TECHNOLOGY QUARTERLY

CASE HISTORY

United we find

Mar 10th 2005
From The Economist print edition

Computing: Collaborative filtering software is changing the way people choose music, books and other things, by helping them find things they like, but did not know about

EACH year, thousands of films are released and tens of thousands of books published. A big city has thousands of restaurants. How does one deal with such abundance? Reading reviews of films, books and restaurants can provide a guide, but there are more reviews than one has the time to read, and you cannot be sure that the reviewer's taste matches your own. Word-of-mouth recommendations can help in that regard; friends, after all, are often friends because they share similar tastes.

For many people, technology now plays an increasing role in making such choices and navigating through large numbers of alternatives. But while this might sound like a job for an internet search engine, keyword-based search engines (such as Google) have a fundamental constraint: they can only help you find something if you already have an idea of what it is. Two people's idea of “good music” may differ substantially, but Google would return the same results to both of them. To find things you might like, but are not already familiar with, requires a different technology, known as “collaborative filtering”.

This increasingly pervasive technology looks for patterns in people's likes and dislikes, and uses those patterns to help people find things they did not know they were looking for. Computer scientists term this task, in a welcome respite from jargon, “find good things”. Collaborative filtering also has the power to do the converse, “keep bad things away”, for instance by filtering unsolicited commercial e-mail messages, or spam. Systems that use collaborative filters to keep spam away already exist, though there are many other ways to do the same thing. Finding unknown good things, however, can at present only be done using collaborative filtering.

The idea has been around for over 15 years. Early prototypes at Xerox PARC, a corporate research facility in Palo Alto, California, date back to the early 1990s. But the delay between the genesis of the idea and its widespread implementation turned out to be quite long, for two reasons. First, a successful collaborative-filtering system is computationally demanding and becomes rapidly more so as the number of users increases. A prototype system might have a few thousand users, which is manageable, but a real-world system will have millions—and the difference in scale introduces new challenges, which have been only recently overcome.

The second reason is that for collaborative filtering to reach its potential, it has to be seamless. Early incarnations of the technology required users to state their tastes explicitly, by going to special websites and filling in on-screen forms, before being presented with recommendations. But a system that is integrated into an online store, and recommends one product to you as you are buying another, is far superior because it requires no intervention by the user. The business challenge of collaborative filtering lies as much in creating a seamless interface as it does in generating the right suggestions—so the technology has had to await the widespread adoption of internet shopping, to which it makes a natural adjunct.

Now that both of these conditions have been met, however, collaborative filtering has started to pop up all over the place. Anyone who shops online is used to having books and music recommended to them as they browse and buy; the technology is also used on DVD-rental sites to recommend films. Having changed the way many people choose books, music and films, collaborative filtering is moving into new areas. It can help people to choose which programmes to watch on television, which restaurants to go to, even where to go on holiday. But how does it work? And should users be worried about collaborative filtering's impact on privacy, or the possibility that recommendation systems can be rigged?

