

Association Mining in Large Databases: A Re-examination of Its Measures^{*}

Tianyi Wu¹, Yuguo Chen², and Jiawei Han¹

¹ Department of Computer Science, UIUC

² Department of Statistics, UIUC

{twu5,yuguo,hanj}@uiuc.edu

Abstract. In the literature of data mining and statistics, numerous interestingness measures have been proposed to disclose succinct object relationships of association patterns. However, it is still not clear when a measure is truly effective in large data sets. Recent studies have identified a critical property, *null-(transaction) invariance*, for measuring event associations in large data sets, but many existing measures do not have this property. We thus re-examine the null-invariant measures and find interestingly that they can be expressed as a generalized mathematical mean, and there exists a total ordering of them. This ordering provides insights into the underlying philosophy of the measures and helps us understand and select the proper measure for different applications.

1 Introduction

Despite more than a decade of study over association mining, it has been well recognized that traditional association rules may not disclose truly interesting event relationships [2]. For example, mining a market basket data set may find a rule, “*coffee* \rightarrow *milk*”, with nontrivial support and high confidence (e.g., 80%), but this does not imply that *buying coffee* and *buying milk* are strongly associated because *milk* itself might be popular. Thus, researchers have proposed various measures as constraints to mine true relationships among events [3,11,10,4].

Many association, correlation, and similarity measures have been proposed for analyzing the relationships among discretized events [6,12]. For example, χ^2 is a typical measure for analyzing correlations among discretized events in statistics [7]. However, it may not be an appropriate measure for analyzing event associations in large transaction databases. Notice in a typical transaction database, a particular item i (e.g., *coffee*) appearing in a transaction T (i.e., $i \in T$) is often a small probability event. Since most transactions do not contain item i , they are *null transactions w.r.t. i*. If the association among a set of events being analyzed is affected by the transactions that contain none of them (i.e., null-transactions), such a measure is unlikely to be of high quality. Recent studies

^{*} The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678, NSF BDI-05-15813, and NSF DMS-05-03981. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Table 1. Two-event contingency table

	<i>milk</i>	\overline{milk}	Σ_{row}
<i>coffee</i>	<i>mc</i>	\overline{mc}	<i>c</i>
\overline{coffee}	$m\overline{c}$	$\overline{m\overline{c}}$	\overline{c}
Σ_{col}	<i>m</i>	\overline{m}	total

[12,9,10] have shown that null-(transaction) invariance is critically important for an interestingness measure. We use the following example to illustrate this.

Example 1. In a typical shopping transaction database, a product appearing in a transaction is called an **event**, and a set of products appearing in a transaction is called an **event-set**. Association analysis is to identify interesting (positive or negative) associations among a set of events. It is expected that a particular event happens with a very low probability.

In Table 1, the purchase history of two events *milk* and *coffee* are summarized by their support counts, where, for instance, *mc* denotes the support of the event-set “*coffee and milk*”, i.e., the occurrences of transactions containing them. Table 2 enumerates six data sets in terms of a “flattened” contingency table. We then select six measures: χ^2 , *Lift*, *AllConf*, *Coherence*, *Cosine*, *Kulczynski* (denoted as *Kulc* hereafter), and *MaxConf*, and show their results for each data sets. The definitions for the six measures are given in Table 3. *AllConf*, *Coherence*¹, *Cosine*, and *MaxConf*² are the only ones that are not sensitive to the number of null-transactions (hence called *null-invariant measures*) among over 20 interestingness measures studied in [12]. *Kulc* is another null-invariant measure proposed in [1]. Two popular but not null-invariant measures, χ^2 and *Lift* [3,6], are listed for comparison.

