI/INFORM Global database.
- Wasfi, A. M. A. (1999). Collecting user access patterns for building user profiles and collaborative filtering. *IUI '99: Proceedings of the 4th international conference on intelligent user interfaces*, Los Angeles, California, United States, 57-64. Retrieved November 5, 2005, from the ACM Digital Library, <http://0-doi.acm.org.library.simmons.edu:80/10.1145/291080.291091>
- Zan Huang, Wingyan Chung, B. M., & Hsinchun Chen. (2004). A graph model for E-commerce recommender systems. *Journal of the American Society for Information Science & Technology*, 55(3), 259-274. Retrieved November 18, 2005 from the ASIS&T Digital Library, <http://search.epnet.com/login.aspx?direct=true&db=izh&an=12238895>