PHENOMENAL DATA MINING: FROM DATA TO PHENOMENA

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

2001 Oct 20, 11:39 a.m.

Abstract

Phenomenal data mining finds relations between the data and the *phenomena* that give rise to data rather than just relations among the data.

For example, suppose supermarket cash register data does not identify cash customers. Nevertheless, there really are customers, and these customers are characterized by sex, age, ethnicity, tastes, income distribution, and sensitivity to price changes. A data mining program might be able to identify which baskets of purchases are likely to have been made by the same customers. In this example, the receipts are the data, and the customers are phenomena not directly represented in the data. Once the "baskets" of purchases are grouped by customer, the way is open to infer further phenomena about the customers, e.g. their sex, age, etc.

This article concerns what can be inferred by programs about phenomena from data and what facts are relevant to doing this.[1] We work mainly with the supermarket example, but the idea is general.

[1] In a sense, all data mining is phenomenal; it's just that the phenomenal part is usually done by hand. We want the computer to do the phenomenal part also.

In order to infer phenomena from data, facts about their relations must be supplied. Sometimes these facts can be implicit in the programs that look for the phenomena, but more generality is achieved if the facts are represented as sentences of logic in a *knowledge base* used by the programs.

The result of phenomenal data-mining might include an extended database with additional fields on existing relations and new relations. Thus the relations describing supermarket baskets might be extended with a customer field, and new relations about customers and their properties might be introduced.

1 Introduction

Science and common sense both tell us that the facts about the world are not directly observable but can be inferred from observations about the effects of actions. What people infer about the world is not just relations among observations but relations among entities that are much more stable than observations. For example, 3-dimensional objects are more stable than the image on a person's retina, the information directly obtained from feeling an object or on an image scanned into a computer. [2] Likewise the fact that a customer has children is more stable than the fact that a particular basket includes Roll-ups. The fact that a customer has diabetes is more stable than a particular pattern of food purchases that may allow inferring that he has diabetes. The phenomenal facts, partly because they are more stable than observations, are more predictive of future behavior than simple obsrvational facts.

The extreme positivist philosophical view that science concerns relations among observations still influences the design of learning programs, and that's what *data miners* are. However, science never worked that way, neither do babies and neither should data mining programs. All obtain and use representations of the objects and use observations only as a means to that end.

Data mining involves computer programs that infer relations among different kinds of data in large databases. The goal has been to infer useful

[2] Even very young babies have a lot of innate knowledge of the world. My article **The Well-Designed Child**[3] concerns what innate knowledge children probably do have about the world and what knowledge robots should be given. Elizabeth Spelke, [Spe94], investigates innate knowledge in babies experimentally.

relations that might not have been noticed or at least could not have been confirmed among this data. We use the standard example of a supermarket chain with a database of all the cash register receipts for some long time period—many gigabytes of data. The company wants to *mine* this database for information useful for improving its business.

Data-mining can be made to do more than just find relations among data. Data amounts to observations of the world, and it is possible to infer relations among entities in the world from the data. Such relations are likely to be as useful to know about as are relations among the entities directly represented in the data. In the supermarket chain example, there are people, groups of people, their homes with pantries, refrigerators and freezers and facts about what they cook and what they eat. It should even be possible to infer the existence of entities in the world, such as previously unidentified groups of people with distinct eating and purchasing habits. Another example is to identify bellwether groups; what they buy today, many more will buy tomorrow.

Moreover, the information will usually admit a more compact description in terms of the underlying phenomena than in terms of the data.

Although all common sense level knowledge of the world is potentially relevant to data mining, formalizing common sense has proved to be a difficult AI problem, and progress has been slow. Nevertheless, we can expect that certain phenomena will be related to the information in databases in a straightforward enough way so that information about them can be found by data miners.

2 Phenomena in the World

What phenomena in the world should a data mining program have built into it, be told or be able to discover for itself?

