ISSN: 0828-3494
ISBN: 0-7731-0531-x (print)
ISBN: 0-7731-0532-8 (on-line)

An Application of Peculiar Record Detection to a City Works Database

Ken Konkel, Mahesh Shrestha,
Howard J. Hamilton, and Yiyu Yao
Technical Report CS-2005-04
July, 2005

Department of Computer Science
University of Regina
Regina, Saskatchewan
CANADA S4S 0A2

An Application of Peculiar Record Detection to a City Works Database

Ken Konkel, Mahesh Shrestha, Howard J. Hamilton, and Yiyu Yao
Department of Computer Science
University of Regina
Regina, Saskatchewan
CANADA S4S 0A2
{shresthm, hamilton, yyao, konkel1k} @ cs.uregina.ca

Abstract
Peculiar data are objects that are anomalies within a data set. These anomalies are relatively few in number and hold some interesting qualities that make them stand out. In this paper, we attempt to determine the effectiveness of mining for peculiar data on the Engineering and Works Department database from the city of Regina.

1. Introduction

Peculiar data are objects that are relatively few in number and are significantly different from the rest of the objects [9, 10, 11]. Peculiar data are interesting to the user because the user might have been completely unaware of it. Such data may provide novel information. Our purpose is to determine the effectiveness of mining for peculiar data on the Engineering and Works Department database from the city of Regina.

The remainder of this paper is organized as follows. Section 2 gives background information on previous methods for detecting peculiar data and defines four peculiarity measures. Section 3 outlines our adaptations to the PDD Framework [8]. Section 4 explains our experiments and details our results. Section 5 presents our conclusions.

2. Background

As described in [8], several methods for finding peculiar data have been developed and studied. We review three of these methods here. The data from which the peculiar objects are to be detected is in the form of a table. Let $R = R_1 ... R_n$ be the set of records in the data, and $A = A_1 ... A_n$ be the set of attributes in the data. Let x_{ij} be the value of record R_i for attribute Aj. The objective of the existing peculiar data detection methods is to detect peculiar records from the set R. Let $Distance(R_i, R_j)$ be a distance function that gives the distance between $R_i = (x_{i1} ... x_{in})$ and $R_j = (x_{j1} ... x_{jn})$. Let $Cardinality(S)$ be the number of elements in set S.

Zhong's method [9, 10, 11] calculates the peculiarity factor of each record in R based on the sum of the distance between R and every other record in the data set. We will refer to this as the *cumulative distance (CD)* of a record, defined as

$CD(R_i) = \sum_{j=1}^{m} Distance(R_i, R_j)^p$ where p is a parameter denoting the importance of the

distance, which can be adjusted by the user. By default, $p = 0.5$. The threshold value (*TV*) is a linear combination of the standard deviation and the mean of the peculiarity factors: $TV = \mu + (\alpha * \sigma)$ where μ and σ are the mean and standard deviation of the peculiarity factors, and α is the threshold adjustment factor that can control the number of peculiar data. All records having peculiarity factors greater than the threshold value are classified as peculiar records.

Knoor's method [4, 5, 6] detects peculiar data on the basis of two properties of a record: the magnitude of the distance between the record and the rest of the records and the portion of the data that lie at more than some specified distance. The peculiar data are called *distance based (DB)* outliers. A record R_i is a $DB(p,d)$ outlier if at least *p* of the records in *R* lie greater than distance *d* from R_i. A *radius-neighborhood* of R_i, $RN_d(R_i)$, contains the set of records $R_j \in R$ that are within distance *d* of R_i.

