# Finding Maximal Fully-Correlated Itemsets in Large Databases

Lian Duan
*Department of Management Sciences*
*The University of Iowa*
*Iowa City, IA 52241 USA*
*lian-duan@uiowa.edu*

W. Nick Street
*Department of Management Sciences*
*The University of Iowa*
*Iowa City, IA 52241 USA*
*nick-street@uiowa.edu*

***Abstract*—Finding the most interesting correlations among items is essential for problems in many commercial, medical, and scientific domains. Much previous research focuses on finding correlated pairs instead of correlated itemsets in which all items are correlated with each other. When designing gift sets, store shelf arrangements, or website product categories, we are more interested in correlated itemsets than correlated pairs. We solve this problem by finding maximal fully-correlated itemsets (MFCIs), in which all subsets are closely related to all other subsets. Putting the items in an MFCI together can promote sales within this itemset. Though some exsiting methods find high-correlation itemsets, they suffer from both efficiency and effectiveness problems in large datasets. In this paper, we explore high-dimensional correlation in two ways. First, we expand the set of desirable properties for correlation measures and study the advantages and disadvantages of various measures. Second, we propose an MFCI framework to decouple the correlation measure from the need for efficient search. By wrapping the best measure in our MFCI framework, we take advantage of likelihood ratio's superiority in evaluating itemsets, make use of the properties of MFCI to eliminate itemsets with irrelevant items, and still achieve good computational performance.**

## I. Introduction and Related Work

The analysis of relationships between items is fundamental in many data mining problems. Although we are, in general, interested in correlated sets of arbitrary size, most of the published work with regard to correlation is related to finding correlated pairs [1]. Related work with association rules [2], [3], [4] is a special case of correlation pairs since each rule has a left- and right-hand side. Support and confidence [5] produce misleading results because of the lack of comparison to the expected probability under the assumption of independence. To overcome this, lift [2], conviction [3], and leverage [6] are proposed. The above correlation measures are intuitive and straightforward, but do not have the downward-closed property [5] to reduce the computational expense. Therefore, other alternatives with a downward-closed property, collective strength [7] and all-confidence [4], were proposed. Though these measures reduce the computational expense, collective strength retrieves itemsets with irrelevant items and all-confidence is not really a correlation measure.

However, there are some applications in which we are specifically interested in correlated itemsets rather than correlated pairs. For example, we are interested in finding sets of correlated stocks in a market, or sets of correlated gene sequences in a microarray experiment. Tan [1] compared 21 different measures for correlation. Only four of the 21 measures can be used to measure the correlation within a given $k$-itemset. Dunning [8] introduced a more statistically reliable measure, likelihood ratio, which outperforms other correlation measures. It measures the overall correlation within a $k$-itemset, but cannot identify the itemsets with irrelevant items. Jermaine [9] extended Dunning's work and examined the computational issue of probability ratio and likelihood ratio. But finding correlated itemsets is much harder than correlated pairs because of three major problems. First, computing correlation for each possible itemset is an NP-complete problem [9]. Second, if there are some highly correlated items within an itemset and the rest are totally independent items, most correlation measures still indicate that the itemset is highly correlated. But no measure provides information to identify the itemsets with independent items. Third, there is no guarantee that the itemset has high correlation if any of its strict subsets are highly correlated.

Given an itemset $S = \{I_1, I_2, ..., I_m\}$, if all its subsets are closely related to all other subsets, and no irrelevant items can be removed from $S$, we consider the itemset $S$ to be a *fully-correlated itemset*. Given a fully-correlated itemset $S$, if there is no other item that can be added to generate a new fully-correlated itemset, then $S$ is a *maximal fully-correlated itemset* (MFCI). However, finding MFCIs is not easy. First, there are some overlapping items among different MFCIs. For example, given a set with four people {me, my dad, my mom, my advisor}, intuitively we hope to find the two MFCIs {me, my dad, my mom} and {me, my advisor}. Here I belong to both relationships, so I exist in both sets. Second, an itemset might not be a fully-correlated itemset even if all its strict subsets are fully-correlated itemsets. For example, an itemset like {Intel, AMD, Dell} may have many contracts involving any pair of items, but the number of contracts involving all three all together may be far less. Both Intel and AMD may cooperate with Dell to assemble computers, and Intel and AMD may cooperate on CPU technology research, but the full set does not represent an imoprtant relationship.