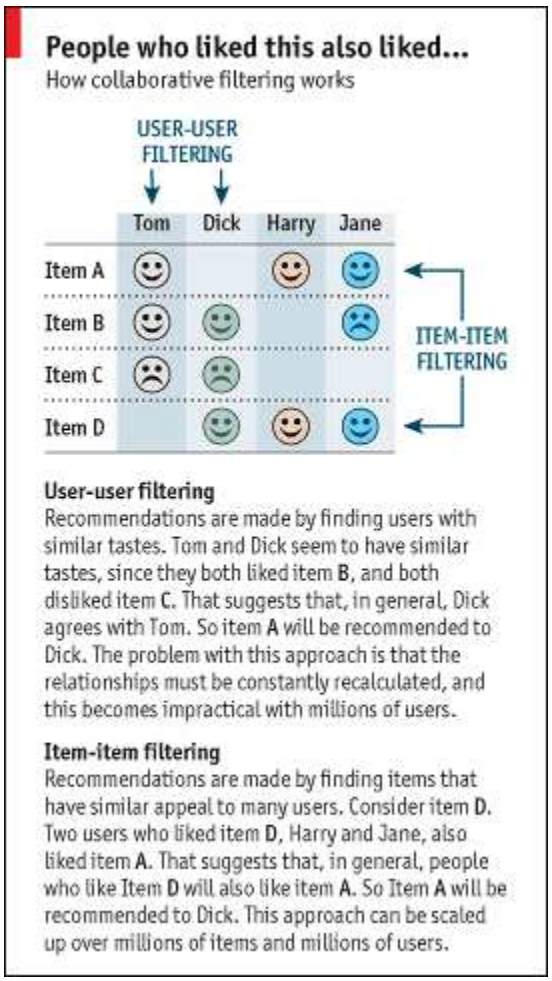
Tell me what I want

Collaborative filtering starts off by collecting data on individuals' preferences. This can be an explicit process, by which a user ranks a book (or CD, or restaurant) on a numerical scale, typically on a scale of one to five. It can also be an implicit process—a purchase, for instance, is a clear indication that an individual is interested in the item in question. But implicit measures can also be more subtle; for instance, the amount of time spent viewing a web page, or even just the "clickstream"—the sequence of links clicked on by a person browsing on the web. These different methods can then either be aggregated into a single score, or stored separately to allow more detailed analysis. And sometimes, consumers will be asked to score the same item in different ways—for instance, what one thought of the food at a restaurant, and what one thought of the service.

The result is a mountain of data, the size of which is the main challenge when it comes to searching it for patterns. But things are helped slightly by the "sparseness" of the data. The vast majority of items do not have a ranking, implicit or explicit, from any given user. Even the busiest users have rarely ranked more than 1% of the items. Amazon, for instance, sells over 2m books through its online store. The sparseness of the data is a saving grace, because it allows various mathematical techniques to be brought into play which vastly speed up the process of generating recommendations.

There are two basic ways of doing this. The first idea was proposed in 1992 by Dave Goldberg and his colleagues at Xerox PARC, who also coined the term "collaborative filtering". Their approach was to recommend items to a user based directly on that user's similarity to other users. If I liked a book and you liked the same book, then I am likely to like things you like. However, this so-called "user-user" collaborative filtering turns out to have very poor performance when scaled up to millions of users. The problem is that the relationships between users must be constantly recalculated, which is too computationally costly.

This is why Badrul Sarwar and his colleagues at the University of Minnesota, in Minneapolis, pioneered so-called "item-item" collaborative filtering systems in 2001. (Other groups, including Amazon, had similar ideas around the same time.) Item-based filtering works by periodically taking a snapshot of everybody's item rankings. It then computes the similarities between items as follows. For a given item, such as a book, it finds all the other items that were also ranked by people who ranked the original book. The filtering software then looks for other items that were given a similar rank to the original item by many people (see diagram).



The details of what it means to be “similar” vary from system to system. Indeed, one key aspect of getting a system to make good recommendations is having an appropriate mathematical definition of similarity. The simplest approach, which is to measure the average difference in rankings, works fairly well. And there are various tricks that can be used to increase performance, such as introducing a bias against very popular items: there is little value in recommending a bestseller such as “The Da Vinci Code” to people, because they have probably heard of it already.

The benefit of item-item filtering is that this elaborate similarity calculation need only be done infrequently. Then, when a user ranks a new item—by purchasing it, ranking it, visiting its web page, or whatever—the system can simply call up a pre-calculated list of items that are also likely to appeal to that user. This is what allows Amazon to handle over 30m customers and give instant recommendations, even as the list of items that have been ranked by a customer changes, since merely calling up the web page for a particular book counts as a ranking. All the calculations are done by Amazon's powerful server, which creates a list of recommended items and seamlessly stitches that list into the next page sent to the user's web browser, neatly excluding items they have already purchased.

The TiVo personal video recorder, on the other hand, which can recommend programs based on your (and other users') previous viewing habits, works in a different way: the recommendations are generated by each TiVo box, not by a central server. The server generates a matrix that relates the popularity of different shows to each other, akin to the pre-calculated item lists used by Amazon to generate recommendations. But the task of making recommendations is then left to the individual TiVo boxes, which use that matrix, combined with the data they have stored locally about the viewer's preferences, to suggest shows that might be of interest. As well as unloading much of the work on to the individual boxes, this has the added virtue of preserving privacy: the central server never stores data about individual users, just aggregated data about viewing trends.