$$RN_d(R_i) = \{R_j \in R \mid Distance(R_i, R_j) \le d\}.$$

Breunig's method [1, 2, 3] calculates the *local outlying factor (LOF)* for each record by taking the ratio of the average density of the neighborhoods of the neighbors of the record over the density of the neighborhood of the record. In Breunig's terminology, the peculiar data are called *distance based local (DBL)* outliers. If the density of the neighborhood of a record is lower than that of its neighbors, its LOF will be higher and it will be called a DBL outlier. LOF is calculated as follows [2]. The *k cardinality-neighborhood* of record R_i, $CN_k(R_i)$, contains the k nearest neighbors of R_i. The *k-distance* of R_i, $D_k(R_i)$, is the distance between R_i and its k^{th}*-nearest neighbor*. The *local reachability density* of R_i, $LRD_k(R_i)$, is

$$LRD_k(R_i) = 1/\left(\frac{\sum_{R_j \in CN_k(R_i)} \max\{D_k(R_j), Distance(R_i, R_j)\}}{Cardinality(CN_k(R_i))} \right)$$

The *local outlying factor* of R_i is

$$LOF_k(R_i) = \frac{\sum_{R_j \in CN_k(R_i)} \frac{LRD_k(R_j)}{LRD_k(R_i)}}{Cardinalty(CN_k(R_i))}$$

After the LOF is calculated for all records in R, the records with the highest LOF are the most peculiar and the records with the lowest LOF are the least peculiar.

The *PDD Framework* [8] uses multiple peculiarity measures to determine the overall peculiarity of data from several different views of the data. There are four measures and six views in total.

The views from which the data are examined are *record*, *attribute*, *frequency*, *interval*, *sequence*, and *sequence of differences*. The *record* view is the standard way of representing the data. The *attribute* view stores the values relevant to each attribute (essentially the transpose of the *record* view). The *frequency* view stores the frequency

of each record. The *interval* view stores the intervals between adjacent records once the data has been sorted according to some function. The *sequence* view stores sequences of a given length once the data has been sorted. The *sequence of differences* view stores the differences between elements of a sequence.

Four measures used to determine peculiarity are *cumulative distance* (CD), *fraction of data outside neighborhood* (FDON), *neighborhood density ratio* (NDR), and the *composite z measure* (CZ) [8]. The CD measure is as defined above. The FDON measure calculates the percentage of records that lie outside of its *radius neighborhood*, which is discussed in Knoor's method.

$$FDON_d(R_i) = \frac{Cardinality(\{R_j \in R \mid Distance(R_i, R_j) > d\})}{Cardinality(R)}$$

The NDR measure calculates the ratio of a record's *k-neighborhood* density to the average density of the *k-neighborhoods* of its *k-neighbors*. The density of $CN_k(R_i)$ is the inverse of the average distance from R_i to a record in $CN_k(R_i)$.

$$Density(CN_k(R_i)) = 1/\frac{\sum_{R_j \in CN_k(R_i)} Distance(R_i, R_j)}{Cardinality(CN_k(R_i))}$$

The NDR of record R_i then, is:

$$NDR_k(R_i) = \frac{AverageDensity(CN_k(R_i))}{Density(CN_k(R_i))}$$

where

$$AverageDensity(CN_k(R_i)) = \frac{\sum_{R_j \in CN_k(R_i)} Density(CN_k(R_j))}{Cardinality(CN_k(R_i))}$$

3. Adaptations to the PDD Framework

In general, we used the PDD Framework [8] to guide our analysis of the Engineering and Works database of the City of Regina. However, several parameters and functions had to be specified when the framework was applied in practice. This section describes the detailed adaptations that were applied to the framework.

The FDON measure from the PDD Framework requires a lot of a priori knowledge or trial and error in order to specify a proper *d* value and obtain meaningful results. To avoid this guesswork and allow for seamless movement between data sets, the user is able enter a *d* value if a suitable one is known, or a fraction of the *k-distance* to be used in its place. When using a fraction of the *k-distance*, the FDON measure becomes a more localized measure since the FDON value becomes more influenced by the specific neighborhood of each record. It still follows that the more dense a neighborhood, the lower the FDON value.

