Collaborative Filtering: Research or IT?

The information technology revolution impacts the practice of marketing research in many ways. A new Internet application known as collaborative filtering is a significant development in this regard since it has the appearance and character of marketing research, along with some important differences. It may pose a threat or an opportunity to conventional research methods. It bridges the gap between research and application in marketing, thus has implications far beyond research itself. It is important therefore to understand what it is and how it is applied as a basis for dealing with it in the future.

This article gives a brief summary of some of the rationale behind the method and its origin, describe how it works from the customer perspective, raises some methodological issues, and then points to some of the implications for research and marketing practice.

What is it?

Collaborative filtering is a method that uses intelligent software agents to gather and analyze preference information that is provided by an individual on a web site. The goal is to provide recommendations directly to the individual about things that would be of interest to the individual.

In marketing parlance, a seller of a line of products wants to know what subset of products should be part of an offer tailored specifically to the customer’s needs and wants in order to maximize purchases. Successful prediction can lead to much more efficient marketing than the alternative, undifferentiated mass marketing approach. The elusive goal of one-on-one, relationship marketing could be significantly advanced through this kind of web-based technique.

Consider the cases of the bookseller. It would be extremely desirable to know the reading tastes of an online customer who is browsing at the store's Web site. For example, if you know that the customer likes business books, especially in marketing, and has made an inquiry about a new brand-management book, then there are several other brand management and brand-equity books you might want to offer up.

One online bookseller asks whether you want to participate in its recommendation service. If you go to that site you get a series of questions. First you are asked about categories of books in which you are interested, e.g. fiction, history, science fiction, home repair. Once you select your categories you are asked whether you have read and liked a list of different books. A simple, five-point categorical scale is used, ranging from “loved it” to “not for me.”

Once you have rated 20 books in total, you are ready to receive recommendations. They are displayed in a separate place on the page, and can be stored for you in case you want to leave and come back at a later point. This makes it very convenient to make a purchase of a recommended book online, although there is no obligation to do so.

You have the opportunity to revise your ratings; perhaps you changed your mind about how good a book was, or you purchased and/or read a new book. You may re-submit your answers and receive a revised set of recommendations.

Now extend this kind of thinking to all the published titles in all content areas. Extend it to music, videos, periodicals, and so on. What started as a method for identifying published articles and information that would be relevant to the interests of the searcher has changed into a set of “virtually” endless commercial possibilities.

The components of this system include an ever-increasing database of user preference information, a friendly Web interface that induces consumer cooperation, and a recommendation engine.

The database is the repository of all the preferences of all the people who have previously come to the Web site and provided information. The larger it grows over time, the more finely the database can be segmented to find people with closer matches to each other.

The Web site is the attraction mechanism. People are not randomly sampled, but rather show up when they choose to shop/browse. It is important that the site is a pleasant experience and that it encourages cooperation in ways that don’t make the individual uncomfortable divulging information.

The analytical tools store and summarize the data. The data are analyzed with filtering and statistical methods, for example, correlation analyses. These models create the predictions and recommendations and offer them to individuals in real time. The data and models are continuously updated for ongoing use. The providers of collaborative filtering systems offer the engine and consulting services to implement the tools.

Is it research?

Collaborative filtering emerged from the information-processing field as a way to deal with the problem of information over-
load. How does an individual sort through massive amounts of information content that are available by searching with identifying key words? Key word searches often are insufficient to produce a small set of topics to read or consider directly.

If a computer is to further help the individual find what is relevant it needs to know something about the individual. Yet, at the time of the first visit to the Web site the computer knows nothing. This is referred to as the "cold start" problem. Why not therefore take advantage of the fact that other people have had the same information need and have searched and found new information that is relevant to them. If we could automate word of mouth referrals a person could save a lot of searching time. The goal is to find what information other people have found useful and recommend it.

The key logical leap seems to be that the individual is a member of a population who has many attributes. Pooling information across a sample from that population provides insight into how the whole population behaves and can be used to predict how individuals will behave. Said differently, database searching can make use of statistical inference.