| Correlation Measure | Formula |
|---|---|
| Support | $tp$ |
| All-confidence | $tp/max(P(I_i))$ |
| $\chi^2$-statistic | $\sum_i \sum_j \frac{(r_{ij}-E(r_{ij}))^2}{E(r_{ij})}$ |
| Simplified $\chi^2$-statistic | $\frac{(r-E(r))^2}{E(r)}$ |
| Probability Ratio | $tp/ep$ |
| Leverage | $tp - ep$ |
| Likelihood Ratio | $Pr(tp, k, n)/Pr(ep, k, n)$ |

Table I
FORMULAS OF CORRELATION MEASURES

| Correlation Measure | P1 | P2 | P3 | P4 | P5 | P6 |
|---|---|---|---|---|---|---|
| Support | | X | | | | |
| All-confidence | | X | | | | |
| $\chi^2$-statistic | X | X | X | | | X |
| Simplified $\chi^2$-statistic | X | X | X | | X | X |
| Probability Ratio | X | X | X | | | |
| Leverage | X | X | X | X | X | |
| Likelihood Ratio | X | X | X | X | X | X |

Table II
PROPERTIES OF CORRELATION MEASURES

| Correlation Measure | Lower bound | Upper bound |
|---|---|---|
| All-confidence | $tp$ | 1 |
| Simplified $\chi^2$-statistic | $-\frac{(tp-(1-\frac{1-tp}{m})^m)^2}{(1-\frac{1-tp}{m})^m}$ | $\frac{(tp-tp^m)^2}{tp^m}$ |
| Probability Ratio | $tp * (1-\frac{1-tp}{m})^{-m}$ | $tp^{(1-m)}$ |
| Leverage | $tp - (1-\frac{1-tp}{m})^m$ | $tp - tp^m$ |
| Likelihood Ratio | $-\frac{Pr(tp,k,n)}{Pr((1-\frac{1-tp}{m})^m,k,n)}$ | $\frac{Pr(tp,k,n)}{Pr(tp^m,k,n)}$ |

Table III
UPPER AND LOWER BOUNDS

Therefore, it is more reasonable to keep three MFCIs {Intel, AMD}, {Intel, Dell}, and {AMD, Dell}.

In this paper, we describe a set of desirable properties for correlation measures and give the definition of fully-correlated itemsets which not only has a downward-closure property to reduce the computational expense but also can be incorporated with any correlation measure. Using the best correlation measure that satisfies all these desirable properties together with the fully-correlated itemset definition helps eliminate itemsets with irrelevant items and find the most interesting itemsets in a reasonable amount of time.

## II. CORRELATION MEASURES

To find high-correlation itemsets, we should find a reasonable correlation measure first. Since it is impossible to compare against every possible measure [10], in this paper we use the six criteria in Section II-B to evaluate seven best or most common measures including support,[1] all-confidence, $\chi^2$-statistic, simplified $\chi^2$-statistic, probability ratio, leverage, and likelihood ratio.

### A. Correlation Measure Formulas

Given an itemset $S = \{I_1, I_2, ..., I_m\}$ in the dataset with the sample size $n$, the actual probability $tp = P(S)$, the expected probability $ep = \prod_{i=1}^{m} P(I_i)$, and the occurrence $k = P(S) * n$, Table I shows the formulas of each measure.

For the simplified $\chi^2$-statistics, the cell $r$ corresponds to the exact itemset $S$. For likelihood ratio, $Pr(p, k, n) = \binom{n}{k} \cdot p^k \cdot (1-p)^{(n-k)}$ is the probability function of the binomial distribution.

### B. Correlation Measure Properties

Given an itemset $S = \{I_1, I_2, ..., I_m\}$, a good correlation measure $M$ should satisfy the following six key properties:

- P1: $M$ is equal to a certain constant number $C$ when all the items in the itemset are statistically independent.
- P2: $M$ monotonically increases with the increase of $P(S)$ when all the $P(I_i)$ remain the same.
- P3: $M$ monotonically decreases with the increase of any $P(I_i)$ when the remaining $P(I_k)$ and $P(S)$ remain unchanged.
- P4: The upper bound of $M$ gets closer to the constant $C$ when $P(S)$ is close to 0.
- P5: $M$ gets closer to $C$ (including negative correlation cases) when an independent item is added to $S$.
- P6: $M$ gets further away from $C$ (including negative correlation cases) with increased sample size when all the $P(I_i)$ and $P(S)$ remain unchanged.