The *Density* function from the NDR measure was also altered in our implementation. With the current definition there is no upper limit on the density of a neighborhood. This

leads to great disparities in the density and NDR values. This disparity has a tendency to magnify small differences and leads to irrelevant records with arbitrarily high NDR values. To eliminate this problem we implemented the *Density* function as

$$Density(CN_k(R_i)) = \frac{1}{1 + AverageDistance(CN_k(R_i)) * (1 + \sigma(CN_k(R_i)))}$$

By implementing the *Density* function in this way we can eliminate several problems. We now have $0 < Density(CN_k(R_i)) \leq 1$, so all densities will fall into the same range. Also, $Density(CN_k(R_i)) = 1$ means that all neighbors in $CN_k(R_i)$ are identical to R_i. Lastly, this implementation denies a record the opportunity to have an arbitrarily large NDR value undeservedly.

Before computing the measures, the values for each attribute were normalized. The normalization process conforms all attributes to the same scale without altering the underlying distribution. This step allows us to compute meaningful results from any combination of attributes, regardless of their range in values. Without this step, the peculiarity factor measures are dominated by whichever attribute has the largest range in values.

4. Experimental Results

For our experiments the data were imported into Microsoft Excel. The PDD Framework [8] was implemented using Microsoft Visual Basic. All four measures were implemented and experiments were conducted using the record view of the data.

The experiments presented here were conducted using the City of Regina Engineering and Works Department database. The database is used to track the work done on the infrastructure of the city related to the water and sewer infrastructure. For the purpose of testing and retrieving meaningful results some data in the tables was interpolated. Where there was no Crew Size entered, the average Crew Size value for the given data set was used instead. If a time field was left blank, a value of 0 was assumed.

Results from each experiment will be presented in the following general format. The leftmost columns from each table are used for record identification. These columns can be used for relating the results back to the original data, and are not used in any of the peculiarity computations. The next columns are the attributes that are included in the peculiarity factor measurements. The attributes for these groups of columns are chosen by the user and can be any number of attributes from the original data. Next to the group of attributes chosen for the peculiarity computations are the corresponding normalized values for each respective attribute. After the columns of normalized values is the Density column, which holds the density of the *k-neighborhood* for each corresponding record. Next are the results from each of the measures, *CD*, *FDON*, *NDR*, and lastly *CZ*. The results are ranked according to the *CZ* measure since it incorporates all previous calculated values.

The first three experiments were conducted on the *tblMainLeaks* data, which contains information regarding work done in relation to water leaks. There are 738 total records in this table.

The records included in the first experiment computations were those with Amount of Pipe Used >0 (292 records). The first experiment was conducted using the attributes Amount of Pipe Used, Crew Size, and Total Hours. All values were normalized using z-scores before computations. A neighborhood of size 10 was used. The records detected as the top ten most peculiar from experiment 1 are presented in Table I.

	Workorder	Property ID	Crew	Pipe Used	Crew Size	Total Hrs	Pipe Used z-norm	Crew Size z-norm	Total Hrs z-norm	Den-sity	CD	FDON	NDR	CZ
1	5	21269	A	34	8	11.5	7.3	0.6	1.7	0.12	2210	0.976	2.38	4.13
2	542040	55767	C	27	8	16	5.5	1.7	3.4	0.16	1989	0.983	1.91	3.21
3	73523	41331	B	4	6	5	-0.2	-3	-0.8	0.26	943	0.99	2.97	3.06
4	72759	59290	C	4	6	5.5	-0.2	-3	-0.6	0.26	932	0.99	2.94	3.01
5	75978	47945	A	8	9	7.5	0.8	4	0.2	0.29	1265	0.835	2.26	2.38
6	71192		B	5	6	8	0	-3	0.4	0.28	926	0.986	2.46	2.38
7	542034	23685	C	4	6	10	-0.3	-3	1.1	0.29	982	0.993	2.22	2.15
8	542071	45171	A	3	8	16.5	-0.5	1.7	3.6	0.28	1239	0.983	1.8	1.97
9	90	31625	A	22	8	8.5	4.3	0.6	0.5	0.23	1339	0.976	1.65	1.91
10	43077	31619	A	24	7	8.5	4.9	-1	0.5	0.23	1497	0.979	1.36	1.78