At first, knowledge of the general phenomena will be built into the data miners (data mining programs), and the programs will infer specific values. Later data miners should use the information expressed in a logical form. This will permit them to use databases of common sense facts about the world. Very ambitious data mining projects might hope to make programs that will come up with entirely new phenomena.

Here are some phenomena and facts relevant to the supermarket domain together with logical expressions for some of these facts. We give just two example formulas, and these are not part of a worked out scheme for constructing a knowledge base.

people There are the shoppers themselves and also family members. The data may not identify them directly, but learning about them is the point of data mining.

$$Shopper(x) \rightarrow Family(x) \subset People. \tag{1}$$

ownership and purchases People buy things and then own them and keep them somewhere. Maybe the facts about where people keep things are not relevant for most data mining. The distinction between durable goods and consumables is important.

$$\begin{array}{l} Durable(x) \wedge Has(person, x, s) \rightarrow Has(person, x, Result(Uses(person, x))) \\ \neg Durable(x) \wedge Has(person, x, s) \rightarrow \neg Has(person, x, Result(Uses(person, x))) \end{array} \tag{2}$$

possessions Freezers, refrigerators, cars and microwave ovens are items that some customers will have and others won't. Having them affects behavior.

events The observed events are purchases in the stores for which we have databases.

Unobserved are the trips to the store and the cooking and eating and the inspections of the larder. Maybe these can usefully be discriminated, but maybe they should be lumped into consumption. Other unobserved events include purchases from the competitors. When a person purchases a freezer, his status changes to that of a freezer owner and that fact will persist. The event of acquiring a freezer is more common than that of giving up the possession of a freezer.

preferences People have preferences among states of affairs—or more specifically among objects.

distributions of properties over people The customers have age, sex, income and ethnic distributions.

customers appear and disappear There are causes for the appearance and disappearance of customers, and supermarket chains will be interested in finding them out. These include moving in or out of the area, change in family circumstances, advertising campaigns by the chain or its competitors, changes in the store or its hours of operation, satisfaction or dissatisfaction with goods, prices or service.

The present state of AI is not up to formulating a full common sense database, but full common sense knowledge is not necessary. We can expect to do a lot with very limited knowledge. A sophisticated data mining system might be able to use the following facts in its formulation of hypotheses. An ambitious logic-based system might use logical expressions of the facts. Less ambitiously, programmers would use them in designing data mining systems.

1. People persist in time. People want objects. People consume objects and want more. Some objects are permanent on the relevant time scale.

2. Objects are created, appear in stores, sold to customers (people) who use them up and need more.

3. There are kinds of people and kinds of objects.

4. People have attributes, and these attributes change, although some are permanent.

5. People buy objects with money. This uses up money and people do not buy at a rate much higher than they get more money.

6. There is an is-a hierarchy of items and and an is-a hierarchy of people. We suppose these are spelled out in some literature.

7. There is an is-a hierarchy of food.

8. Although it is tempting to organize facts into is-a hierarchies, this is not always possible or appropriate. More complicated predicates and functions and logical assertions are sometimes needed to express the facts.

9. People are associated into families. Purchases are made for a family.

10. When food items are purchased, some go into pantries, some into refrigerators, some into freezers and some are eaten right away. When a food object is eaten it is removed from where it was stored.

11. There are bounds on the rate at which people eat. What they don't get from one store they get from another.

12. A person has an age which increases with time. Very young people are children.

13. There are lots of people an lots of stores. The data miner will have information about only some of them.

14. Customers who buy substantial quantities of frozen or freezable goods have freezers.

15. Owners of microwave ovens can be identified.

16. Consistent purchase of the most expensive items indicates prosperity. It can be asked whether consistent purchase of expensive items is all the data miner wants to know anyway. I don't know about that.

17. Everybody eats, so food not bought at one store is bought at another.

18. Suppose a customer comes rarely and always buys frozen spinach in bags and a few other items. Inference: the store where he buys most of his food doesn't sell frozen spinach in bags.

The point is that all the above are a priori facts that may be used to infer phenomena. We suppose that only some phenomena need be taken into account. For this phenomenal mining we ignore birth and death, physical motion, and shape. Mass is taken into account only in connection with quantities purchased and rates of consumption.