Let’s examine D_1 and D_2 , where *milk* and *coffee* are positively associated because *mc* is considerably greater than \overline{mc} and $m\overline{c}$. The results of the five null-invariant measures show that *m* and *c* are strongly positively associated in both data sets. However, *Lift* and χ^2 generate dramatically different values, due to their sensitivity to \overline{mc} . In such cases, since \overline{mc} is usually huge and unstable, a good interestingness measure should not be affected by it. Similarly, in D_3 , the five null-invariant measures correctly show that *m* and *c* are strongly negatively associated; whereas *Lift* and χ^2 judge it in an incorrect or controversial way. For D_4 , both *Lift* and χ^2 indicate a highly positive association between *m* and *c*, whereas the others³ a neutral association, because $mc : \overline{mc} = mc : m\overline{c} = 1 : 1$. This means that given the event *coffee*, the probability of the event *milk* is exactly 50% and vice versa. Note that *milk* and *coffee* are statistically independent if and only if $P(mc) = P(m)P(c)$, where \overline{mc} cannot be ignored. In fact,

¹ Notice that *Coherence*(*a, b*), though introduced lately [10] and defined differently, is essentially the popularly used Jaccard Coefficient [12].

² We use *MaxConf* instead of *Confidence* as in [12] to avoid any confusion with the directional “confidence” measure in traditional association rule mining.

³ The neutral point of *Coherence* is at 0.33 instead of 0.5 (see [9]).

Table 2. Example data sets

Data set	mc	\overline{mc}	$m\overline{c}$	$\overline{m\overline{c}}$	χ^2	Lift	AllConf	Coherence	Cosine	Kulc	MaxConf
D_1	10,000	1,000	1,000	100,000	90557	9.26	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0	1	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	670	8.44	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	24740	25.75	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	8173	9.18	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	965	1.97	0.01	0.01	0.10	0.5	0.99

Table 3. Interestingness measure definitions

Measure	Definition	Range	Null-Invariant
$\chi^2(a, b)$	$\sum_{i,j=0,1} \frac{(e(a_i, b_j) - o(a_i, b_j))^2}{e(a_i, b_j)}$	$[0, \infty]$	No
Lift(a, b)	$\frac{P(ab)}{P(a)P(b)}$	$[0, \infty]$	No
AllConf(a, b)	$\frac{sup(ab)}{max\{sup(a), sup(b)\}}$	$[0, 1]$	Yes
Coherence(a, b)	$\frac{sup(ab)}{sup(a) + sup(b) - sup(ab)}$	$[0, 1]$	Yes
Cosine(a, b)	$\frac{sup(ab)}{\sqrt{sup(a)sup(b)}}$	$[0, 1]$	Yes
Kulc(a, b)	$\frac{sup(ab)}{2} (\frac{1}{sup(a)} + \frac{1}{sup(b)})$	$[0, 1]$	Yes
MaxConf(a, b)	$max\{\frac{sup(a)}{sup(ab)}, \frac{sup(b)}{sup(ab)}\}$	$[0, 1]$	Yes

statistical independence requires $\overline{m\overline{c}}$ to be 1000. Therefore, the neutrality of the null-invariant measures does not necessarily suggest statistical independence. ■

Our subsequent discussions will be focused only on the five null-invariant measures in Table 3. Although a comparative study of interestingness measures has been done in [12], there are still many important yet unanswered questions on the null-invariant measures, such as “Are there inherent relationships among them?”, and “Which measure is better for evaluating interesting associations among small probability events?”. This motivates us to conduct an in-depth study to weave a well-organized picture. Specifically, we make the following contributions: (i) We show that there exists a total ordering among these measures based on a mathematical analysis. This not only explains their inherent relationships and underlying philosophy, but also provides a unified view of association analysis in large transaction datasets. And, (ii) we propose a generalized measure to handle multiple events under the unified framework. The rest of the paper is organized as follows. Section 2 presents the re-examination and generalization. Section 3 overviews the related work. Finally, Section 4 concludes the study.

