Personalization and Recommender Systems in the Larger Context: New Directions and Research Questions

Clifford Lynch
Coalition for Networked Information
21 Dupont Circle, Washington, DC, U.S.A.
clifford@cni.org

Introduction

My background is in building large-scale library automation systems and distributed networked information systems; these are areas where personalization has been very slow to arrive, and where recommender systems have met both cultural opposition (in part based in privacy concerns) and technical barriers. I have been a strong advocate of personalization in information retrieval systems; this seems to me to be a logical extension of the secular move towards personal machines and intelligent systems generally, but the promises here are still unrealized. Recommender systems have fascinated me since I first encountered them as a mechanism for automating a very fundamental method of discovering information.

My interests are less in the algorithms used for constructing recommender systems than in the broader questions of how they fit architecturally into large scale information discovery systems, particularly in a highly distributed environment (which relies on standards and protocols for data interchange); how they may interact with and complement other methods for discovering information of interest to individuals, and with the social contexts of these methods; and how they may provide insight into the social processes of information discovery and diffusion. In this talk I will explore all three of these areas and I'll try to highlight what I find to be interesting research challenges that could advance both the usefulness and broad deployment of personalization techniques.

Recommender Systems and Information Discovery

The classic models of information discovery or retrieval are based on searching. One enters a series of search criteria, and they are either compared to the contents of objects in the database (content-based retrieval) and/or to metadata which describes the objects in the database (bibliographic style retrieval); matching objects are returned, most commonly these days in some ranked order based on closeness of match. Responding to a query is purely between the user posing the query, the objects, and perhaps the people who have assigned descriptive metadata to the objects. Other information users aren't involved.

Recommender systems try to automate aspects of a completely different information discovery model where people try to find other people with similar tastes and then ask them to suggest new things. Actually, most recommender systems do something just a little different; they really hide the people (in the interests of preserving privacy and anonymity) and just pass on the recommendations derived from the opinions or actions of those with similar tastes.

There is a third set of approaches that are starting to emerge based on reputation management; this focuses on the identification of people that you respect (or that someone you respect respects in turn) and then using the opinions or actions of these selected individuals as a basis for recommendations. The reputation based model captures a different aspect of socially based information discovery, where one makes assessments about people in a more general way rather than comparing your behavior to theirs directly, and has its roots in social practices involving reviewers or "opinion leaders".

And there's yet another approach that deals in popularity; you may want to know about the most popular works within a given community as another means of finding out what you should pay attention to. This has precedents in best-seller lists and has seen some interesting applications recently in Amazon's purchase circles. But a variation on this idea is also showing up in search engines such as Google, which uses a content-based search approach but then assigns a higher weight to sites that are linked by many other sites in ranking the search result, on the basis that these sites are more authoritative.

All of these approaches are valid and relevant; no one of them solves the full range of information discovery problems in a world of too much information. One major research challenge is to ask how we can make these varied approaches work together most effectively. Amazon, which has been a real pioneer in this area, simply presents all of the methods simultaneously, which often becomes overwhelming; I suspect that few Amazon users pursue all of the different paths to additional recommendations that are offered. Google, as indicated, is much more subtle, though what it does is much simpler in some sense. Clearly, recommender systems can be used to find people who might be of interest in a reputation-based system; searching can be used to supplement the output of recommender systems; and popularity based identification has close relationships to recommender systems. But we need a much better framework for synthesizing the various approaches, and a lot more experience in presenting systems that offer multiple approaches to users at different levels of sophistication and with differing time horizons.

As we think about designing systems that help people to cope with information overload, we need to carefully investigate what we are doing when we retrieve or recommend information. Systems like Amazon use purchase decisions as a surrogate for liking a work; the idea is that they will suggest works that they think you'll like (purchase). In other information retrieval or discovery settings, the key issue is awareness; you want to be aware of certain kinds of information, even if upon examination you determine that you don't like it, or disagree with it, or think that it's rubbish. At present many of our information retrieval systems conflate relevance with quality (as do many users), and socially-based systems are particularly prone to this. Some careful work is needed to disentangle these, and this may have important implications for personalization and recommender systems.

Personalization in a Distributed Information Environment

Today, personalization is something that occurs separately within each system that one interacts with. Recommender systems are one technique for personalization; in essence the personalization occurs slowly as each system builds up information about your likes and dislikes, about what interests you and what fails to interest you. There are numerous other personalization techniques; most of these rely either on collection of system usage history which is then employed to change the behavior of the system, or on the user taking the time and trouble to explicitly personalize the behavior of the system in various ways by setting parameters, making selections or engaging in dialogs with the system.