Table I: Top Ten Peculiar Records for the attributes Pipe Used, Crew Size and Total Hours from the tblMainLeaks data

We can see that the first two records are unique in the Amount of Pipe Used and the total hours (we can see this from the high values for the z-norms for the respective categories). The next five records have abnormal crew sizes (in the entire data set, only 1 record has a crew of size 9, and only 5 have a crew of size 6), and notice their NDR (neighborhood density ratio) is quite high, meaning there is something that sets these records apart from other records in their neighborhood. Record eight has very high Total Hours, and records nine and ten both have very high Amount of Pipe Used.

The records included in the second experiment were those with Amount of Pipe Used = 0 (446 records). We selected the columns of Crew Size and Total Hours to be used in the calculations. All values were normalized using z-scores before computations. A neighborhood of size 10 was used. The records detected as the top ten most peculiar from experiment 2 are presented in Table II.

	Workorder	Property ID	Repair Material	Crew Size	Total Hrs	Crew Size z-norm	Total Hrs z-norm	Density	CD	FDON	NDR	CZ
1	542008	41116	Clamp	8	21	1.62	7.06	0.14	3241	0.993	3.32	5.73
2	74846	26560	Other	7	19	-0.7	6.18	0.17	2792.4	0.987	2.65	4.41
3	542046	48715	Clamp	9	4	3.89	-0.7	0.25	1810.4	0.993	3.26	3.99
4	542047	53755	Clamp	9	4	3.89	-0.7	0.25	1810.4	0.993	3.26	3.99
5	71191	5947	Clamp	9	5	3.89	-0	0.27	1784.9	0.993	3.07	3.74
6	71726	41329	Clamp	6	5	-2.9	-0.3	0.26	1371.8	0.989	2.69	2.8
7	542080	15716	Clamp	6	4	-2.9	-0.5	0.27	1380.8	0.989	2.66	2.78
8	542041	19181	Other	6	5	-2.9	-0	0.27	1370.3	0.989	2.66	2.77
9	72763	49529	Clamp	6	3	-2.9	-1.1	0.28	1450.1	0.991	2.5	2.68
10	542098	63929	Clamp	6	7	-2.9	0.64	0.28	1408.1	0.991	2.5	2.63

Table II: Top Ten Peculiar Records for the attributes Crew Size and Total Hours from the tblMainLeaks data

The top two peculiar records are detected because of their abnormally large Total Hours. The next three records have very unique Crew Size values (a size of 9 is very rare in the data set). Also of note is record nine, which has low Crew Size (6) and low Total Hours (3).

All records were included in the third experiment, which used the attributes of Date Received, Date Completed, Crew Size and Total Hours. All values were normalized using z-scores before computations. A neighborhood of size 10 was used. The records detected as the top ten most peculiar from experiment 3 are presented in Table III.