2 A Re-examination of Null-Invariant Measures

2.1 Inherent Ordering Among the Measures

Given two events a and b , the support of a is denoted as $sup(a)$, and we use a for $sup(a)$ when there is no ambiguity. Let M be any of the five null-invariant

measures. From Table 3, we immediately have the following fundamental properties: **(P1)** $M \in [0, 1]$; **(P2)** M monotonically increases with $sup(ab)$ when $sup(a)$ and $sup(b)$ remain unchanged; and it monotonically decreases with $sup(a)$ (or $sup(b)$) when $sup(ab)$ and $sup(b)$ (or $sup(a)$) stay the same; **(P3)** M is symmetric under event permutations; and **(P4)** M is invariant to scaling, i.e., multiplying a scaling factor to $sup(ab)$, $sup(a)$, and $sup(b)$ will not affect the measure. These four properties justify the “conventional wisdom” about association analysis in large databases, and therefore are desirable. Specifically, $P1$ states that the value domain of M is normalized so that 0 indicates no event co-occurrence and 1 indicates that events always appear together⁴. $P2$ is consistent with the basic intuition that more co-occurrences would result in greater measure values, and vice versa. $P3$ and $P4$ show the robustness of M .

Despite such “conventional wisdom”, there are subtle cases that cannot be resolved by our common sense. Turn to D_5 and D_6 in Table 2, where m and c have unbalanced conditional probabilities – $P(m|c) > 0.9$ and $P(c|m) < 0.1$. *AllConf*, *Coherence*, and *Cosine* view both as negatively associated, *Kulc* is neutral, and *MaxConf* claims strongly positive associations. One may ask, “Which measure intuitively reflects the true relationship?” Unfortunately, there is no commonly agreed judgment for such cases due to the “balanced” skewness of the data.

Interestingly, we show that there is a total ordering that discloses the inherent relationships among the measures and thus may help the user’s decision-making. To begin with, we rewrite the definitions in Table 3 into the form of conditional probabilities in Table 4 ($P(a|b) = sup(ab)/(sup(ab) + sup(\bar{a}b))$). The rewriting of *Coherence* need the assumption that $sup(ab) \neq 0$, and for simplicity, we assume that all these measures are equal to 0 when $sup(ab) = 0$.

Table 4. Null-invariant measures defined using conditional probabilities

Measure	Definition	Exponent
<i>AllConf</i> (a, b)	$\min\{P(a b), P(b a)\}$	$k \rightarrow -\infty$
<i>Coherence</i> (a, b)	$(P(a b)^{-1} + P(b a)^{-1} - 1)^{-1}$	$k = -1$
<i>Cosine</i> (a, b)	$\sqrt{P(a b)P(b a)}$	$k \rightarrow 0$
<i>Kulc</i> (a, b)	$(P(a b) + P(b a))/2$	$k = 1$
<i>MaxConf</i> (a, b)	$\max\{P(a b), P(b a)\}$	$k \rightarrow +\infty$

Following from the rewritten definitions, we generalize all five measures using the mathematical generalized mean [7]. Each of them can be represented by the generalized mean of the two conditional probabilities $P(a|b)$ and $P(b|a)$ as

$$\mathbb{M}^k(P(a|b), P(b|a)) = \left(\frac{P(a|b)^k + P(b|a)^k}{2} \right)^{1/k}, \tag{1}$$

⁴ The only exception is that $MaxConf(a, b) = 1$ may not indicate that a and b always co-occur.

where $k \in (-\infty, +\infty)$ is the exponent of the generalized mean. As in Table 4, each measure can be generalized to Eq. (1) with the corresponding exponent.

