Transactions on Machine Learning
and Data Mining
Vol.1, No 2 (2008) 83-96
© ISSN: 1865-6781 (Journal),
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks

Ron S. Kenett[1] and Silvia Salini[2]

[1]KPA Ltd., Raanana, Israel and University of Torino, Torino, Italy
ron@kpa.co.it
[2]Department of Economics, Business and Statistics, University of Milan, Italy
silvia.salini@unimi.it

Abstract. Association rules are one of the most popular unsupervised data mining methods. Once obtained, the list of association rules extractable from a given dataset is compared in order to evaluate their importance level. The measures commonly used to assess the strength of an association rule are the indexes of support, confidence, and lift. Relative Linkage Disequilibrium (RLD) was proposed in as an approach to analyse both quantitatively and graphically association rules RLD can be considered an adaptation of the lift measure with the advantage that it presents more effectively the deviation of the support of the whole rule from the support expected under independence. Moreover RLD can be interpreted graphically using a simplex representation leading to powerful graphical display of association relationships. In this paper we demonstrate the strength of RLD by applying it to two large data sets. One data set consists of 2008 aircraft accident and incident occurrences recorded in the FAA data base. The other data set consists of operational risks captured by a large financial institution operating under Basel II regulations.

Keywords: association rules, simplex representation, text mining, Relative Linkage Disequilibrium

1. Introduction

Relative Linkage Disequilibrium (RLD), as an association measure for assessing association rules was first presented at the ICDM conference in 2008 (see [1]). In this paper, we first review RLD in Section 2 and the related simplex graphical representation and then proceed to present two applications to large textual data sets.

The first example consists of description of aircraft accidents, described in Section 3. The second example in Section 4 is from a large financial organization tracking operational risks. We conclude with direction for further research in Section 4.

2. Relative Linkage Disequilibrium and Simplex Representations

Relative Linkage Disequilibrium (RLD) is an association measure motivated by indices used in population genetics to assess stability of the genetic composition of populations and exploratory analysis methods applied to contingency tables (see [1] and [2]). To define RLD consider a set transactions with items A and B on the Left Hand Side (LHS) and Right Hand Side (RHS) of an association rule. These two events generate four combinations whose frequencies are described in Table 1 below:

Table 1: The association rules contingency table

	B	^B
A	x_1	x_2
^A	x_3	x_4

$$\sum_{i=1}^{4} x_i = 1,\ 0 \le x_i, i = 1...4.$$

$x_1 =$ the relative frequency of occurrence of both A and B

$x_2 =$ the relative frequency of transaction (item sets) where only A occurs

$x_3 =$ the relative frequency of transaction (item sets) where only B occurs

$x_4 =$ the relative frequency of transaction (item sets) where neither A or B occur

There is a natural one to one correspondence between the set of all possible 2x2 contingency tables, such as Table 1, and points on a simplex (see Figure 1). We exploit this graphical representation to map out association rules. The tables that correspond to independence in the occurrence of A and B, correspond to a specific surface within the simplex presented in Figure 1. By "independence" we mean that knowledge of frequencies of A and B is sufficient to reconstruct the entire table, the items A and B do not interact.

Let $D = x_1x_4 - x_2x_3$ $f = x_1 + x_3$ and $g = x_1 + x_2$.

$f =$ relative frequency of item B

$g =$ relative frequency of item A

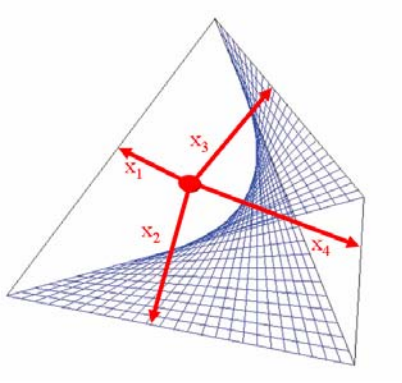

Figure1: The surface of independence (D=0)

It can be easily verified that:

$$x_1 = fg + D$$
$$x_2 = (1-f)g - D$$
$$x_3 = f(1-g) - D$$
$$x_4 = (1-f)(1-g) + D$$

The geometric interpretation of D makes it an appealing measure of interaction. The surface on Figure 1 represents all association rules with $D = 0$. However points closer to the edges of the simplex will have intrinsically smaller values of D. The Relative Linkage Disequilibrium standardizes D by the distance D_M from point corresponding to the contingency table in the simplex to the surface D=0 in the direction (1, -1, -1, 1). RLD is therefore computed as D/D_M.