	Workorder	Property ID	Date Re-ceived	Date Repaired	Crew Size	Total Hrs	Date Received z-norm	Date Repaired z-norm	Crew Size z-norm	Total Hrs z-norm	Den-sity	CD	FDON	NDR	CZ
1	542038	55163	7/11/2005	7/12/1995	7	9	5.18	-0.8	-0.6	1.21	0.19	4195	0.999	3.4	5.76
2	542008	41116	1/30/1995	2/2/1995	8	21	-1.1	-1.1	1.67	5.82	0.17	4742	0.995	2.62	4.98
3	542046	48715	8/6/1995	8/6/1995	9	3.5	-0.8	-0.8	3.95	-0.9	0.25	3323	0.997	2.63	3.82
4	542047	53755	8/6/1995	8/6/1995	9	3.5	-0.8	-0.8	3.95	-0.9	0.25	3323	0.997	2.63	3.82
5	74846	26560	11/2/1998	11/5/1998	7	19	1.18	1.21	-0.6	5.05	0.25	4141	0.993	2.07	3.61
6	72763	49529	8/17/1998	8/17/1998	6	2.5	1.05	1.07	-2.9	-1.3	0.28	2827	0.995	2.35	2.95
7	75978	47945	2/25/1999	2/25/1999	9	7.5	1.36	1.4	3.95	0.63	0.29	3523	0.924	2.08	2.93
8	542071	45171	9/28/1995	9/29/1995	8	17	-0.7	-0.7	1.67	4.09	0.28	3548	0.996	1.93	2.9
9	542025	21272	4/21/1995	4/21/1995	6	11	-0.9	-1	-2.9	1.79	0.26	2980	0.996	2.15	2.77
10	542080	15716	11/6/1995	11/6/1995	6	4	-0.6	-0.6	-2.9	-0.7	0.27	2616	0.995	2.31	2.72

Table III: Top Ten Peculiar Records for the attributes Date Received, Date Repaired, Crew Size and Total Hours from the tblMainLeaks data.

A very useful part of peculiarity detection is finding anomalies. This is represented here. The most peculiar record found has a Date Received and Date Repaired that are 10 years apart with Date Received greater than Date Repaired. This is no doubt very peculiar, and most likely an error in the data entry. Records two, five, and eight represent work orders

a with very high Total Hours, as well as records which have a Date Repaired not equal to Date Received. Records three, four, seven, and nine represent abnormal crew sizes of 9 and 6 (only fourteen records in the entire data set of a Crew Size of 9 or 6). Record six has a Crew Size of 6 and a Total Hours of only 2.5, which are very low values for the respective attributes.

Experiment four was conducted on the *tblServiceConnections* data, which consisted of 5435 total records. We selected the Date Received and the Date Repaired categories to be used in the computations. A neighborhood size of 5 was used, and no normalization was done since all data are of the same type and scale (i.e. dates). The records detected as the top ten most peculiar from experiment 4 are presented in Table IV.

	Workorder	Property ID	Date Received	Date Repaired	Density	CD	FDON	NDR	CZ
1	4737	19421	11/15/2004	11/16/1994	8.96E-08	2.4E+07	0.9998	3E+05	0.184
2	82094		10/29/2000	1/13/2000	4.15E-03	1.7E+07	0.9998	216.7	0.112
3	82854	48316	3/14/2000	3/14/2000	0.5	1.6E+07	0.9998	1.47	0.104
4	82820		3/13/2000	3/13/2000	0.47336	1.6E+07	0.9998	1.341	0.104
5	82819	15127	3/10/2000	3/13/2000	0.24137	1.6E+07	0.9998	2.109	0.104
6	82851	29036	3/8/2000	3/8/2000	0.42231	1.6E+07	0.9998	0.959	0.104
7	82814		3/6/2000	3/8/2000	0.35916	1.6E+07	0.9998	0.983	0.104
8	82777	53561	3/2/2000	3/2/2000	0.45174	1.6E+07	0.9998	1.039	0.103
9	82810		3/1/2000	3/1/2000	0.45174	1.6E+07	0.9998	0.952	0.103
10	82764	31775	2/29/2000	3/1/2000	0.43479	1.6E+07	0.9998	0.935	0.103

Table IV: Top Ten Peculiar Records for the attributes Date Received and Date Repaired from the tblServiceConnections Data

It should be noted that these are the top ten peculiar records detected that had both a Date Received and a Date Repaired. The actual records that were detected as the top peculiar results were records that had no Date Received entered. Clearly the first result here is quite peculiar, as it represents a time span of ten years, and the Date Received is later than the Date Repaired. The second record represents the next largest time span in the table, again the Date Received is later than the Date Repaired. The rest of the results are detected as peculiar for less obvious reasons. For example, record 3 does not appear extraordinary, yet when examining the neighborhood of that record we find that there were nine other records that had the same Date Received, and yet this is the only record that had a Date Repaired on the same date.

