Interactive Data Mining at the Speed of Thought: A Position Paper

Vladimir Zelevinsky
Principal Research Scientist
Oracle
101 Main St, Cambridge, MA 02142
vladimir.zelevinsky@oracle.com

ABSTRACT
We describe a system that combines guided navigation with the computation of Pearson correlation coefficients to support the task of interactive data mining by creating dynamic previews of possible navigation states.

Author Keywords
Interactivity; faceted search; guided navigation; data mining; correlations.

ACM Classification Keywords
G.3 [**Probability and Statistics**] – *Correlation and regression analysis*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Relevance feedback, Selection process*.

General Terms
Algorithms; Theory; Measurement.

INTRODUCTION

At present, data mining is a process that requires careful model-building and deep initial insight into data; the subsequent steps include writing custom code and running complex computations for longer than expected; at the end of which, a result is produced that is either obvious or wrong. At best, data mining requires expert input and guidance, returning results that only experts can understand and analyze.

The purpose of this paper is to walk through one solution that utilizes existing – and commonplace! – technologies and century-old mathematics to arrive at a system that performs data mining at the speed of thought, while making it usable by (and useful to) any layperson.

GUIDED NAVIGATION AND INTERACTIVITY

The industry standard of guided navigation (otherwise known as faceted search [1]) is usually implemented via the widget (see Figure 1). The count in the parentheses is the number of documents one would obtain after narrowing down the current result set by selecting the corresponding refinement. This count is effectively a preview of the potential refined data set. This practice dominates the implementations of guided navigation: it is well known, well-understood, and its assumptions are too infrequently questioned.

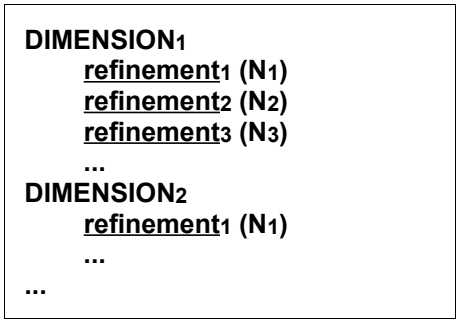

Figure 1. The standard guided navigation widget

Generally speaking, such a future preview does not have to be limited to the cardinality of the future refined set: it can be any analytic query performed on such a set and its derived attributes. For example, in some data exploration scenarios, it might be useful to display such a preview as SUM(revenue), or SUM(revenue) – SUM(expenses). In this case, the user can narrow down to those slices of data that have either a particular total revenue or a particular profit.

In such scenarios, more common in the world of business intelligence, the ultimate information retrieval need is not a particular record or a set of records, as tends to be the case in the e-commerce world, but an insight. To be fair, the customer's experience with an e-commerce site must include the process of arriving at one or more insights as well, whether these insights involve the discovery of an item the customer wants, a realization that a particular brand has high reliability ratings, or an understanding of certain trade-offs (for example, that higher user ratings tend to correlate with higher prices).

Such iterative navigation of the available data (by choosing one of the suggested refinement links while using the record count as a preview of the potential refined set) is, in its essence, an elementary case of data mining – with the advantage of interactivity, which helps the user follow the

Proceedings of HCIR 2012, Cambridge MA, Oct 4-5, 2012

information scent of a particular information retrieval need or an insight.

We'll describe a system that offers a powerful data-mining experience while following the interactive (and iterative) pattern of guided navigation, requiring no technical knowledge from the user, and being able to run the computations at the (almost) speed of thought.

DATA SET

For our research, we selected the famous Boston 1978 data set [2]: for each census tract in the metropolitan Boston area (506 total), 17 variables were recorded (see Table 1; the labels and their explanations are given as provided with the source data). The advantage of this set is that the data-mined conclusions tend to be easily verifiable via common sense. In our research, we ignored the information-free columns of tract number, as well as those of latitude and longitude.

TRACT	tract number
LON	approximate longitude
LAT	approximate latitude
CMEDV	median value of owner-occupied homes in $1000s
CRIM	per capita crime rate
ZN	portion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$, where Bk is the proportion of blacks
LSTAT	% lower economic status of the population

Table 1. Boston 1978 data columns

As an aside: writing this paper in 2012, we cannot help but express a certain measure of bewilderment at the variable B, defined as $1000(Bk - 0.63)^2$, where Bk is the proportion of black residents. This formula, while appearing inconspicuous at the first glance, takes on a different meaning when one realizes that in all of Boston the variable Bk is always *less* that 0.63, making the number in parentheses always negative. As a result, B *decreases* with the increasing values of Bk, implying a rather troubling political position.