There are several problems with this model, at least from the user's point of view. Investment in personalizing one system (either through explicit action or just long use) are not transferable to another system. (Of course, from the system operator's point of view, this may be very desirable; it increases switching costs for users and thus helps lock in a user base.) Information such as likes and dislikes or usage patterns are scattered across multiple systems and can't be combined to obtain maximum leverage. And the user does not have control of the information bases that define his or her "profile". If you want to buy books from multiple online booksellers this is annoying. But if we are concerned with developing information discovery systems to assist users in a world of information overload, these problems are critical. People obtain information from a multiplicity of sources, and personalization has to happen close to the end user; this is the only place where there is enough information to do personalization effectively, to keep track of what's new and what isn't, what has and has not proven useful. The user needs to become a hub and a switch, moving data to allow accurate personalization from one system to another.

It's reasonable to think about recommender system algorithms as distinguishing features and specific advantages of individual systems; it's also reasonable to think that aggregate databases of objects and opinions might be unique to each information finding system (and I use this term very broadly, to include

for example shopping systems). But for the end user it would be a much better world if he or she could simply have a program pass a collection of history and opinion data to each system he or she wishes to interact with and instantly obtain personalized behavior and where appropriate recommendations from it.

There has been some work along these lines. At a very mundane level, it has been clear for some time that one of the biggest barriers to getting consumers to make purchases on the net is the tedium of entering personal name and address and credit card information over and over for each new site, registering for the site and obtaining (and remembering) a username and password to be used on future visits. This has led to ideas like electronic wallets which can store this information in a structured form and transfer it to a new site on demand. In the recommender system related area, there have been experimental systems that look at web browser bookmark files, and the Alexa system (now owned by Amazon) actually follows web browsing trails.

Much more can potentially be done here. Imagine that users could maintain a series of local (personal) databases that defined their trust systems (we have the very primitive beginnings of this in the identity trust codified in systems like PGP or certificate collections in web browsers); their reputation and rating systems; their preferences of various kinds; interest profiles; and their opinions about objects of various kinds (web sites, books, sound recordings, films, people, etc). Imagine that there were standards for transferring this kind of information to systems that one wanted to interact with, and also standards for updating the locally held, personal databases to reflect actions taken in remote systems (such as purchasing a book, or visiting a web site); and, of course, tools to edit the personal databases and to establish rules about their update and dissemination. This would potentially permit instant personalization of any system that one wanted to interact with; it would allow users to visit new sites and immediately obtain potentially useful recommendations from them.

Moving from this vision to implementation is a very complex research problem, with lots of difficult details, but one that I think merits investigation. I would just note in passing that a good deal of infrastructure and numerous standards are needed to make this work; for example, if one wants to move opinions about objects between systems, one needs a global way of referring to objects (URIs for web pages or International Standard Book Numbers or similar identifiers for books; for music, or journal articles, or other types of content matters get much more complicated.) If one wants to deal with reputations of actors who may be present in multiple systems, one needs global identities for these actors (such as email addresses); public key infrastructure will also play a role here if the identities are to be trusted. This has interesting interactions with social practices involving the use of (perhaps multiple) pseudonyms with and across systems, and with privacy. Also, the personal local database for each user contains a wealth of information (even though the user presumably has good editorial control over it) and users are going to want to be quite careful about what parts of it are shared with what systems under what conditions (and perhaps something like P3P, perhaps augmented with some legal requirements on systems to require that they follow their advertised privacy policies, may be necessary to make users comfortable enough to employ such a system). And, of course, to be really useful virtually all of the work has to happen automatically in terms of information transfer between the user's client and the various systems, while still leaving the user in control.

Personalization and Privacy

Libraries have always been very strong defenders of privacy. A key best practice for libraries in recent years, at least in the United States, has been to destroy patron-related data unless there is a compelling business need to keep it. Thus, for example, circulation systems typically break the link between a patron and a book that has been borrowed when that book is returned. While the book is out on loan, the library needs to keep a record of who has borrowed it in order to protect the library's assets (and thus, with an appropriate court order, it is possible to find out what books a given patron has currently borrowed); but the library simply doesn't keep a record of what has been borrowed and subsequently returned by a given patron, and thus cannot be compelled to reveal this information. It simply doesn't exist to be subpoenaed.

This explains why one doesn't see recommender systems in libraries very often.