That is just one way to address what is, for privacy advocates, a major concern about collaborative filtering: that to make recommendations, it is necessary to gather information about many people in a central repository. But there are other ways too. Indeed, a scheme proposed by John Canny, of the University of California at Berkeley, shows that it is, in fact, possible for a group of individuals to pool their opinions and generate recommendations without revealing their own personal preferences to others.

“A search-engine user hunts alone; the user of a collaborative-filtering system is part of a crowd.”

Each individual encrypts their data using what is called a one-way hash—a function that is very easy to compute in one direction, but virtually impossible in the other (without a key, at least). The computations are then performed using the encrypted data. This is possible because many modern encryption schemes have the helpful property that performing calculations on encrypted data produces the same answer as manipulating the unencrypted data and then encrypting the result. The resulting matrix of recommendations is then decrypted incrementally, since each user can only decrypt a small part of it. Eventually, the whole matrix is decrypted and made available to everyone. But, says Dr Canny, “at no stage does unencrypted information about a user's preferences leave their own machine.”

This sort of scheme has the advantage, he says, that users can store personal information themselves, without having to surrender it to a central authority (such as an online retailer), while still benefiting from the power of collaborative filtering. At the moment, users' personal information is sprinkled around on several different sites. Dr Canny worries that this favours retail monopolies, since they will have the most data from which to generate recommendations. His scheme demonstrates that personal data could, instead, be aggregated by users themselves. Your taste in books can then be used to generate recommendations, by aggregating your purchasing histories from several online bookstores.

Fiddling the filters

A second concern about collaborative filtering is that as it grows in importance, people may increasingly try to manipulate it: publishers, for example, might start recommending their own books. Last November, Michael O'Mahony of University College, Dublin, published a paper demonstrating that even today's most advanced collaborative filtering systems are not all that robust when subjected to malicious users seeking to subvert their ranking systems. None of the existing systems is explicitly designed to combat malicious use. Can such “recommendation spam” be prevented?

Nolan Miller, of Harvard University's Kennedy School of Government, and his colleagues believe that it can, and have outlined a way to do it. Their scheme uses probabilistic techniques to determine whether a score is likely to be “honest”, by spotting unusual-looking patterns in scoring. Dozens of accounts created on the same day, all of which give high scores both to a bestseller and a new book, for example, might be an orchestrated attempt by a publisher to get fans of the former to buy the latter. Honest users

are rewarded, and dishonest ones punished, through a points-based system akin to a loyalty scheme, so that honest users might earn discounts or store credit.

The scores used to compute recommendations are the ones corrected for honesty, not the original, potentially malicious scores. Dr Miller's system is not yet ready for commercial application; it makes assumptions about the statistical distribution of people's recommendations that may not correspond to their real-world behaviour, for example. But it points out a line of research that could preserve the integrity of collaborative-filtering systems under attack. If the rise of spam e-mail is any guide, it makes sense to think about such problems now, before they become widespread.

But even if the problems of privacy and dishonesty can be overcome, there may be a limit to how accurate the recommendations made by collaborative-filtering systems can be. This arises from the fact that people's opinions change. You may enjoy a new album at first, and give it a good score, but change your mind after a few weeks once the novelty has worn off. But your old score still stands.

A recent study by Jonathan Herlocker of Oregon State University and his colleagues evaluated several film-recommendation systems based on collaborative filtering. Using a five-point scale, it compared the scores users would be expected to give particular films, based on their known preferences, with the scores they actually gave. The predicted and actual scores differed by at least 0.73 points. Dr Herlocker speculates that this might be evidence for a fundamental limit to the accuracy of recommendation systems based on collaborative filtering. There is no point in making suggestions any more finely tuned than the variations in an individual's own opinions. Dr Herlocker may well be correct, or the technology may just have further to go.

But the value of collaborative filtering has, in any case, already been established. It helps people find things they might otherwise miss, and helps online retailers increase sales through cross-selling. Where the user of a search engine is on a solitary quest, the user of a collaborative-filtering system is part of a crowd. Search, and you search alone; ramble from one recommendation to another, and you may feel a curious kinship with the like-minded individuals whose opinions influence your own—and who are, in turn, influenced by your opinions.