INSIGHT FROM CORRELATIONS

What kind of an insight can a set such as this offer to a user (for example, a recent transplant to the Boston area who is looking to buy a house)? One way to compute interesting characteristics of such a set would be to calculate correlations between the numerical columns.

We use the standard Pearson correlation coefficient formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

This expression has two highly useful features. Since the bulk of computation involves summing two variances and one covariance over every row of data, such a computation can be easily parallelized. In addition, it is highly amenable to sampling.

In some scenarios (e.g., if the data set is that of manufacturing and sales), one might be interested in affecting a particular column that cannot be directly manipulated, like profit. Pearson correlation is capable of demonstrating which columns correlate (or anti-correlate) with the column of interest: for example, the number of vacation days, which can be directly influenced, is positively correlated with productivity, or the rate of machine X being introduced in the factories can be anti-correlated with the failure rate of the manufactured item. It is well known that correlation does not imply causality; however, such correlations can point to valuable insights (in the examples above, suggesting that managers increase the number of vacation days and utilize machine X).

What if the user is not interested in influencing a particular column but is looking for an arbitrary insight? The need to find answers to questions the user does not know how to ask tends to be well-addressed by guided navigation. In our case, this can be accomplished by computing $r(C_x, C_y)$ for every pair of columns of quantitative data.

DIS	**NOX**	–0.769
INDUS	**NOX**	+0.764
AGE	**DIS**	–0.748
CMEDV	**LSTAT**	–0.741

Table 2. Top correlations

We have performed this computation for the top correlations (see Table 2). The results do confirm that the air pollution (NOX) is anti-correlated to the distance (DIS) from the places of employment, which, in 1978, were likely to involve industrial production; that the prevalence of industry (INDUS) in towns results in more air pollution; that the older towns (AGE) are more centrally located; and that expensive houses (CMEDV) are rarely afforded by poor people (LSTAT). Slightly further down the list is the anti-correlation between house prices and pupil-teacher ratio: r(CMEDV, PTRATIO) = –0.506, suggesting that expensive houses (high CMEDV) do correlate with good schools (low PTRATIO). As evidence of victory for civil rights, the B column is not strongly correlated with any other column.

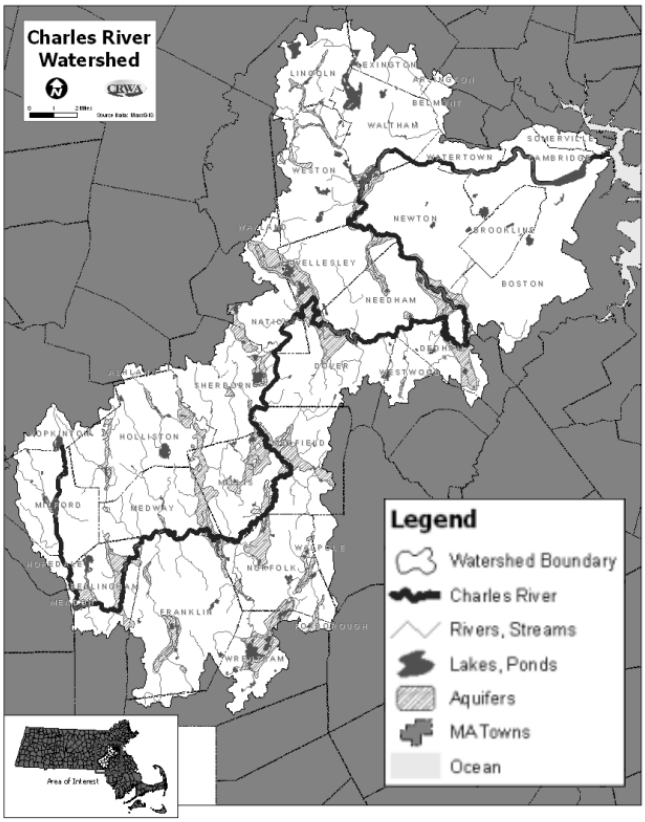

Figure 2. The course of Charles River (courtesy of Charles River Watershed Association)

One data column, however, is not like the others: this is the binary CHAS variable, which is set to one if the tract border the Charles River and zero otherwise. The correlations between this variable and any other are minor: all correlations r(Cx, CHAS) are less than 0.176 by absolute value. For the explanation of the low predictive power of this variable, one only needs to look at the somewhat unpredictable course of the Charles (see Figure 2).