The computation of RLD can be performed through the following algorithm:

If $D > 0$
then
if $x_3 < x_2$
then $RLD = \dfrac{D}{D + x_3}$
else $RLD = \dfrac{D}{D + x_2}$

$$
\begin{aligned}
&else \\
&if \quad x_1 < x_4 \\
&\quad then \quad RLD = \frac{D}{D - x_1} \\
&\quad else \quad RLD = \frac{D}{D - x_4}
\end{aligned}
$$

Some asymptotic properties of RLD are available [2] and can be used for statistical inference.

The **arules** extension package for R (see [3]) provides the infrastructure needed to create and manipulate input data sets for the mining algorithms and for analyzing the resulting *itemsets* and rules. Since it is common to work with large sets of rules the package uses sparse matrix representations to minimize memory usage. The 2008 version 06-6 now incorporated RLD which has thus become officially incorporated into the R infrastructure (see [4]).

The function interestMeasure() in **arules** is used to calculate a variety of measures such as *support*, *confidence* and *lift* including missing information derived from the transactions used to mine the associations.

The *support* for a rule A => B is obtained by dividing the number of transactions which satisfy the rule, N{A=>B}, by the total number of transactions, N

$$support\ \{A=>B\} = N\{A=>B\} / N = x_1$$

The support is therefore the frequency of events for which both the LHS and RHS of the rule hold true. The higher the *support* the stronger the information that both type of events occur together.

The *confidence* of the rule A => B is obtained by dividing the number of transactions which satisfy the rule N{A=>B} by the number of transactions which contain the body of the rule, A.

$$confidence\ \{A=>B\} = N\{A=>B\} / N\{A\} = x_1 /(x_1 + x_2) = x_1 / g$$

The confidence is the conditional probability of the RHS holding true given that the LHS holds true. A high *confidence* that the LHS event leads to the RHS event implies causation or statistical dependence.

The *lift* of the rule A => B is the deviation of the support of the whole rule from the support expected under independence given the supports of the LHS (A) and the RHS (B).

$$
\begin{aligned}
&lift\ \{A=>B\} = confidence\{A=>B\} / support\{B\} \\
&= support\{A=>B\}/support\{A\}support\{B\} \\
&= x_1 /((x_1 + x_2) * (x_1 + x_3)) = x_1 /(g * f)
\end{aligned}
$$

Lift is an indication of the effect that knowledge that LHS holds true has on the probability of the RHS holding true. When lift is exactly 1 we see no effect (LHS and RHS independent), no relationship between events. For lift greater than 1 we have positive effect (given that the LHS holds true, it is more likely that the RHS holds true), i.e. positive dependence between events. If lift is smaller than 1 we have negative effect (when the LHS holds true, it is less likely that the RHS holds true), i.e. negative dependence between events.

The traditional measures available in arules for *itemsets* are:

- All-confidence [5]
- Cross-support ratio [6]
- Chi square measure[7]
- Conviction [8]
- Hyper-lift and hyper-confidence [9]
- Leverage [10]
- Improvement [11]
- Miscellaneous measures from [12] (e.g., cosine, Gini index, φ- coefficient, odds ratio)

RLD, the Relative Linkage Disequilibrium we developed ([1], [2]) has been recently introduced to the arules package, see [4]

In this paper we implement the Relative Linkage Disequilibrium measure (RLD) available in the function InterestMeasure() and we use the function quadplot() and triplot() of the library **klaR** [13] to produce the simplex 3D and 2D representation.

As an example consider the classical R "groceries" market basket analysis data set. The groceries data set ([4]) contain 30 days of point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions and the items are aggregated to 169 categories. In Table 2, the first 20 rules sorted by lift are displayed. For each rule the RLD, the odds Ratio and the Chi Square are reported.

