Data Mining for Health Advice

By Shad Kish

You may have heard the computer buzz word data mining—a pattern-seeking process that Wikipedia places at the crossroads between artificial intelligence, machine learning, statistics, and database systems. Even if you have missed the term, you have almost certainly witnessed the commercial results of this “mining” in your local supermarket and on your own smartphone, tablet, laptop, or desktop. Academic researchers have been actively mining data from nearly every domain, as well, and important public health applications are rapidly emerging.

Perhaps you’ve noticed that supermarkets sometimes display unrelated items on the shelves next to each other. Wine and cheese are a common example, and of course they have an obvious gastronomic association. But what about beer and baby diapers—another pairing you might spot at Walmart? The hidden key to this seemingly illogical twosome is data mining, a prime research topic for Dr. Hui Yang, an associate professor of computer science at San Francisco State University. Computer scientists working for large chain retail stores can employ data mining techniques to recognize and extract meaningful patterns from customers’ purchase behavior.

This data, in turn, helps store managers to increase their sales by selecting and pairing sometimes-unlikely sets of products and shelving them together to promote sale through association or convenience. The results of data mining also help customers to make decisions more quickly and easily while shopping. During a recent interview at her ninth-floor office in Thornton Hall, Yang was able to further illuminate the “beer and baby diaper” example. Walmart’s computer programmers applied data mining techniques to recognize their customers’ purchasing patterns, she says. To do so, they wrote a computer program that assigns each customer a virtual “shopping basket” capable of tracking all of the items the customer purchases during each store visit. Computer researchers then analyzed tens of thousands of such baskets and millions of individual items. Their analysis led to the surprising discovery of a positive association between purchasing baby diapers and buying beer. Further analysis revealed that it is mostly husbands who purchase those items in the same “basket.” A plausible explanation for this purchase pattern, says Yang, is that husbands sometimes babysit, watch TV sports, and drink beer at the same time. This data-mined insight then informed the decision of marketing managers at Walmart to shelve the items side by side to increase the sales of each.

Yang offers a similar example from Amazon.com, again based on data customers’ co-purchasing histories. “Have you ever noticed the following feature on Amazon.com: “Customers Who Bought

Health Advice

You may have heard the computer buzz word data mining—a pattern-seeking process that Wikipedia places at the crossroads between artificial intelligence, machine learning, statistics, and database systems. Even if you have missed the term, you have almost certainly witnessed the commercial results of this “mining” in your local supermarket and on your own smartphone, tablet, laptop, or desktop. Academic researchers have been actively mining data from nearly every domain, as well, and important public health applications are rapidly emerging.

Perhaps you’ve noticed that supermarkets sometimes display unrelated items on the shelves next to each other. Wine and cheese are a common example, and of course they have an obvious gastronomic association. But what about beer and baby diapers—another pairing you might spot at Walmart? The hidden key to this seemingly illogical twosome is data mining, a prime research topic for Dr. Hui Yang, an associate professor of computer science at San Francisco State University. Computer scientists working for large chain retail stores can employ data mining techniques to recognize and extract meaningful patterns from customers’ purchase behavior.

This data, in turn, helps store managers to increase their sales by selecting and pairing sometimes-unlikely sets of products and shelving them together to promote sale through association or convenience. The results of data mining also help customers to make decisions more quickly and easily while shopping. During a recent interview at her ninth-floor office in Thornton Hall, Yang was able to further illuminate the “beer and baby diaper” example. Walmart’s computer programmers applied data mining techniques to recognize their customers’ purchasing patterns, she says. To do so, they wrote a computer program that assigns each customer a virtual “shopping basket” capable of tracking all of the items the customer purchases during each store visit. Computer researchers then analyzed tens of thousands of such baskets and millions of individual items. Their analysis led to the surprising discovery of a positive association between purchasing baby diapers and buying beer. Further analysis revealed that it is mostly husbands who purchase those items in the same “basket.” A plausible explanation for this purchase pattern, says Yang, is that husbands sometimes babysit, watch TV sports, and drink beer at the same time. This data-mined insight then informed the decision of marketing managers at Walmart to shelve the items side by side to increase the sales of each.

Yang offers a similar example from Amazon.com, again based on data customers’ co-purchasing histories. “Have you ever noticed the following feature on Amazon.com: “Customers Who Bought

39
Yang likes to think of data mining as “finding the treasure that is hidden in information/knowledge/data.” “This information,” she continues, is newly generated: it “is not something that is previously known.” It also requires finding “meaningful patterns among the newly discovered information.” Computer programmers use various names for the process: knowledge mining from data, knowledge mining, knowledge extraction, data (pattern) analysis, and knowledge discovery from data (KDD). But it’s all the same general pattern seeking with significant implications.