The first three properties proposed by Piatetsky-Shapiro [6] are mandatory for any reasonable correlation measure $M$, while the last three are novel and desirable properties. The fourth property means it is impossible to find any strong positive correlation from itemsets occurring rarely. For the fifth property, we want some penalty for adding independent items in order to make fully-correlated itemsets stand out. For the last property, we hope the correlation measure can increase our confidence about the positive or negative correlation of the given itemset $S$ when we get more sample data from the same population. Table II is a summary of measures with regard to all the six properties. Only likelihood ratio and leverage satisfy Property 4. Table III shows the upper and lower bounds of all the measures.

Given the itemset $S$, support is the proportion of transactions containing $S$, and all-confidence is the ratio of its probability to that of the item with the highest probability in $S$. Although both support and all-confidence possesses the downward closure property to facilitate search, they are not designed as correlation measures. Theoretically, they lack comparison to the expected probability under the independence assumption and satisfy only the second of the six de-

[1]We use "support" and "actual probability" interchangably in this paper.

sired correlation measure properties.[2] Pratically, they share three problems. First, they are biased toward itemsets with high-support items. If an itemset $S$ consists of independent, high-support items, $Support(S)$ and $AllConfidence(S)$ will also be high despite the independence. This problem is exacerbated if we extend our search to include the presence of some items and the absence of others, since absence of a rare item is itself a high-support item. Second, they are biased to short itemsets as their value decreases monotonically with increasing itemset size. Third, the anti-monotone property makes it difficult to compare correlation among itemsets of different sizes.

The $\chi^2$ test is arguably the most popular statistical check for correlation. If an itemset contains $n$ items, $2^n$ cells in the contingency table must be considered for the $\chi^2$ statistic. The computation of the statistic itself is intractable for high-dimensional data. However, we can still use the basic idea behind the $\chi^2$-statistic to create a simplified $\chi^2$-statistic. Even if we can solve the computational problem for a given itemset, the $\chi^2$-statistics' general applicability for testing correlation within the itemset framework is still very doubtful. The problem stems from the fact that each possible event should be expected to occur at least five times for the $\chi^2$ test of independence to be valid [11].

The probability ratio is the ratio of its actual probability to its expected probability under the assumption of independence. This measure is straightforward and means how many times the itemset $S$ happens more than expected, but it favors the itemsets containing a large number of items rather than significant trends in the data.

Leverage measures the difference between $tp$ and $ep$ if all the items in $S$ are independent from each other. Since $ep$ is always no less than 0, leverage can never be bigger than $tp$. Therefore, leverage is biased to high-support itemsets. In addtion, the leverage of a given itemset remains the same no matter how many samples we get from the same population.

The likelihood ratio is similar to a statistical test based on the loglikelihood ratio described by Dunning [8]. The concept of a likelihood measure can be used to statistically test a given hypothesis, by applying the likelihood ratio test. The likelihood ratio strikes a balance between the probability ratio and the actual occurrence $k$. It favors itemsets with both high probability ratio and high occurrence. Itemsets containing a small number of items tend to have high support, but the probability ratio tends to be low, while itemsets containing a large number of items tend to have high probability ratio, but the actual occurrence tends to be low. Likelihood ratio favors middle-sized itemsets which strike a balance between the probability ratio and the actual occurrence.

[2]Due to page limitation, all mathematical proofs are omitted.

| | C | | not C | |
|---|---|---|---|---|
| | B | not B | B | not B |
| A | 25 | 25 | 25 | 425 |
| not A | 25 | 25 | 25 | 425 |

Table IV
AN INDEPENDENT CASE

## III. MAXIMAL FULLY-CORRELATED ITEMSETS

Even likelihood ratio, which penalizes itemsets with independent items, cannot detect whether a given itemset contains items that are independent of the others. To achieve that, we define fully-correlated itemsets as follows.