We apply to the data the apriori algorithm setting minimum support 0.1 and minimum confidence 0.8 ([14]). We obtain 200 rules. The aim of this example is to show the intuitive interpretation of RLD through his useful graphical representation. Figure 2 shows the simplex representation of the contingency tables corresponding to the 200 rules and the top 10 rules sorted by RLD. We pick up from this analysis the important association between "citrus fruit" and "tropical fruit" and "other vegetables", something not obvious when considering support and lift. We can also represent the relative position of these rule graphically. The corners represent four tables with relative frequency 1 = (0,1,0,0), 2 = (1,0,0,0), 3 = (0,0,1,0) and 4 = (0,0,0,1).

Table 2: First 20 rules for the groceries data, sorted by Lift.

lhs	rhs	supp	conf	lift	RLD	odds	chi
{whole milk, yogurt}	{curd}	0.010	0.180	3.372	0.141	4.566	184.870
{citrus fruit, other vegetables}	{root vegetables}	0.010	0.359	3.295	0.281	4.958	188.438
{other vegetables,yogurt}	{whipped/sour cream}	0.010	0.234	3.267	0.175	4.450	177.154
other vegetables}	{root vegetables}	0.012	0.343	3.145	0.262	4.679	206.042
{root vegetables}	{beef}	0.017	0.160	3.040	0.250	4.631	277.341
{beef}	{root vegetables}	0.017	0.331	3.040	0.250	4.631	277.341
{citrus fruit, root vegetables}	{other vegetables}	0.010	0.586	3.030	0.487	6.183	175.058
{tropical fruit,, root vegetables}	{other vegetables}	0.012	0.585	3.021	0.485	6.195	207.203
{other vegetables, whole milk}	{root vegetables}	0.023	0.310	2.842	0.225	4.390	330.231
{other vegetables, whole milk}	{butter}	0.012	0.154	2.771	0.143	3.639	146.317
{whole milk, curd}	{yogurt}	0.010	0.385	2.761	0.286	4.088	132.726
{whipped/sour cream}	{curd}	0.011	0.146	2.742	0.135	3.539	129.718
{curd}	{whipped/sour cream}	0.011	0.197	2.742	0.135	3.539	129.718
{other vegetables, whole milk}	{whipped/sour cream}	0.015	0.196	2.729	0.140	3.702	183.728
{other vegetables, yogurt}	{root vegetables}	0.013	0.297	2.729	0.212	3.791	163.187
{whole milk, yogurt}	{whipped/sour cream}	0.011	0.194	2.709	0.132	3.500	131.650
{other vegetables, yogurt}	{tropical fruit}	0.012	0.283	2.701	0.199	3.688	151.333
{root vegetables, other vegetables}	{citrus fruit}	0.010	0.219	2.645	0.148	3.407	119.391
{other vegetables, rolls/buns}	{root vegetables}	0.012	0.286	2.628	0.199	3.568	141.814
{tropical fruit, whole milk}	{root vegetables}	0.012	0.284	2.602	0.196	3.514	136.436

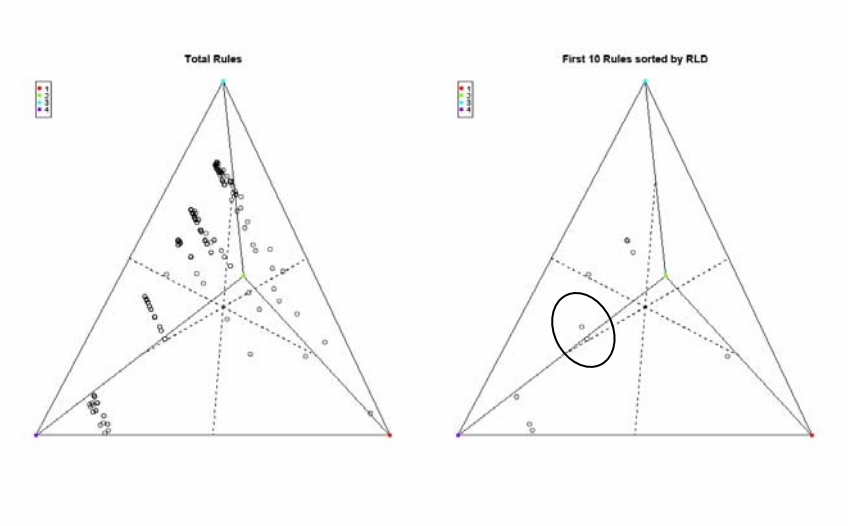

Figure 2: Simplex representation of groceries association rules

3. RLD Application to Aircraft Accident Data

An aircraft accident is an occurrence associated with the operation of an aircraft in which people suffer death or injury, and/or in which aircraft receives substantial