Proof of Correctness. We first prove $AllConf(a, b) = \lim_{k \rightarrow -\infty} \mathbb{M}^k(P(a|b), P(b|a))$. Without loss of generality, let's assume that $P(a|b) \leq P(b|a)$. The proof follows from $\lim_{k \rightarrow -\infty} \mathbb{M}^k(P(a|b), P(b|a)) = \lim_{k \rightarrow -\infty} \left(\frac{1+(P(b|a)/P(a|b))^k}{2} \right)^{1/k} P(a|b) = P(a|b) = AllConf(a, b)$. The proof for $MaxConf$ has a similar argument. For $Cosine$, let $x = P(a|b)/P(b|a)$. We have $\lim_{k \rightarrow 0} \ln \left(\frac{x^k+1}{2} \right)^{1/k} = \ln(x^{1/2})$, because $\lim_{k \rightarrow 0} \ln \left(\frac{x^k+1}{2} \right)^{1/k} = \lim_{k \rightarrow 0} \frac{\ln \left(\frac{x^k+1}{2} \right)}{k} = \lim_{k \rightarrow 0} \frac{\frac{1}{2}x^k \ln x}{(x^k+1)/2} = \frac{1}{2} \ln x$. Therefore, $Cosine(a, b) = \lim_{k \rightarrow 0} \mathbb{M}^k(P(a|b), P(b|a))$. The proof for $Kulc(a, b) = \mathbb{M}^{-1}(P(a|b), P(b|a))$ is trivial. For $Coherence$ however, the equation $Coherence(a, b) = \mathbb{M}^{-1}(P(a|b), P(b|a))$ does not hold. In fact, we have $Coherence(a, b) = (2/\mathbb{M}^{-1} - 1)^{-1}$. For simplicity, we define a new measure $Coherence' = \mathbb{M}^{-1}(P(a|b), P(b|a))$ as a replacement of $Coherence$ in our following discussions. This is a reasonable replacement because $Coherence'$ preserves the ordering of $Coherence$; that is, $Coherence'(a_1, b_1) \leq Coherence'(a_2, b_2) \Leftrightarrow Coherence(a_1, b_1) \leq Coherence(a_2, b_2)$. ■

All five measures can be expressed nicely as the generalized mean of $P(a|b)$ and $P(b|a)$ except that $Coherence$ (or $Jaccard Coefficient$) need a order-preserving transformation. The generalization to $\mathbb{M}^k(P(a|b), P(b|a))$ (note that these measures only differ in terms of the exponent k) gives us two implications, summarized into the following lemmas.

Lemma 1. For any $k \in (-\infty, +\infty)$, $\mathbb{M}^k(P(a|b), P(b|a))$ always satisfies the fundamental properties P1–P4 and the null-invariance property.

PROOF. Both $P(a|b)$ and $P(b|a)$ have range $[0, 1]$, so their mean must have the same range. Also, $\mathbb{M}^k(P(a|b), P(b|a))$ is monotone *w.r.t.* $sup(a)$, $sup(b)$, and $sup(ab)$, and is invariant to event permutation, scaling, and null-transactions. ■

Lemma 2. Given any two events a and b , we have

$$AllConf(a, b) \leq Coherence'(a, b) \leq Cosine(a, b) \leq Kulc(a, b) \leq MaxConf(a, b). \tag{2}$$

PROOF. Given any exponents k and k' ($k < k'$), we have $\mathbb{M}^k(a, b) \leq \mathbb{M}^{k'}(a, b)$ [7], where the equality holds if and only if $P(a|b) = P(b|a)$. ■

These two lemmas provide insights into both sides of the coin. The first one provides a general justification to the *common*, desirable properties of the null-invariant association measures, whereas the second lemma presents an organized picture of the *differences* between them. The total ordering of the measures clearly exhibits their relationships. First, higher-order (*i.e.*, with larger k) measures provides an upper-bound to lower-order (*i.e.*, with smaller k) measures. Therefore, given a fixed interestingness threshold (*e.g.*, 0.9), the patterns output by a higher-order measure must be a superset of those by a lower-order one. This is helpful to association pattern mining, in that computationally expensive

measures such as *Cosine* that involves square root computation, is bounded by computationally cheaper measures like *Kulc*, which can be pushed deep into the mining process. Intuitively, a lower-order measure is more strict (i.e., prune more patterns), because a small k tends to mitigate the impact of the larger one of the two conditional probabilities, whereas a large k tends to aggravate it.

While the generalized mean represents a family of null-invariant measures, there is no universally accepted one for association analysis in large databases, because no particular value of k is generally better. Thus, an appropriate value of k should be determined on a case-by-case basis. It is worth noticing that each of the measures being examined is a special case in the whole spectrum of exponent k . That is, *AllConf* ($k \rightarrow -\infty$) and *MaxConf* ($k \rightarrow +\infty$) correspond to the minimum and maximum of the conditional probabilities, whereas *Coherence'* ($k = -1$), *Cosine* ($k \rightarrow 0$), and *Kulc* ($k = 1$) correspond to the *harmonic mean*, *geometric mean*, and *arithmetic mean* of the conditional probabilities.