One way to look at collaborative filtering is marketing research embedded in a customer transaction process at a Web site. Information technology enables this application to work. Its purpose is to sell something to an individual, rather than to learn something in order to develop marketing strategy and tactics. While it is clearly different from research, it appears to use methods that are at least analogous to, if not identical to, traditional research methods.

Questionnaires are administered, responses are used to build predictive models, and recommendations are given to suggest who should be offered what product. It uses the same logic that is used in segmentation and targeting methods: find people who share similar profiles, offer them products that they have yet to purchase, but others like them have already acquired.

Collaborative filtering has some limitations as research. One limitation is that it may suffer from sampling bias. While it may capture many of the appropriate people, i.e. the seller's potential market, it may miss some shoppers who frequent other sites, and it may not represent those who visit in the proper proportions. For example, should shoppers be weighted in proportion to frequency of visits, purchases, or demographics of the population?

A possible counter to this classical sampling argument is that the purpose is to achieve sales or customer relationship goals immediately rather than understanding the true level of preference in the population. The criterion for success is lift in profitable sales, rather than statistical criteria such as unbiasedness or consistency. Additionally, the effectiveness is being continuously adjusted, presumably improved, as more data become available. This is analogous to sequential sampling in which conclusions are drawn after each new observation is recorded.

Another limitation relates to the domain of products (or, generally, the content items) that is evaluated. In current applications the predictions and recommendations are made only for products that exist and have achieved some amount of actual sales, e.g., books that have already been purchased and read. In contrast, marketing research has the capability to simulate products that don't currently exist, using techniques such as concept testing and experimental design. Many marketing research applications do not require (nor could they obtain) a calibration against actual sales.

A possible rejoinder to this argument is that collaborative filtering has, at the moment, a narrowly defined scope. Within that scope it is highly accurate, and in particular, potentially more accurate and more useful than conventional forms of marketing research.

**IS IT MARKETING?**

Another way to look at this phenomenon is to note how many marketing functions it actually does or could perform. Collaborative filtering is more than a research technique or an information technology application. It plays or likely will play a role in branding, advertising, pricing, and channels.

One of the values of brands is that they reduce searching and searching costs of consumers. They signal quality and consumers use them to filter out alternative products they might rationally consider. Are intelligent agents really brands in action? They could provide quality claims and deliver against them, for example, guaranteeing the best product recommendations. Could they develop brand names of their own? Will they reduce the importance of or compete with existing brands?

The owner of the database has something of great value to advertisers. Information about the preferences of large numbers of specific individuals should be more effective than broad demographics used for mass media buying. There are potential privacy issues due to the fact that detailed records of individuals are being kept and the individual may not want all their information revealed. One possible solution is to ensure that informed consent is obtained from each individual before information is disseminated.

The recommendation engine itself could be a vehicle for advertising. Suppose a set of five recommendations is derived through the computer analysis. One could envision that some recommendations are highlighted more than others, e.g., listed first, given more visibility on screen. Perhaps subtler relative weightings of alternative recommendations could be given based on the price paid by the advertiser.

A natural extension of the search for the preferred content is the search for the best price for a particular item of content. There are Web software applications in which a search is conducted for the lowest available airfare as the time of departure approaches. Others search for the best price for books and compact disks. This type of application may have the effect of reducing prices in general, increasing the perception of products as commodities, and perhaps eroding current brand preferences.

The implications for channels are potentially huge. If shopping is reduced to searching the Internet the role of retail distribution will be radically altered. While it is premature to think that malls and other retail outlets will disappear any time soon, more and more shopping can and is being done online, and it is often better informed, faster, and cheaper.

**WHAT'S NEXT?**

It is far from certain where this particular information technology application
will go. Will it prove to be accurate in predicting what consumers will want to buy? Will it be able to do so in a way that adequately protects privacy?

At the very least, marketing researchers should understand what it does in order to be able to evaluate its impact. It may compete with traditional research methods. If a recommendation engine can accomplish product offer tailoring and targeting to attractive customers, then perhaps there will be less need for traditional market research.