damage; an incident is an occurrence which is not an accident but is a safety hazard and with addition of one or more factors could have resulted in injury or fatality, and/or substantial damage to the aircraft [15]. Previous research on aircraft accidents has focused on studying accident data to determine factors leading to accidents ([16], [17], and [18]). The low rate of accidents however, makes it difficult to discover repeating patterns of these factors. Other research has analyzed larger sets of data available on incidents to determine the causal factors of incidents ([19]]). Nazerri et al present an application of association rules for analyzing such data ([20]).

All accidents in the United States involving civil aircraft are investigated by the National Transportation Safety Board (NTSB), an independent organization, and are reported in the NTSB database. Accident data, therefore, can be assumed complete and free of bias. Incident data however, are under-reported and subject to self-reporting bias. To address these constraints, one can analyze the underlying factors of accidents and incidents *qualitatively*. The historical data on incidents is large enough to represent these factors qualitatively. Also, we can consider *all* factors that have been present in an event, regardless of their primary or contributory role in leading to the event. This minimizes the impact of the bias in reporting the factors.

We apply here RLD to the qualitative textual reports. We consider the event in 2008 (unknown event type, occurrence, accident, incident). We consider 40 variables that can be clustered in 5 factors: weather, location and period, flight characteristics, aircraft characteristics, injury. In the data preparation phase we discretize the numerical variable and we obtain a new data set with 1291 transactions and 984 items (the levels of the 40 variables). The apriori algorithm produce 234,869 rules. We focus on the rules with RHS event type "accident" and we calculate for each rule RLD. Figure 3 show the first 200 rules with high level of RLD.

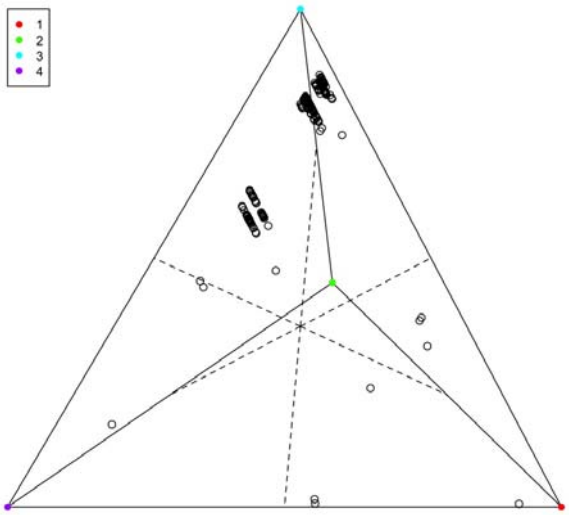

Figure 3: Simplex representation of the first 200 rules sorted by RLD for aircraft accident data set

The more frequent variables associated with event type accident are connected to weather conditions (wind speed, wind direction, sky conditions, visibility) and injury type and number. In some rules, variables related to location and period appear occasionally. Apparently the traffic conditions and the communications aspects related to Air Traffic Control (ATC) are associated with such accident events. This confirms the results of Nazeri et al ([20]). Moreover, it seems that the type of flight and the type of aircraft are not relevant in the accident event. In the Nazeri analysis the Aircraft characteristics was associated with incidents and not with accidents.

4. RLD Application to Operational Risk Data

Operational risk in the banking industry is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events ([21]). These include:

- Internal fraud
- External fraud
- Employment practices & workplace safety
- Clients, products & business practices
- Damage to physical assets
- Business disruption & system failures
- Execution, delivery & process management
- Includes legal risk.

Operational risks do however exclude reputational and business/strategic risk.

The rising interest of the banking industry in operational risks is due, among other reasons, to the globalization of the financial markets, the growth of IT applications, and the increasing diffusion of sophisticated financial products. The Basel II capital accord requires banks to put aside a minimum capital requirement which matches its exposure to credit risk, market risk and operational risk. Specifically, a 12% of minimum capital requirement needs to be allocated to operational risks ([21]).