Experiment five was conducted on the *tblHydrantReplacement* data, which consisted of 291 total records. We selected Meters of C900, CrewSize, RegTime Spent, and OT Time Spent to be used in the computations. A neighborhood of size 10 was used, and z-scores were used for normalization. The records detected as the top ten most peculiar from experiment 5 are presented in Table V.

	Workorder	Problem	m of C900	Crew Size	Reg Time	OT Time	m of C900 z-norm	Crew Size z-norm	Reg Time z-norm	OT Time z-norm	Density	CD	FDON	NDR	CZ
1	700434	Damaged Boot Assembly	1.3	88	0		-0.26	13.4	-2.51	-1.1	6.33E-03	3967	0.993	94.3	8.45
2	72654	Damaged Boot Assembly	1.5	66	6		-0.2	9.75	0.14	-1.1	0.02186	2871	0.993	33.9	4.02
3	700101	New Hydrant Installation	48	7	23		12.9	-0.09	7.64	-1.1	6.57E-02	4418	0.997	7.57	3.87
4	700528	New Hydrant Installation	20	8		5	4.99	0.07	-2.51	0.76	0.16449	1760	0.997	2.64	1.26
5	31400	Other	8	8		11	1.62	0.07	-2.51	2.99	0.28743	1455	0.99	1.81	0.92
6	700434	Vertical Crack on Lead	2	7		12	-0.06	-0.09	-2.51	3.36	0.32171	1446	0.938	1.63	0.76
7	795440	Damaged Boot Assembly	4.7	8		9	0.7	0.07	-2.51	2.24	0.43779	1212	0.976	1.12	0.63
8	71346	Other	1	8		6	-0.34	0.07	-2.51	1.13	0.65536	958	0.986	1.12	0.44
9	13	Damaged Boot Assembly	2	7.6		6	-0.06	0	-2.51	1.13	0.67275	957	0.986	0.97	0.43
10	76933	Leadite Joints on Tee	2.8	7	6	8	0.16	-0.09	0.14	1.87	0.45	916	0.993	1.14	0.42

Table V: Top Ten Peculiar Records for the attributes Meters of C900, Crew Size, RegTime Spent and OT Time Spent from the tblHydrantReplacement data.

Analyzing these results we can see that the top two peculiar records had crew sizes of 88 and 66, respectively. These are definitely anomalies in the data since the average Crew Size is 8. The third and fourth most peculiar records have Meters of C900 of 48 and 20, respectively, again very high for that category. Also, the third most peculiar record has very high Regular hours value. These are the most interesting records, however the remaining records in the top ten have significantly low Regular hours (usually 0) and relatively high OT hours, which may be of interest to the user.

Experiment six was conducted on the *tblValveRepairs* data, which consisted of 1587 total records. We selected the Roadways Cut, Cold Mix Cut, and Fillcrete categories to be used in the computations. Each of these categories is a True(1) and False(0) category. Z-values were still used to normalize the data, although it is not necessary in this case since we know that all fields have the same range of values. The records detected as the top ten most peculiar from experiment 6 are presented in Table VI.

	Workorder	Problem Type	Roadways Cut	Cold Mix Cut	Fillcrete	Roadways Cut z-norm	Cold Mix Cut z-norm	Fillcrete z-norm	Density	CD	FDON	NDR	CZ
1	72159	Valve Box	1	1	0	2.054	2.51	-0.7	0.3	5860	0.999	3.54	15
2	71180	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
3	71181	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
4	71991	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
5	71992	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
6	72142	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
7	72143	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
8	72144	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
9	73116	Valve Box	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1
10	74283	Valve	0	1	1	-0.487	2.51	1.42	1	5247	0.98	1	1

Table VI: Top Ten Peculiar Records for the attributes Roadways Cut, Cold Mix Cut, and Fillcrete from the tblValveRepairs data.