INSIGHT FROM CORRELATIONS

If the Charles River variable fails to offer any insight into the nature of the data, it does suggest the most promising direction in applying correlation computations to interactive data mining.

So far, we have silently ignored one data column: namely, the name of the town. While one cannot compute a correlation between a numerical (quantitative) column and a text (nominative) column, there is a standard transformation technique that allows to apply the Pearson formula to such data.

The technique involves splitting each nominative column into as many columns as there are different values (see Table 3), creating, in effect, a set of binary variables that operates exactly like the CHAS column (is the given tract in Boston? Cambridge? Brookline?). The formula remains the same, reducing to the case of point-biserial correlations.

Nominative column	Boston column	Cambridge column	Brookline column
Boston	1	0	0
Boston	1	0	0
Boston	1	0	0
Cambridge	0	1	0
Cambridge	0	1	0
Brookline	0	0	1

Table 3. Converting a nominative column into a series of binary (quantitative) columns

This will cause a performance hit: now we have N columns instead of one. If the process is computing all pairwise correlations, the penalty is multiplicative: two columns with three different values in each will make the computations nine times slower. However, the impact is linear if we are computing the correlation with one fixed column of interest; in addition, the formula is greatly simplified in the case of binary values. Our previous points on parallelization and sampling apply here as well.

After running the complete pairwise computations, we arrive at a set of new correlations (see Table 4). Finally, there is a strong correlation with the variable B, for the

Boston Roxbury	**B**	–0.574
Cambridge	**PTRATIO**	–0.436
Cambridge	**NOX**	+0.417
Brookline	**PTRATIO**	–0.393
Boston Charlestown	**LSTAT**	+0.326

Table 4. Top correlations for town names

largely black neighborhood of Roxbury: $r = -0.574$ (the sign is negative since B decreases with the increase of black population). Other correlations offer immediately useful information to our potential user who is looking where in Boston to buy a house: Cambridge has high air pollution but (at least, in 1978) good schools; Brookline also has good schools; and Charlestown, as has been pointed out in [3], has a high crime rate.

When we have this information, all that is left to do is to format it in a somewhat different, yet easily recognizable, manner (see Figure 3, and compare with Figure 1).

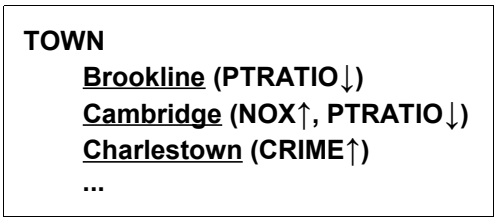

Figure 3. Guided Analytics widget

There are certainly other ways to present this information to users: refinements can be grouped according to their main (anti-)correlations, as opposed to the industry-standard grouping by their taxonomic categories; the refinements can be sorted in the order of decreasing correlation; or the user can be simply alerted to the top correlating pairs.

We have achieved what we set to obtain: a system that automatically generates a preview of each possible refined state, deriving this preview entirely from the data's most salient features, requiring no human input, and performing only simple computations. It is essential to keep in mind that we are working in the framework of guided navigation: the computations that were performed above on the entire data set will be re-computed on the basis of the narrowed subset after a refinement is selected. This way, if the user decides to limit selection to towns with good schools (Brookline and Cambridge in 1978), the PTRATIO variable will correlate less strongly (being low for both towns), and another variable is likely to be selected as a highly-correlating feature for the subsequent refinement preview.

The usefulness of interactive data mining extends beyond business analytics applications: the same approach can certainly apply to e-commerce, where it can distill the list of, e.g., information-free brand names into relevant facts (for example, brand X is strongly correlated with cheap products, while brand Y has high user ratings).

DIRECTIONS FOR FUTURE WORK

There is a considerable number of other correlation metrics (Pearson, in particular, performs best on normally distributed data), not to mention more complicated measurements that involve more than two columns. We expect a lot of potential in utilizing other signal-extraction computations in this framework.

The computational requirements of such a system are considerable. At present, there remains a wealth of challenges to overcome before we can truly achieve the "at the speed of thought" query-response times.

REFERENCES

1. Tunkelang, D. 2009. Faceted Search: Synthesis Lectures on Information Concepts, Retrieval, and Services. DOI: 10.2200/S00190ED1V01Y200904ICR005

2. Harrison, D. and Rubinfeld, D.L. 1978. Hedonic prices and the demand for clean air, *J. Environ. Economics & Management*, vol. 5, 81-102. Retrieved from: http://archive.ics.uci.edu/ml/datasets/Housing

3. http://www.imdb.com/title/tt0840361/