The Basel II agreement splits operational risk exposures and losses into a series of standardized business units, called '*business lines*', and into groups of operational risk losses according to the nature of the underlying operational risk event, called '*event types*'. In [22], a comprehensive Loss Data Collection Exercise (LDCE) initiated by the Basel II Committee, through the work of its Operational Risk Subgroup of the Accord Implementation Group (AIGOR), is described. The exercise follows other similar exercises sponsored by the Basel Committee and individual member countries over the last five years. The 2008 LDCE is a significant step in the Basel Committee's efforts to address Basel II implementation and post-implementation issues more consistently across member jurisdictions. While similar to two previous international LDCEs, which focused on internal loss data, this LDCE is the first international effort to collect information on all four operational risk data elements - internal data, external data, scenario analysis, and business environment and internal control factors (BEICFs) - used in an Advanced Measurement Approach (AMA) for calculating operational risk capital charges under Basel II. As an independent contribution to the LDCE we present here the application of RLD to internal operational risk data

collected by a large banking institution. Our goal is to demonstrate, with a concrete example, how RLD can be used to assess risks reported in such organizations using textual reports.

We consider a data set of operational risk events with 20 variables, some categorical, some continuous and some textual with a description of the loss event. Examples of such descriptions are:

"Booked on fixed income trade that was in the wrong partfund code. Have cancelled trade resultant in error of 15000"

"Cash contribution not invested due to incorrect fax number used by client. Not our error but noted due to performance impact on the fund."

"The client sent a disinvestment instruction that was incorrectly processed as an investment. Due to a positive movement in the equity markets the correction of the error led to a gain.

In the data preparation phase we discretized the continuous variables (expected and actual values of loss) and, using the library **tm** of R [23], we selected the textual description variables, in particular, *activity*, *process* and *risk* type. Than, the data was processed for an association rules analysis. Following these steps we obtain a new data set with 2515 transactions and 235 items (the levels of the variables). The apriori algorithm produces 345,575 rules[1]. With such a large number of rules traditional measures of association typically cannot identify "interesting" associations. Too many rules with too little a difference between them. Moreover, with traditional measures of association, it is often difficult to explore and cluster rules in an association rules analysis. RLD and its complementary simplex representation help us tackle this problem. For each rule, we calculate RLD and sort the rules accordingly. Figure 4 shows the first 200 rules with the highest level of RLD.

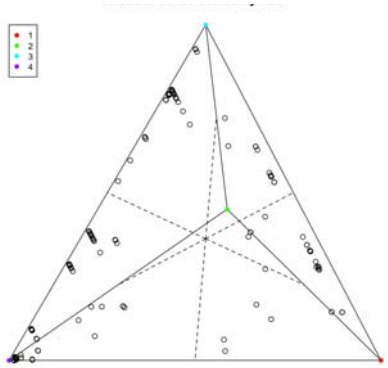

Figure 4: Simplex representation of the first 200 rules sorted by RLD for operational risk data set

[1] We modify the default level of support in the arules algorithm of R, we set a very low level of support 0,01. This is useful in operational risk application, because we expect that the loss event are not so frequent.