*Definition 1:* Given an itemset $S$ and a correlation measure, if the correlation value of any subset containing no less than two items of $S$ is higher than a given threshold $\tau$, i.e., all subsets of $S$ are closely related to all other subsets and no irrelevant items can be removed, this itemset $S$ is a fully-correlated itemset.

This definition has three very important properties. First, it can be incorporated with any correlation measure. Second, it helps to rule out an itemset with independent items. For example in Table IV, $B$ is correlated with $C$, and $A$ is independent from $B$ and $C$. Suppose we use the likelihood ratio and set the correlation threshold to be 2. The likelihood ratio of the set $\{A, B, C\}$ is 8.88 which is higher than the threshold. But the likelihood ratio of its subset $\{A, B\}$ which is 0 doesn't exceed the threshold. According to our definition, the set $\{A, B, C\}$ is not a fully-correlated itemset. The fully-correlated itemset should be $\{B, C\}$ whose likelihood ratio is 17.93. Third, there is a desirable downward closure property which can help us to prune unsatisfied itemsets quickly like Apriori [5]. The pseudocode for generating maximal fully-correlated itemsets is shown in Algorithm 1.

## IV. EXPERIMENTS

In this section, we describe the performance study used to evaluate our method. The algorithm was run on three real life datasets from different areas. The first is a subset of the Netflix data set (http://www.netflixprize.com). Since we also want to compare top-$k$ correlation patterns with MFCIs, the correlation among only the first 100 movies (according to the code sequence in the Netflix dataset) is checked. The second dataset is the anonymous Microsoft web data set from the UCI Repository [12] containing web browsing records of 38,000 anonymous users to 294 areas of the web site. The third dataset contains 15,000 nursing careplans from a Midwest hospital for the year 1996.

### A. Comparison of MFCIs to Top-k Correlation Itemsets with Likelihood Ratio

Since different MFCIs and top-$k$ correlation itemsets will be retrieved using different correlation measures, in this

**Algorithm 1** Find Fully-Correlated Itemsets

**Main:** FindFullyCorrelatedItemsets($L_2$: fully-correlated 2-itemsets)
  //$LC$ is the Vector containing maximal fully-correlated itemsets
  **for** $k = 3; L_{k-1} \neq \emptyset; k++$ **do**
    $C_k$ = GenerateHighLevelItemset($L_{k-1}$);
    **for** each candidate $c \in C_k$ **do**
      **if** $CalculateCorrelation(c) > correlationThreshold$ **then**
        add $c$ to $L_k$;
      **end if**
    **end for**
    $LC = LC \cup FindMaximalItemset(L_{k-1}, L_k)$
  **end for**
  **return** $LC$;

**Procedure:** GenerateHighLevelItemset($L_{k-1}$: $(k-1)$-itemsets)
  **for** each itemset $l_1 \in L_{k-1}$ **do**
    **for** each itemset $l_2 \in L_{k-1}$ **do**
      **if** $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge ... \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$ **then**
        $c = l_1 \bowtie l_2$; //join step: generate candidates
        **if** not HasUncorrelatedSubset($c$, $L_{k-1}$) **then**
          add $c$ to $C_k$;
        **end if**
      **end if**
    **end for**
  **end for**
  **return** $C_k$;

**Procedure:** HasUncorrelatedSubset($c$: $k$-itemset candidate; $L_{k-1}$: $(k-1)$-itemsets)
  **for** each $(k-1)$-subset $s$ of $c$ **do**
    **if** $s \notin L_{k-1}$ **then**
      **return** TRUE;
    **end if**
  **end for**
  **return** FALSE;

**Procedure:** FindMaximalItemsets($L_{k-1}$: $(k-1)$-itemsets; $L_k$: $k$-itemsets)
  **for** each $k$-itemset $l1 \in L_k$ **do**
    **for** each $(k-1)$-subset $l2$ of $l1$ **do**
      **if** $l2 \in L_{k-1}$ **then**
        remove $l2$ from $L_{k-1}$;
      **end if**
    **end for**
  **end for**
  **return** $L_{k-1}$;