Expanding Applications
Data has exploded in recent decades, says Yang, from kilobytes to megabytes, gigabytes, terabytes, and petabytes. This proliferation has made data collection and data availability more complicated and unmanageable; this, in turn, has encouraged computer scientists to develop automated data collection tools and database systems. Data mining has become particularly useful to fields with large data archives such as business (for use in e-commerce and transactions); science and medicine (for remote sensing, bioinformatics, scientific simulation); education and academia (for online teaching, learning, and grading); consumer culture (for news coverage, YouTube, and digital photography), and in a host of other applications, including web site personalization, customer care, and fraud detection.

Rather than relying on their own instincts, data “crunchers” of all types now have tools with which to extract valuable knowledge hidden in large data archives and then use it to inform their business strategies, research conclusions, even arrests for illegal activity. They can get glimpses of future behavior, as well, that will allow them to make forecasts and plan new directions to follow.

In her own research, Hui Yang is mining published biomedical research articles in the field of nutritional genomics. Specifically, she is seeking to learn the relationships between food, diseases and genes. In other work, she is visualizing and analyzing microarray data on various genes as well as information on protein-folding. Recently, for example, she used biomedical text mining to explore the relationship between milk drinking and ulcers in scholarly journal articles. A patient might feel stomach pain stemming from what medical tests reveal to be an ulcer and then he or she receives a doctor’s prescription for medication. But, Yang explains, this patient may want to change to a healthier diet and as a result, tries an Internet search for foods that prevent ulcer, worsen ulcers, or cure ulcers. Almost certainly, says Yang, the patient would have difficulties analyzing the Internet search results, which are usually voluminous, she says, “and may lead to confusion. Moreover, there are always conflicts between data. Say you have found on the Internet that drinking some type of milk can prevent ulcer, but now you want to know the
best type of milk to drink, and whether this is the prevalent expert opinion or advice from a minority of physicians and/or nutritionists.

In the past two years or so, Yang and her graduate students have designed software for a web-based prototype that will enable users to post phrases about subjects for which they seek information, recommendations or answers (whether to drink milk, for example, or skip it, and what types are best and worst.) The system then searches English-language journal articles stored in online data bases such as PUBMED. The result, says Yang, is “a graphical user interface that visualizes the automatically extracted food-centric networks,” and allows users to interact by refining their questions and exploring in new directions. With Yang’s prototype, a user could, for example, receive a summary of recommendations about ulcer and milk drinking. “It will tell you the percentage of people saying to drink 2 percent milk, or skim milk, or whole milk,” Yang explains. For instance, half of the comments might say drinking non-fat or 2 percent milk could alleviate ulcer’s pain. One-fifth of the comments might discourage all milk drinking to avoid worsening the ulcer, and about one-third might recommend drinking only skim milk for ulcers. Yang concludes that using biomedical text mining to analyze literature articles to extract, rate, and summarize relationships between food (in this case milk) and disease (in this case ulcer) can provide useful information to both patients and to health professionals who need to know overall health effects in order to establish healthy life habits.

Yang’s biomedical text mining project offers research opportunities for graduate students in the computer science department at SF State. They program and analyze the journal articles using JAVA programming language (because of its portability). Yang also collaborates with scientists from other universities such as Cal Poly at Pomona who bring nutritional and genomics backgrounds to the project. SF State faculty in other disciplines such as biology participate, as well, through the Center for Computing for Life Science (CCLS) in Thornton Hall. This large, well-lighted space is filled with computers, white boards, and comfortable furniture that encourage collaboration, brainstorming, and long work sessions.

Yang and her extended team, plans to expand the search for nutritional information beyond scientific and medical journals to blogs and other databases, and to include information published in languages other than English. They also intend to improve and evaluate their web-based prototype system and to open it to public usage. Her ultimate goal, she says, is for people to “be able to access it anywhere through the Internet to gain knowledge regarding the health benefits of a given food.”

A schematic description of the web-based text mining system developed by Dr. Yang’s research group. The system extracts health-related knowledge from public data repositories such as PubMed. It employs a host of techniques including machine learning, natural language processing (NLP), and text mining. It consists of five main functional modules ranging from relevant article retrieval to relationship integration and visualization. The system also is also equipped with a web-based graphic user interface (GUI) allowing end-users to interact with the extracted knowledge.

Graphic courtesy of Hui Yang

A BRIEF HISTORY OF DATA MINING

Data mining is a relatively young field but it rests on half a century of technological development. Collecting and digitizing data began in the early 1950s. The 1960s saw the creation of data bases to improve information collection and management in a range of formats. As the mass of data grew, it became vital to better organize and run data bases so computer researchers developed Data Base Management System (DBMS) such as FileMaker or Microsoft Access. Later in the 1960s, IBM introduced Information Management System (IMS) as a joint hierarchical database and information management system with an emphasis on transaction processing. Data collection continued to mature in the 1970s and 80s, resulting in the evolution of the World Wide Web and multimedia databases in 1990s. Parallel techniques emerged for maintaining cohesive backup and storage, and this allowed data mining and data warehousing to emerge. Since 2000, data mining applications have mushroomed and become more and more integrated with Internet technology and global information systems.