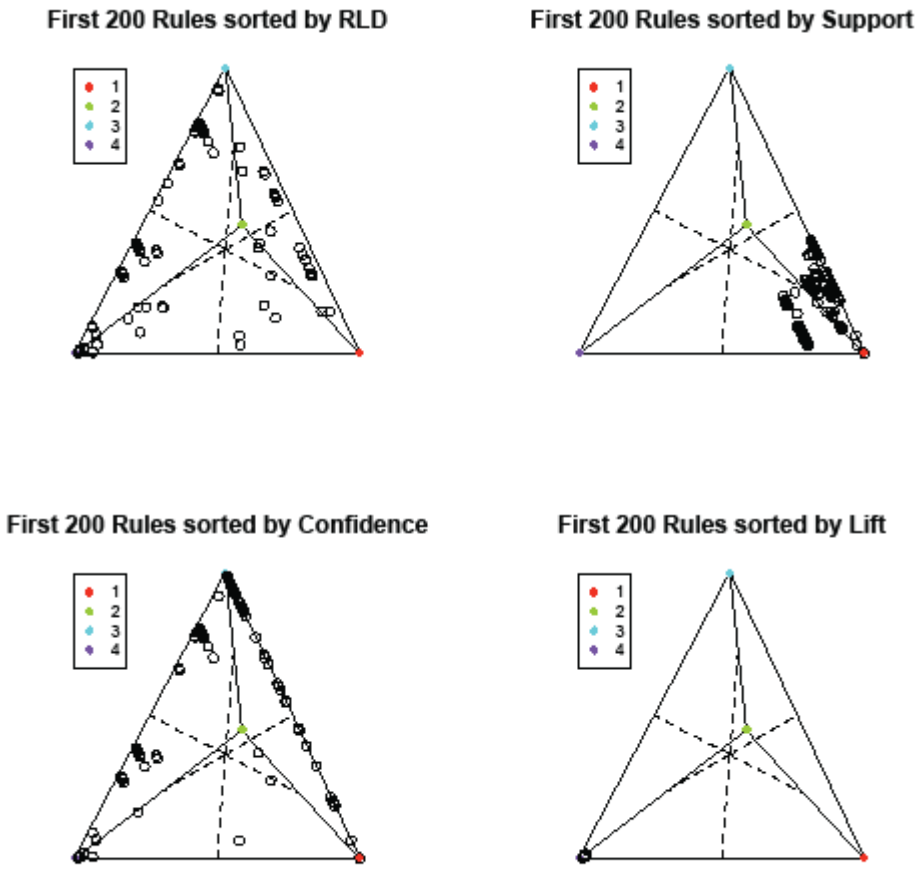

Figure 5: Comparison of the first 200 rules sorted by RLD, support, confidence and lift for the operational risk data set

We can contrast the top 200 rules derived from sorting association rules by support, confidence and lift with RLD (see Figure 5). RLD clearly provides the highest resolution and interesting spread.

We proceed with an automatic clustering of the rules. This is applied here to the first 200 rules sorted by RLD, but can be done for all rules.

The hierarchical cluster analysis is applied to the elements in the association rules contingency table shown in Table 1, that shows the numbers that we use in the

calculation of RLD. Figure 6 shows the cluster dendrogram with a highlight of 12 clusters of association rules.

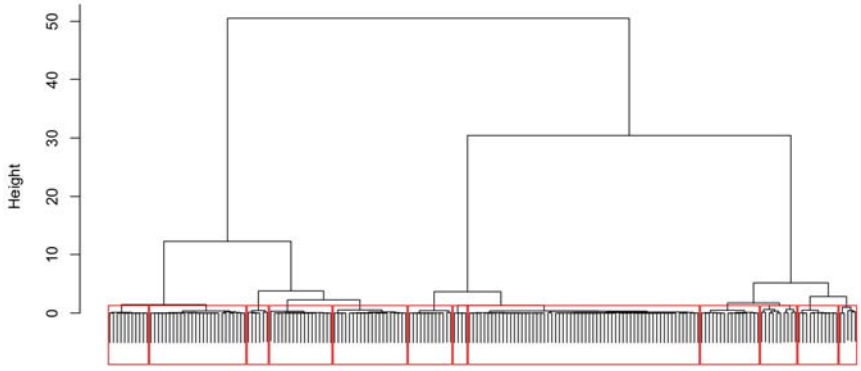

Figure 6: Cluster dendrogram for the 200 rules for operational risk data set

Now we produce a simplex representation for each one of the clusters. Figure 7 shows these plots. Rules in the same cluster have similar type of association. All the rules in these plots have a very high level of RLD, near 1, but different values for the other association measure. For example the rules in the left bottom corner of the clusters 5, 10 and 12 are characterized by very low support and very high lift. On the contrary rules in clusters 2 and 3 have high support, high confidence and low lift. In the cluster 11 there are the rules with confidence equal to 1, lift nearer 1 and very low support, etc…

This example demonstrates the unique property of RLD, using a real data set. We conclude with a summary and some direction for future work.

5 Summary and Future Work

Relative Linkage Disequilibrium is a useful measure in the context of association rules, especially for its intuitive quantitative and visual interpretation. An inherent advantage to informative graphical displays is that the experience and intuition of the experimenter who collects the data can contribute to the statistician's data analysis.