| ID | Netflix | Website | Careplan |
|---|---|---|---|
| 1 | 03, 97 | 1000, 1014 | 005, 037 |
| 2 | 12, 90 | 1008, 1034 | 005, 078 |
| 3 | 17, 26 | 1009, 1035 | 010, 064 |
| 4 | 18, 57 | 1009, 1037 | 012, 037 |
| 5 | 33, 68 | 1009, 1074 | 013, 087 |
| 6 | 46, 78 | 1014, 1040 | 025, 064 |
| 7 | 47, 79 | 1025, 1026 | 026, 054 |
| 8 | 57, 79 | 1027, 1041 | 039, 067 |
| 9 | 58, 78 | 1032, 1056 | 054, 064 |
| 10 | 05, 69, 91 | 1036, 1040 | 054, 087 |
| 11 | 05, 91, 92 | 1038, 1053 | 066, 069 |
| 12 | 16, 58, 77 | 1041, 1070 | 013, 067, 069, 099 |
| 13 | 18, 44, 83 | 1052, 1060 | 015, 065, 067, 069 |
| 14 | 28, 48, 84 | 1026, 1038, 1041 | 065, 067, 068, 069 |
| 15 | 28, 58, 77 | 1001, 1003, 1018, 1035 | 065, 067, 069, 099 |

Table V
MAXIMAL FULLY-CORRELATED ITEMSETS

section we only compare those retrieved by likelihood ratio which satisfies all six desirable properties and works well practically. Like maximal frequent itemsets, there is no ranking of MFCIs. Rather, we get a different number of them under different thresholds. Therefore, we find the threshold under which 15 MFCIs are retrieved to compare with top-20 correlation itemsets, retrieved by calculating every possible combination. The 15 MFCIs of each dataset are shown in Table V, and the top-20 correlation sets are listed in Table VI.[3] Our MFCIs have several advantages over the top-20 correlation sets. First, some top-$k$ correlation sets are redundant since they are subsets of other top-$k$ correlation sets. For example, the first correlation set of Netflix is a subset of the second one; in Website, the fifth set is a subset of the third; in Careplan, the third set is a subset of the first. Second, some top-$k$ correlation itemsets contain irrelevant information. Among the top-20 correlation sets in Netflix, 16 are subsets of the 15 MFCIs. The four remaining itemsets all contain one more movie, "Something's Gotta Give (2003)" (code 30), than the corresponding maximal correlation sets. In fact, "Something's Gotta Give (2003)" is the most favored movie among all the 100 films and has almost no correlation with any other movie. Removing it from these four itemsets results in higher correlation values. Similary in the website dataset, the likelihood ratio between "Microsoft.com Search" (code 1004) and "isapi" (code 1018) is 0.70, but both of them are contained in the 17th correlation set. In the careplan dataset, "Health Maintenance" (code 28) and "Knowledge Deficit" (code 52) are assigned almost in every

[3]Codes used in the three datasets are as follows. **Netflix:** (03) Character, 1997; (05) The Rise and Fall of ECW, 2004; (12) My Favorite Brunette, 1947; (16) Screamers, 1996; (17) 7 Seconds, 2005; (18) Immortal Beloved, 1994; (26) Never Die Alone, 2004; (28) Lilo and Stitch, 2002; (30) Something's Gotta Give, 2003; (33) Aqua Teen Hunger Force: Vol. 1, 2000; (44) Spitfire Grill, 1996; (46) Rudolph the Red-Nosed Reindeer, 1964; (47) The Bad and the Beautiful, 1952; (48) Justice League, 2001; (57) Richard III, 1995; (58) Dragonheart, 1996; (68) Invader Zim, 2004; (69) WWE: Armageddon, 2003; (77) Congo, 1995; (78) Jingle All the Way, 1996; (79) The Killing, 1956; (83) Silkwood, 1983; (84) The Powerpuff Girls Movie, 2002; (90) The Lemon Drop Kid, 1951; (91) WWE: Royal Rumble, 2005; (92) ECW: Cyberslam, 2002; (97) Mostly Martha, 2002; **Website:** (1000) regwiz; (1001) Support Desktop; (1003) Knowledge Base; (1004) Microsoft.com Search; (1008) Free Downloads; (1009) Windows Family of OSs; (1014) Office Free Stuff; (1017) Products; (1018) isapi; (1025) Web Site Builder's Gallery; (1026) Internet Site Construction for Developers; (1027) Internet Development; (1032) Games; (1034) Internet Explorer; (1035) Windows95 Support; (1036) Corporate Desktop Evaluation; (1037) Windows 95; (1038) SiteBuilder Network Membership; (1040) MS Office Info; (1041) Developer Workshop; (1052) MS Word News; (1053) Jakarta; (1056) sports; (1060) MS Word; (1070) ActiveX Technology Development; (1074) Windows NT Workstation; **Careplan:** (005) Anxiety; (010) Breathing Pattern Ineffectiveness; (012) Pain Acute; (013) Communication, Impaired Verbal; (015) Coping Ineffectiveness; (025) Gas Exchange Impairment; (026) Grieving, Anticipatory; (028) Health Maintenance, Altered; (037) Infection, Risk For; (039) Injury, High Risk For; (052) Knowledge Deficit; (054) Nutrition: Less Than Body Requirements Altered; (064) Activity Intolerance; (065) Self Care Deficit, Dressing/Grooming; (066) Self Care Deficit, Feeding; (067) Self Care Deficit, Bathing/Hygiene; (068) Physical Mobility Alteration; (069) Self Care Deficit, Toileting; (078) Skin Integrity, Impaired; (087) Thought Process, Altered; (099) Aspiration, Risk.