Alternatively, the impact of this technology could lead to the conclusion that "research" is even more critical to marketing and business in general. What is needed is for researchers to participate in developing and implementing these kinds of applications to make them better, i.e., more accurate, insightful, and effective. In this view the researcher should welcome the influence of technology (which is probably going to increase in any case) and strive to harness it. Is the right set of variables being used to predict? Are scales reliable and valid? Can the information searching process be understood behaviorally? Does it achieve long term customer relationship development? Does it increase profitability?

Another reason to understand and master this type of application is that other information technology influences are likely to come along, soon. There are examples of new product development taking place on the Internet. In some market situations, prototyping and in-market trials, enabled by information technology, are replacing the more comprehensive marketing research. In sum, it is best to anticipate technological changes and understand what drives them than to observe them from a distance until they have taken hold.

**ADDITIONAL READING**

The best references are available online. Search for collaborative filtering, recommendation engine, and intelligent agents.

---

**DEVILLE CONTINUED FROM PAGE 33**

separate data extract procedure). This functionality is provided by a decision tree implementation that is based upon and which generates SQL.

SQL is the industry-standard query and reporting language that forms the underpinnings of all major database management systems including Oracle, Microsoft SQL, Informix and Sybase. If you want to access enterprise data, SQL is the favored retrieval tool. More and more extensions are being introduced to SQL to expand its role as a query tool into reporting as well. SQL has no analysis capabilities, however; so an environment that offers analysis capabilities and is integrated with SQL would be well placed to serve as a complete enterprise-wide query, analysis, and reporting tool. This is the position that KnowledgeSTUDIO has staked out for itself.

The level of integration that KnowledgeSTUDIO offers with SQL is remarkable and unique. All the data that are fed into the analysis engine is accessed through SQL. KnowledgeSTUDIO generates the SQL so it may be saved, modified, and extended. Since SQL is the industry-standard mechanism for data access and retrieval, all database vendors invest time and resources to ensure that it works reliably and efficiently in their database environments. In data warehousing applications, the vendors work hard to ensure rapid SQL queries and data access across huge (terabyte, petabyte) collections of data. Since KnowledgeSTUDIO exploits SQL as its native data access mechanism, it is in a position to routinely piggy back on any efforts that any database system vendor makes to improve access speed in large collections of data. In fact, several vendors are working in this area—notably Tandem Computers, Informix and Whitecross (of the United Kingdom). ANGOSS has relationships with all these vendors and can demonstrate a correspondingly powerful facility to conduct exceptionally fast analysis sessions with huge collections of data. This "in place" data mining and analysis capability is sometimes viewed as the ultimate goal of data warehousing and data mining (although many vendors, such as SAS, promote a sampling-based approach instead).

**SUMMARY**

The current KnowledgeSTUDIO release from ANGOSS International Ltd. includes improvements to the interface and adds new data mining algorithms (primarily neural networks). The important thing about KnowledgeSTUDIO is that it exploits the DCOM and SQL architectures. It is therefore in a position to run an analysis on the most effective computer platform (client or server) and can plug into the most effective SQL engine (or any SQL engine) in order to retrieve the data for analysis. It is in a position to receive and exploit any DCOM-compliant functionality that it may be missing. And it can contribute its DCOM to any DCOM-compliant application. It is therefore possible to conduct a data retrieval and analysis session from an Excel work book-based application, for example. It is also possible with Visual Basic, for example, to use, extend and modify KnowledgeSTUDIO components to create custom applications. These applications may include DCOM-compliant data mining components provided by KnowledgeSTUDIO or any other DCOM compliant component vendor.

The software components, including the analyst workstation, Software Development Kit and In-Place Mining database access engines are available from ANGOSS International at the following prices:

KnowledgeSTUDIO Workstation @ $25K
US for five seats
KnowledgeSTUDIO Software Development kit @ $25K US for five seats
KnowledgeSERVER In-Place Mining
Driver @ $38K US for first driver
KnowledgeSERVER In-Place Mining
driver @ $1K US for add-on drivers

The ANGOSS web site is www.angoss.com. They may be reached at 416-593-2400. Their address is: 34 St. Patrick Toronto, Ontario, Canada