The context for applications of RLD ranges from web sites, customer satisfaction surveys, operational risks, call centers and many other sources of textual data. The paper showed how RLD is used to select and cluster association rules. An important future research area is the exploration of statistical properties of RLD in the context of association rules in large data sets expanding the results presented in [2]. Moreover, RLD can contribute to identify rare events in large text files, events called "black swans" by N. Taleb (see [24]). Combining RLD with simplex representations can help display item sets with low support exhibiting significant association patterns. Such association rules correspond to points distant from the simplex corner labeled "2"

facing us, in the simplex representations of the figures above. This paper provides an introduction to Relative Linkage Disequilibrium with applications. Hopefully it will

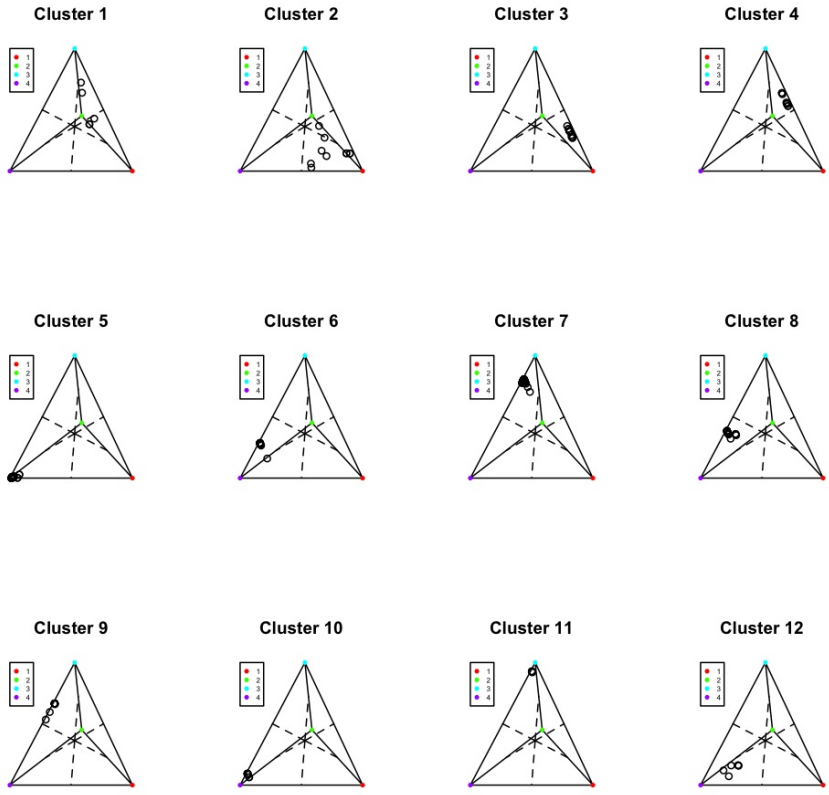

Figure 7: Cluster simplex plot for the 200 rules for operational risk data set

stimulate more research on association rules and their close relationship with contingency tables.

Acknowledgments

The first author was partially supported by the FP6 project MUSING 2006 contract number 027097, 2006-2010. The motivation for developing RLD in the context of association rules is based on discussions and needs expressed by the MUSING

partners. Contributions from P. Lombardi, P. Giudici, S. Figini, P. Cerchielo and R. Pike are gratefully acknowledged.