It is clear that a very large number of facts are relevant to getting information out of databases of customer purchases. These include general facts of common sense and specific facts about consumer properties, consumer goods and consumer behavior. I see no alternative to a big project like CyC [LG90] for them into a knowledge base by hand. However even a small knowledge base may be useful and adequate for experiments.

3 Grouping supermarket purchases by customer

We propose programs to determine from the cash register receipts which baskets were purchased by the same customer. The putative customers can then be given identifiers. Programs can infer more facts about customer characteristics and behavior with facts about purchases of an identified customer over time than could be inferred from mere statistics about the baskets themselves.[4]

This example of *phenomenal data mining* is straightforward in that it is reasonably clear what a successful result would be and how it might be used. We hope to make it plausible that enough information is present in the data to usefully distinguish customers. However, experiment is needed to verify that feasible algorithms exist.

Demographic information about customers is known to be useful, e.g. their ages, occupations, sexes and incomes. When this information is supplied, e.g. in mail order situations where credit is granted, it is extensively used. However, in our supermarket chain example, that information is not in the database of transactions. Let us consider inferring it; it might then be used in any of the presently conventional ways.

There are several approaches to associating baskets purchased by the same customer.

3.1 Minimizing anomaly in assignments of baskets to customers

One approach involves minimizing *total anomaly* in the assignment of baskets to customers.

Definition: A *partial assignment* α groups some of the baskets of purchases according to whether they were purchased by the same customer. Each group also includes an identifier c for the customer and a *classification*

[4] It has been suggested that grouping baskets by customer is an example of *clustering* as treated in learning theory. This is incorrect, although there are some similarities. Consider two large identical baskets purchased ten minutes apart. Clustering would assign them to the same category, but these baskets would almost certainly have been purchased by different customers. Identical baskets purchased far enough apart would have an increased probability of having been purchased by the same customer, but it wouldn't be certain. Still, the literature on clustering might tell us something useful for the present problem.

$class(c)$ of the customer. The set of baskets associated with the putative customer c will be denoted by $baskets(c)$.

Definition: A *complete assignment* groups all of the purchase baskets.

If there are N baskets in the database, there are something like 2^{2^N} complete assignments—less because the customers may be permuted.

Definition: Associated with each assignment will be a numerical *total anomaly* measuring how anomalous the assignment is. The program's goal is to find an assignment (or maybe many assignments) that minimize the total anomaly.

The total anomaly $anom(\alpha)$ of an assignment α is the sum of two main terms,

$$anom(\alpha) = anom1(\alpha) + anom2(\alpha). \tag{3}$$

$anom1(\alpha)$ is itself a sum

$$anom1(\alpha) = \sum_c anom11(c), \tag{4}$$

where the variable c ranges over the set of customers to which the baskets are assigned. $anom2(\alpha)$ concerns global properties of the set of assignments.

Definition: Associated with an assignment α and a customer c is a *characterization* $char(c, \alpha)$ of the putative customer. The characterization may include qualitative characeristics like sex or owning a freezer, quantitative characeristics like age or income group and other *customer characteristics* like a certain purchase signature. The anomaly $anom11(c)$ associated with a customer c depends on the characterization $char(c, \alpha)$. Thus buying chewing tobacco or baby food is more anomalous for some customers than others. A program that generates assignments will generate characterizations as it groups the baskets by customer. The characterization itself will contribute to the anomaly if it is an unusual characterization.

Definition: A *signature* is a set of choices among alternate brands or sizes of certain commodities. The commodities most useful for signatures are those for which variety is not normally considered desirable. While a person may want variety in food he is unlikely to want variety per se in dishwashing soap, toilet paper or size of dog food. Signatures are included in the characterization of a customer.

The part of the anomaly $anom11(c)$ associated with the putative customer c is computed relative to the characterization. Thus if c is characterized as single young female, a purchase of chewing tobacco should have a higher anomaly score than for a male.