2.2 Multiple Events

In this subsection, we extend these measures to multiple events. In order to preserve the fundamental properties and take the “generalized mean” approach for balancing conditional probabilities, we have the following definition.

DEFINITION 1. (Generalized Association Measure) Let X be an event-set containing n ($n \geq 2$) events $\{a_1, a_2, \dots, a_n\}$. The generalized measure is

$$\begin{aligned} \mathbb{M}^k(X) &= \mathbb{M}^k(P(a_2 \cdots a_n | a_1), \dots, P(a_1 \cdots a_{n-1} | a_n)) \\ &= \sqrt[k]{\frac{\text{sup}(X)^k}{n} \left(\frac{1}{\text{sup}(a_1)^k} + \cdots + \frac{1}{\text{sup}(a_n)^k} \right)}. \end{aligned}$$

■

The generalized association measure is the generalized mean of the conditional probabilities of all events. The total ordering still applies to this extension in that the smaller k is, the smaller result the measure will produce. It is worth mentioning that *AllConf* has been defined [10,9] on more than two events, which also matches this definition.

2.3 Empirical Evaluation

We choose the DBLP⁵ data set for our empirical evaluation. We extract papers from several selected data mining and database conferences including *KDD*, *SIGMOD*, and *VLDB* in recent 10 years and generate a transaction database. We show in Table 5 ten typical skewed pairs of productive authors with at least 10 papers and rank them according to their number of joint papers (i.e., $\text{sup}(ab)$). While *AllConf* and *MaxConf* have a straightforward philosophy, their results are omitted and we list the measure value of the other three measures,

⁵ <http://www.informatik.uni-trier.de/~ley/db/>

Table 5. Experiment on DBLP data set

ID	Author <i>a</i>	Author <i>b</i>	<i>sup(ab)</i>	<i>sup(a)</i>	<i>sup(b)</i>	<i>Coherence</i>	<i>Cosine</i>	<i>Kulc</i>
1	Hans-Peter Kriegel	Martin Ester	28	146	54	0.163 (2)	0.315 (7)	0.355 (9)
2	Michael Carey	Miron Livny	26	104	58	0.191 (1)	0.335 (4)	0.349 (10)
3	Hans-Peter Kriegel	Joerg Sander	24	146	36	0.152 (3)	0.331 (5)	0.416 (8)
4	Christos Faloutsos	Spiros Papadimitriou	20	162	26	0.119 (7)	0.308 (10)	0.446 (7)
5	Hans-Peter Kriegel	Martin Pfeifle	18	146	18	0.123 (6)	0.351 (2)	0.562 (2)
6	Hector Garcia-Molina	Wilburt Labio	16	144	18	0.110 (9)	0.314 (8)	0.500 (4)
7	Divyakant Agrawal	Wang Hsiung	16	120	16	0.133 (5)	0.365 (1)	0.567 (1)
8	Elke Rundensteiner	Murali Mani	16	104	20	0.148 (4)	0.351 (3)	0.477 (6)
9	Divyakant Agrawal	Oliver Po	12	120	12	0.100 (10)	0.316 (6)	0.550 (3)
10	Gerhard Weikum	Martin Theobald	12	106	14	0.111 (8)	0.312 (9)	0.485 (5)

Coherence, *Cosine*, and *Kulc*, for each pair and its rank in the parenthesis to demonstrate their similarities and differences.