Yet it also frames a series of interesting design challenges. It should be possible to offer the benefits of personalization while still maintaining a very strong and principled position about user privacy. One obvious possibility is to allow users to opt into a history database with informed consent; presumably such a database would be used only for recommendations, and not (without further consent) to make contacts among individuals, or for reputation-based information identification. Presumably such a database would be protected to the limits of the law, and only released under a legitimate legal order. And presumably one could design an interface that would allow transactional opt-out (i.e. don't include the borrowing of this particular book in my history), or perhaps require transactional opt-in if one wanted to be truly conservative about collecting information. We know very little about how to really design such a system and to make it understandable to casual library patrons, or about what is acceptable to such patrons; research is clearly needed here.

It may be possible to do better. Suppose one explicitly separates a recommendation database from a circulation database. Let users keep (and, if they wish, edit) lists of books that they've borrowed and found useful (again, here, we see the ambiguity between relevance and quality discussed earlier). If library patrons could simply submit these lists into the recommendation database (without necessarily any link to their own identity that's maintained by the library) it would be possible for patrons to submit queries against this recommendation database (including their own list of books in the query) and obtain recommendations without the library holding personal information that could be disclosed. There are problems to be worked out; for example, since the recommendation database is based on contributions rather than actual, known to be valid records of user behavior, malicious people could submit inaccurate lists of books to the database and degrade its quality. And building a user interface to such a system that hides most of the work from the user while still leaving the user informed and in control is an interesting challenge.

This sort of privacy-friendly recommender system builds on the kind of distributed architecture ideas discussed earlier. Indeed, people could integrate lists of books that they had purchased and books that they had borrowed to provide a more comprehensive picture of their opinions, thus improving the quality of the recommendations. Presumably libraries are a well-trusted and honest party that might provide an excellent testbed for distributed recommender systems where the critical data is held and maintained close to the user rather than in external systems (in this example, the library holds only anonymized opinion data records). We need to gain some experience with such prototype systems.

Understanding Social Processes of Information Dissemination

During the 20th century a certain amount of mathematical modeling work has been done on the spread of rumors and ideas. This turns out to be closely connected to research in the modeling of the spread of infectious diseases and plagues. Unfortunately, the mathematics gets very nasty very quickly, even with relatively simple models, and the models don't tell us that much -- at best, they provide some insight into when a disease will spread widely as an epidemic, and when it will stay limited to a small subset of the population or die out entirely. To keep the mathematics tractable, these models tend to assume homogeneous populations, or populations made up of a few different homogeneous subgroups. We have other bodies of observational research, such as bibliometrics and citation analysis which tell us something about rates of diffusion of knowledge as represented by citations.

Yet understanding the spread of information or knowledge through populations is of great importance. We don't understand, for example, why some works become hits and capture the interest and imagination of communities while others are largely ignored. No doubt some of this has to do with intrinsic quality, but some of it has to do with whether key opinion leaders become aware of the work and promote it, with how quickly and how broadly knowledge of the work spreads and the paths that this knowledge follows. And there are subtle second-order effects: when works become too broadly popular, for example, this is a negative within sophisticated, elite communities.

Currently, we tend to think about recommender systems as stateless computation engines in the sense that given a database and a user profile they perform a computation and provide some recommendations to that

user. But in fact new objects are entering the database all the time, and users become aware of them not only through recommendations, but also through the other information finding approaches discussed earlier, and also through exogenous factors such as advertising.

A great research challenge is to construct databases and software systems that model large numbers of individuals and objects and which take into account how these individuals become aware of and rate new objects, considering not only recommendations but also advertising, reputation based discovery, and other processes. Such systems would ideally be able to model and identify dynamic trends such as emerging "hits" and to take these dynamic behaviors over time into account in making recommendations.

Probably the place to start is with modeling and simulation; once we have built accurate models of the introduction and dissemination of works like books or sound recordings throughout a population and can identify emergent behavior patterns within these models, these identified behavior patterns can be applied to improving our ability to make recommendations. These systems will look a bit like recommender systems, but they will include time as a factor, and will presumably employ multi-agent simulation techniques. I know of virtually no good research in this area, even though we have the computational resources, and even though work on recommender systems may offer a significant part of the algorithmic basis for developing such systems.

Finally, let me point out that as such models mature, there are additional avenues of research that need to be pursued. For example, we need to understand the extent to which such systems are vulnerable to deliberate gaming or subversion, or less threateningly, what steps should be taken to try to ensure that a new work becomes a hit. Also, it would be interesting to generalize these systems beyond the context of single decisions (such as whether to purchase a book or a sound recording); for example, one could look at modeling investing decisions (stock purchases and sales) in this context. Here one finds complex feedback phenomena. It would be valuable to understand whether such phenomena move us into an entirely different modeling regime or whether information dissemination models remain relevant.