One way of looking at minimizing anomaly of assignments is that we want to explain as much of the purchasing behavior as possible by allowable characterizations of the customers.

We regard the notions of minimizing anomaly in the space of assignments as a guiding theoretical idea. Programs may find complete assignments, but they are unlikely to do it by comparing a large number of alternative complete assignments. Instead they are likely to do hill climbing in the space of partial assignments.

Here are some kinds of terms that may be associated with the customer part of the anomaly function.

1. A measure of the temporal irregularity of the customer's purchases. Perishable, non-freezable items like milk need to be purchased at a fairly regular rate. If baby food is purchased, it also is consumed at a regular rate, although it can be stored. Some customers will be very irregular, but an assignment shouldn't make most of them irregular.

2. A measure of the extent to which the grouped baskets do not fit the characterization $char(c, \alpha)$.

3. Signatures involving a large variation in brands of certain items should contribute to the anomaly.

4. A lot of variation in a putative customer's purchase quantity of a frequently bought item. This suggests that the same person didn't buy all those baskets.

5. A customer buys food, stores it for a while and eats it. Thus the contents his larder is a function of time. The database tells about the purchasing but not directly about the eating or the state of the larder. We can attribute a *larder function* of time to a customer as part of the ascription and use some measure of its irregularity as a component of the anomaly.

Here are some ideas about programs for finding assignments.

1. We hill climb in the space of partial assignments. For example, moving a purchase from one customer to another may reduce the anomaly of both customers' ascribed larder functions.

2. We might proceed chronologically, assigning each basket to either a previously postulated customer or to a new one.

3. At first new customers would predominate. However, when the number of postulated customers begins to get too large for the number of baskets, the program would try to reduce the number by combining baskets.

How can it be inferred that several cash purchases involved the same customer? We only need to be correct often enough so that the statistics come out right. Each customer has his own pattern of purchases. Here are some considerations.

1. The *signature* $sig(c, \alpha)$ is a purchase pattern unique to the customer c. Consider items where variety is not normally desired, e.g. dishwasher soap. There are several brands, but a customer will normally stick with one for quite a long time. If there are 5 brands and 50 such kinds of items, there are enough possible signatures to distinguish far more customers than a store or even a chain is likely to have. Of course, a customer is unlikely to purchase a complete signature package each time he goes to the store, so *partial signatures* will have to be used.

2. The ingredients for particular recipes are sometimes diagnostic, especially when the recipe is unique to the customer or is a standard recipe varied in a unique way.

3. An important intermediate variable for a customer is the state of his larder at a given time. He likes to have certain items in stock in his refrigerator or freezer.

4. The customer makes choices in a certain pattern, e.g. buys creamy rather than chunky peanut butter. Which choices are made is more indicative than whether peanut butter is bought at all on a particular occasion, since the customer may not have run out yet.

5. Suppose a store has 10,000 items and has 12,000 customers. Suppose purchases average 20 items. My information theory intuition suggests that there is enough information to identify the customers over some 20 shopping trips. The information theory numbers can be analyzed, but experiment is still required to determine feasibility.

6. Sometimes it will be impossible to assign a basket to a customer. As an extreme example, suppose that withing ten minutes two customers each buy a six pack of the same brand of beer and nothing else. Which one made which purchase will be impossible to tell, but it won't matter which purchase is assigned to which customer.

4 The Customer as a Stochastic Process

The methods discussed in section 3 group purchases by customer. However, the specific purchases made by the customer are of interest only in so far as they enable prediction of his future behavior and how he might respond to things the store might do, e.g. advertisements, sales, changes in products offered, changes in prices.

In general, we might regard the customer as a stochastic process, i.e. what he will buy (and whether he will come to the store at all), depends probabilistically on the state of his larder, and the actions of the store.

A regular customer may visit the store once per week for 5 years, i.e. make 250 visits. Some may make as many as 1,000 visits. Nevertheless, there often won't be enough information to make a very sophisticated model of a customer. Therefore, simplified models are worth considering.