In this data set there was only one significantly peculiar record. Since the range of values for the fields is so small [0,1], there are many records that have identical attribute values (and hence a Neighborhood Density of 1). However, the only record which has *both* Roadways Cut=True(1) and Cold Mix Cut=True(1) would have no similar records, and hence stands out as being peculiar.

5. Conclusion

In this paper we have presented our results obtained from applying the PDD Framework [8] to the City of Regina Engineering and Works Department Database. This database tracks the work done in relation to the city's water and sewer infrastructure. The purpose in these experiments was to determine the applicability of the results obtained through peculiarity mining. Six different experiments were conducted using a diverse selection of attributes from the database.

The results from Tables I, II, and III show how different results can be obtained by filtering the data or by including or excluding certain attributes which you may or may not want to influence the peculiarity of a record. This is represented by the fact that the top peculiar records from the first two experiments are completely different than the top peculiar records from the third experiment, even though the measures were ran on the same data. The results from the fourth and fifth experiment indicate that peculiarity mining can also be used to find records that are anomalies because they may contain errors. The most peculiar results were either missing data in a field, contained contradictions, or contained values that were a significant deviation from the normal. The sixth experiment was run on data that were very simple, with a very small range. This should make it more difficult to detect irregularities, since the records are very similar, yet the one record that differs from all the others is the top peculiar record by a substantial amount.

These experiments demonstrate the various benefits and uses of peculiarity mining. This implementation proved effective in detecting irregularities in the various forms of data presented to it. Experiments were conducted using data from various sources and using combinations of attributes that required normalization to be effective. Careful examination of the records identified as peculiar showed that in many cases a human expert agreed that the data were peculiar. Each experiment resulted in meaningful and plausible peculiarities that may be of interest.

References

[1] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. "OPTICS-OF: Identifying Local Outliers," *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases,* Prague, Czech Republic (1999) 262-270.
[2] Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J. "LOF–Identifying Density-Based Local Outliers," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX (2000) 93-104.
[3] Jin, W., Tung, A.K.H., and Han, J. "Mining Top-n Local Outliers inLarge Databases," *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California (2001) 293-298.

[4] Knorr, E.M., and Ng, R.T. “Algorithm for Mining Distance-Based Outliers in Large Datasets,” *Proceedings of the 24th International Conference on Very Large Databases*, New York (1998) 392-403.
[5] Knorr, E.M., and Ng, R.T. “A Unified Notion of Outlier: Properties and Computation,” *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, San Diego, CA (1997) 219-222.
[6] Knorr, E.M., Ng, R.T., and Tucakov, V. “Distance-Based Outliers: Algorithms and Applications,” *International Journal on Very Large Data Bases*, Volume 8, Springer (2000) 237-253.
[7] Portnoy, L. *Intrusion Detection with Unlabeled Data Using Clustering*, Undergraduate Thesis, Columbia University (2000).
[8] Mahesh Shrestha, M., Hamilton, H.J., Yao, Y.Y., Konkel, K., and Zhong, N., “The PDD Framework for Detecting Categories of Peculiar Data,” Submitted, June 2005.
[9] N. Zhong, N., Liu, C., Yao, Y.Y., Ohshima, M., Huang, M., and Huang, J. ”Relational Peculiarity Oriented Data Mining,” *Proceedings of the Fourth IEEE International Conference on Data Mining* Brighton, UK (2004) 575-578
[10] Zhong, N., Yao, Y.Y., Ohshima, M. “Peculiarity Oriented Multidatabase Mining,” *IEEE Transactions on Knowledge and Data Engineering*,Vol. 15, No. 4 (2003) 952-960
[11] Zhong, N., Yao, Y.Y., Ohshima, M., and Ohsuga, S. “Interestingness, Peculiarity, and Multi-Database Mining,” *Proceedings of the 2001 IEEE International Conference on Data Mining*, San Jose, CA, (2001) 566-573.