It can be seen from the support in the table that at least 3 pairs of authors (ID = 5, 7, 9) demonstrate a relationship of the “advisor-advisee” style because $sup(a) \gg sup(b)$ and *b* always coauthors with *a*, but conversely, *a*, as an advisor, only coauthors a small portion of his/her papers with *b*. While *Kulc* shows relative preferences for such very skewed patterns by ranking them the top-3 most strongly associated pairs, *Cosine* and *Coherence* rank relatively balanced data higher. On the other hand, the author pairs ranked top 3 (ID = 1, 2, 3) by *Coherence* are considered to be the bottom 3 by *Kulc*, because these 3 pairs have relatively large $sup(ab)$ but the conditional probabilities are more balanced. The *Cosine* measure, as expected, stands in the middle of the other two: the top *Cosine* patterns (ID = 5, 7, 8) are ranked by *Coherence* as 4th, 5th, and 6th, and by *Kulc* as 1st, 2nd, and 6th. The same observation can be made to the bottom 3 patterns of *Cosine*. In conclusion, *Kulc* tends to give more credits to skewed patterns (e.g., advisor-advisee relationships), *Coherence* prefers balanced patterns (e.g., two comparable collaborators), and *Cosine* lies in-between.

3 Related Work

Both association and correlation mining are essential to the discovery of interesting, inherent relationships among large sets of events in a wide spectrum of applications. Various existing metrics and newly proposed measures have been studied to facilitate such analysis [11,3,12,10,6]. There are statistical correlation analysis methods. χ^2 [2] is borrowed from statistics [7] to identify correlations, considering both the absence and presence of items for interesting rules. TAPER [13], an algorithm for efficiently finding strongly correlated pairs of items, is grounded on the *Pearson’s* coefficient. Another class of work belongs to constraint-based association mining [4], where measures like *Confidence* and *Lift* are used to assist in rule generation. In [10,9] a new interestingness measures *AllConf* is defined based on a few desired properties. There are also measures widely used in other scenarios. For instance, *H-Measure* [5] is tailored for correlation analysis of deep Web-based query templates. Similarity metrics like *Cosine* distance function and

Coherence (or *Jaccard Coefficient*) are also popularly used. *Kulc*, proposed in [8], has been used in chemistry research [1].

An extensive investigation of the implications and connections between different measures has been conducted in [12], which compares a list of twenty-one interestingness metrics. The study describes three desired properties and five other key properties to compare different measures, and claims that no measure is generally better. Thus, one should match the application background against the intrinsic measure properties. Our paper can be viewed as a continued study of [12] in the context of small probability events.

4 Conclusions

We have presented a comprehensive study of null-invariant interestingness measures for mining small probability events. We show a generalization of the measures in one mathematical framework and a total ordering among them that provides an organized view. We also extend their definitions to support multiple events. For future research, it would be interesting to see how this work may influence real-world problems, such as social network analysis and clustering.

References

1. Bradshaw, J.: YAMS - Yet another measure of similarity. EuroMUG (2001), <http://www.daylight.com/meetings/emug01/bradshaw/similarity/YAMS.html>
2. Brin, S., Motwani, R., Silverstein, C.: Beyond market basket: Generalizing association rules to correlations. In: SIGMOD, pp. 265–276 (May 1997)
3. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket analysis. In: SIGMOD, pp. 255–264 (May 1997)
4. Grahne, G., Lakshmanan, L., Wang, X.: Efficient mining of constrained correlated sets. In: ICDE, pp. 512–521 (February 2000)
5. He, B., Chang, K.C.-C., Han, J.: Discovering complex matchings across web query interfaces: A correlation mining approach. In: KDD, pp. 148–157 (2004)
6. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer Academic Publishers, Dordrecht (2001)
7. Kachigan, S.: Multivariate Statistical Analysis: A Conceptual Introduction. Radius Press (1991)
8. Kulczynski, S.: Die pflanzenassoziationen der pieninen. Bulletin, 57–203 (1927)
9. Lee, Y.-K., Kim, W.-Y., Cai, Y.D., Han, J.: CoMine: Efficient mining of correlated patterns. In: ICDM, pp. 581–584 (November 2003)
10. Omiecinski, E.: Alternative interest measures for mining associations. IEEE Trans. Knowledge and Data Engineering 15, 57–69 (2003)
11. Savasere, A., Omiecinski, E., Navathe, S.: Mining for strong negative associations in a large database of customer transactions. In: ICDE, pp. 432–443 (1998)
12. Tan, P.-N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: KDD, pp. 32–41 (2002)
13. Xiong, H., Shekhar, S., Tan, P.-N., Kumar, V.: Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs. In: KDD, pp. 334–343 (2004)