The simplest model is that customer c has probability $p(c,i)$ of buying item i. The matrix $||p(c,i)||$ is likely to be approximable by a matrix of much lower rank, i.e. the customers form a space of lower dimension. If this is true, customers can be approximately characterized by a much smaller number of parameters than are needed for a complete probability distribution. This in turn means that accurate information about the customers can be obtained with smaller samples that would otherwise be required. If the assumption of independence of the members of the signature is valid, it still takes quite a lot of information to characterize the customer.

The next more elaborate model might take into account the state of the customer's larder. He won't buy more of certain items until he has consumed what he previously bought. If we regard the customer's state as given by the contents of his larder, we can regard his purchases as determined by a Markov process.

The model might be further elaborated to take into account his probable response to sales, etc. Economists would be tempted to try to ascribe a demand curve, most likely just two numbers—the demand at a base price and an elasticity.

We will not pursue these elaborations further in this article, but it seems likely that the most useful information to supermarket companies doing data mining will involve the probabilities of response of different kinds of customers to different stimuli.

5 Mail Order Bookstores

Consider a store selling books by mail. The customers are identified, so we don't have that problem.

However, we can suppose that many characteristics of customers are not identified such as age, income, occupation. Literary tastes can be identified in so far as they correspond to a tendency to buy books in predefined classifications. However, the data may allow inferring new classifications with respect to which the customers behave more consistently than with respect to the traditional classifications.

Some classes of customers are leaders in that their preferences today can be used to predict the market for a book later. Identifying such customers and classes of customers may be useful.

The above phenomena—age, etc.—are not in the data *per se*, but they are rather close to it. Their identification should not be as ambitious a project as identifying the customers of a supermarket. Almost all of the computations will involve the individual customers. The leadership phenomenon involves more but still has a rather simple character.

6 Proposed Experiments

Grouping supermarket purchases by customer as proposed in Section 3 can be tested with the aid of a supermarket database that does contain customer identification. We discard the customer identification, run our grouping algorithm and compare the results with the genuine customer data.

My present opinion is that grouping baskets by customers is likely to be a difficult but feasible task. As will be seen, it will involve taking advantage special features of the behavior of supermarket customers. In this respect, it may resemble cryptanalysis which often takes advantage of special features of the behavior of senders of messages. Moreover, the results cannot be perfect in terms of identifying the purchasers, but the uncertainties about who bought what may not affect the interesting statistics of customer behavior.

Here are some ideas about how to proceed.

1. It may be best to start the experiments with a relatively small store. That way there will be fewer assignments to try and fewer similar signatures.

2. Very likely we should start with a date in the middle of the operation of the system and try to extend identifications both forward and backward in time.

3. At any time in the computation, there will be a certain collection of putative customers and a set of possible assignments of some of the baskets to customers. Maybe the computational resources will be adequate to deal with hundreds to thousands of possible assignments. Each of these assignments will have an anomaly computed on the basis of what has been assigned so far.

4. Since many people shop on a weekly basis, it may be worthwhile to try to find some putative customers who buy on a particular day of the week.

5. It may be possible to find some signatures for some customers that are repeated every week. For example, a shopper may buy both whole milk and skim milk every time, because of the needs of different family members.

6. The algorithm may grow assignments forward and backward in time. As it goes it will eliminate certain assignments.

7. When it cannot decide among the assignments over some lengthy period, say two months, it should probably just pick one in order to keep down the number of open choices.

8. Perhaps there will be a compact way of keeping certain choices open in order to use long term aspects of the signature.

7 The Logic of Phenomenal Data Mining

The ways in which mathematical logic has been used in database theory and database systems are likely to require extension for phenomenal data mining..

Database theory and database system commonly use mathematical logic to represent facts. However, subsets of logic are adequate for most present database systems. For example, the databases can often be considered as collections of ground literals, i.e. predicate symbols applied to constants. More general sentences are used as rules and given a special status.