References

1. Kenett, R. and Salini, S.: "Relative Linkage Disequilibrium: A new measure for association rules", in: P. Perner (Ed.), *Advances in Data Mining: Medial Applications, E-Commerce, Marketing, and Theoretical Aspects*, ICDM 2008, Leipzig, Germany, July, 2008. Lecture Notes in Computer Science, Springer Verlag, Volume 5077, 2008.
2. Kenett, R.: "On an Exploratory Analysis of Contingency Tables". *The Statistician*, 32, pp. 395-403, 1983.
3. Hahsler, M., Gr¨un, B., and Hornik, K.: "arules – A computational environment for mining association rules and frequent item sets". *Journal of Statistical Software*, 14(15):1–25. ISSN 1548-7660. URL http://www.jstatsoft.org/v14/i15/, 2005.
4. Hahsler, M., Gr¨un, B., and Hornik, K.: "The arules package: Mining Association Rules and Frequent Itemsets, version 0.6-6", http://cran.r-project.org/web/packages/arules/index.html, 2008.
5. Omiecinski, E.: "Alternative interest measures for mining associations in databases". *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.
6. Xiong H., Tan P.-N., and Kumar V.: "Mining strong affinity association patterns in data sets with skewed support distribution". In B. Goethals and M. J. Zaki, editors, *Proceedings of the IEEE International Conference on Data Mining*, November 19–22, Melbourne, Florida, pages 387–394, 2003.
7. Kenett, R. and Zacks, S..: *Modern Industrial* ***Statistics: Design and Control of Quality and Reliability***, Duxbury Press, San Francisco, 1998, Spanish edition 2000, 2nd paperback edition 2002, Chinese edition, 2004.
8. Brin S., Motwani, R., Ullman, J., and Tsur, S.: "Dynamic itemset counting and implication rules for market basket data". In SIGMOD 1997, *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA, 1997.
9. Hahsler, M., Kurt Hornik, and T. Reutterer: "Implications of probabilistic data modeling for mining association rules". In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization*, pages 598–605. Springer-Verlag, 2006.
10. Piatetsky-Shapiro, G.: "Discovery, analysis, and presentation of strong rules". In: *Knowledge Discovery in Databases*, pages 229–248, 1991.
11. Bayardo R., Agrawal R., and Gunopulos D.: "Constraint-based rule mining in large, dense databases". *Data Mining and Knowledge Discovery*, 4(2/3):217–240, 2000.
12. Tan, P-N, Kumar, V., and Srivastava, J.: "Selecting the right objective measure for association analysis". *Information Systems*, 29(4):293–313, 2004.
13. Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., Zentgraf , M., "The klaR package: Classification and visualization, version 0.5-7", http://cran.r-project.org/web/packages/klaR/index.html, 2008.
14. Borgelt, C.: "Apriori – Finding Association Rules/Hyperedges with the Apriori Algorithm". Working Group of Neural Networks and Fuzzy Systems, Otto-von-Guericke-University of Magdeburg, Universit ¨atsplatz 2, D-39106 Magdeburg, Germany,. URL http://fuzzy.cs.uni- magdeburg.de/~borgelt/apriori.html, 2004.
15. Federal Aviation Administration, Air Traffic Organization, *Aircraft Accident and* Incident Notification, Investigation, and Reporting. Order 8020.16

http://www.faa.gov/airports_airtraffic/air_traffic/publications/at_orders/media/AAI.pdf, 2008
16. Ladkin, P.: *ATM Related Accidents*. Eurocontrol. http://www.eurocontrol.int/corporate/public/standrd_page/cb_safety.html, 2008
17. Majumdar, A., Dupuy, M.D., and Ochieng, W.O.:"A framework for the Development of Safety Indicators for New Zealand Airspace: the Categorical Analysis of Factors Affecting Loss of Separation Incidents". Transportation Research Board (TRB) annual conference. (2006)
18. Hansen, M. and Zhang, Y.: "Safety Efficiency: Link between Operational Performance and Operation Errors in the national Airspace System". Transportation Research Record, *Journal of Transportation Research Board*, no. 1888, p 15. (2004)
19. National Aeronautics and Space Administration, Air Traffic Management System. (2007) From website: http://quest.arc.nasa.gov/aero/virtual/demo/ATM/tutorial/tutorial1.html
20. Nazeri, Z., Barbara, D., De Jong, K., Donohue, G., Sherry, L.: "Contrast-Set Mining of Aircraft Accident and Incident Data", in P. Perner (Ed.), *Advances in Data Mining: Medial Applications, E-Commerce, Marketing, and Theoretical Aspects*, ICDM 2008, Leipzig, Germany, July, 2008. Lecture Notes in Computer Science, Springer Verlag, Volume 5077, 2008.
21. Basel Committee on Banking Supervision, *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework*, http://www.bis.org/publ/bcbs107.htm, 2004.
22. Basel Committee on Banking Supervision, *Operational Risk - 2008 Loss Data Collection Exercise*, http://www.bis.org/publ/bcbs_nl13.htm, 2008
23. Feinerer I (2007). "tm: Text Mining Package. R package version 0.3", URL http://CRAN.R-project.org/package=tm.
24. Taleb, N. (2007), *The Black Swan: The impact of the highly improbable*, Random House, NY.