One example that immediately arises in the supermarket problem is the fact that the customers who bought particular baskets are unknown, and it is not known *a priori* whether two given baskets were bought by the same customer. In Prolog and similar systems, the *unique names hypothesis*, i.e. that distinct constant expressions represent distinct objects, is usually built into the system.

Consider

$$buyer(b1) = buyer(b2),$$

which asserts that baskets $b1$ and $b2$ were purchased by the same customer. Unlike the common situation in database systems, the truth of this assertion is not in the database. Neither is a unique identifier for *b1* available.

The set of customers is unknown, although it is known to exist. Facts about it may be known or conjectured.

7.1 Ontology

We follow Quine in taking the *ontology* of a system to be given by the collection of domains over which variables range. In the supermarket example, we may have

products These can be represented by their product codes.

purchased items The particular instances of items purchased as part of a particular basket.

baskets The collection of items purchased on a particular occasion.

customers The set of customers is unknown but is known to exist .

8 Remarks.

1. Suppose a customer of type i has a probability P_{ij} of including item j in a basket. We can infer an approximate number of types by looking at the approximate rank of the matrix P_{ij}.

2. Classifying customers into discrete types may not give as good results as a more complex model that take into account the age of the customer as a continuous variable.

3. A linear relation between phenomena and observations is the simplest case, and such relations can probably discovered by methods akin to factor analysis.

4. We could infer that there were two subpopulations if we didn't already know about sex.

5. We might infer from data from our stores in India, that there was a substantial part of the population that didn't purchase meat products. We can tell this from a situation in which everyone buys meat but less, because certain other purchase patterns are associated with not buying meat.

6. Tire mounting services are purchased in connection with the purchase of tires. The phenomenon is that tires are useless unless mounted. Does knowing this fact give more than just the correlation?

7. Suppose a new item, e.g. a hula hoop, is increasing its sales rapidly, and 5 percent of the customers have bought it. Suppose, however, that the customers that buy it rarely buy another, and these customers are only those with young girls in the family, and those customers have almost all bought one. Under these hypotheses, which identifying customers might verify, it is reasonable to conclude that the fad for hula hoops has reached its peak, and that if a lot more are ordered, the store is likely to be stuck with them.

8. Suppose we have the baskets grouped by customer—either because the data was given or because we have inferred it as described above. Can we determine how far the customers live from the store? The information might be useful in anticipating how much business might be lost to

a newly opened competitor. No immediate idea occurred to me when I thought of the question. However, it is rash to conclude that it can't be done. Someone cleverer than I, or who knows more about customers of supermarkets, might figure a way. One just shouldn't jump to negative conclusions.

9. Grouping by customer might permit observing that no-one who buys item 531 ever buys anything from that store again. Such a fact would not show up as a direct correlation in the data unless item 531 were bought in quantities that significantly affected sales of some other items.

10. If a customer buys a certain product but doesn't buy a necessary complementary product, we can infer that he buys the complementary product from someone else.

The only experimental work with phenomenal data mining is reported in [LT98].

9 Acknowledgments

I am indebted to Rakesh Agrawal of the IBM Almaden Research Laboratory for introducing me to data mining in conversation and through his papers. I also had useful discussions with Ted Selker, Joe Halpern, Jeff Ullman, Nimrod Megiddo, Rajeev Motwani, Brad Efron, Christos Papadimitriou, Ron Fagin, Tom Costello, and Gregory Tseytin.

This work was partly supported by ARPA (ONR) grant N00014-94-1-0775 and partly by IBM while I was visiting faculty during the summer of 1996 and since.

References

[LG90] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project.* Addison-Wesley, 1990.

[LT98] Donal Lyons and Gregory S. Tseytin. Phenomenal data mining and link analysis. In David Jensen and Cochairs Henry Goldberg, editors, *Artificial Intelligence and Link Analysis, 1998 Fall Symposium*, pages 68–75. AAAI, AAAI Press, 1998.

[Spe94] Elizabeth Spelke. Initial knowlege: six suggestions. *Cognition*, 50:431–445, 1994.