nursing care plan. The likelihood ratio between {028} and {065,067,068,069} is 0.27 and the likelihood ratio between {052} and {065,067,068,069} is -1.3 which means they are almost independent, and both the 2nd and the 7th correlation sets contain irrelevant information. Therefore, the MFCIs are more reasonable than the top-$k$ correlation itemsets.

| Ranking | Netflix | Website | Careplan |
|---|---|---|---|
| 1 | 58, 77 | 1001, 1003, 1018 | 065, 067, 068, 069 |
| 2 | 28, 58, 77 | 1009, 1018, 1035 | 028, 065, 067, 068, 069 |
| 3 | 05, 91 | 1001, 1003, 1018, 1035 | 065, 067, 069 |
| 4 | 44, 83 | 1008, 1009, 1018, 1035 | 028, 065, 067, 069 |
| 5 | 12, 90 | 1001, 1018, 1035 | 067, 068, 069 |
| 6 | 18, 44, 83 | 1003, 1018, 1035 | 039, 065, 067, 068, 069 |
| 7 | 16, 58, 77 | 1001, 1003 | 052, 065, 067, 068, 069 |
| 8 | 28, 58 | 1001, 1003, 1035 | 028, 039, 065, 067, 068, 069 |
| 9 | 28, 30, 58, 77 | 1018, 1035 | 065, 068, 069 |
| 10 | 05, 91, 92 | 1009, 1018, 1035, 1037 | 028, 052, 065, 067, 068, 069 |
| 11 | 33, 68 | 1009, 1017, 1037 | 065, 067, 068 |
| 12 | 30, 44, 83 | 1009, 1037 | 028, 067, 068, 069 |
| 13 | 47, 79 | 1026, 1038, 1041 | 015, 028, 065, 067, 068, 069 |
| 14 | 46, 78 | 1001, 1009, 1018, 1035 | 015, 028, 039, 065, 067, 068, 069 |
| 15 | 18, 57 | 1026, 1038 | 015, 039, 065, 067, 068, 069 |
| 16 | 18, 30, 44, 83 | 1025, 1026 | 015, 065, 067, 068, 069 |
| 17 | 30, 58, 77 | 1001, 1003, 1004, 1018 | 039, 065, 067, 069 |
| 18 | 05, 69, 91 | 1008, 1018, 1035 | 028, 065, 068, 069 |
| 19 | 05, 92 | 1003, 1009, 1018, 1035 | 015, 028, 039, 052, 065, 067, 068, 069 |
| 20 | 16, 58 | 1001, 1003, 1009, 1018, 1035 | 039, 052, 065, 067, 068, 069 |

Table VI
CORRELATION RANKING LIST

### B. Comparison of Likelihood Ratio to Other Measures

If we specify a threshold to retrieve 10 MFCIs using probability ratio, most of them just coincidently happen once or twice. From Table VII, the average support of the 10 MFCIs is 9.4 in Netflix, 1.4 in Website, and 1 in Careplan. The result gets much worse when we check the top-10 correlation itemsets. Each of them contains at least 16 movies in Netflix, 12 areas in Website, and 9 nursing diagnoses. All the average supports of the top-10 correlation itemsets are 1. None of them are interesting patterns, so we conclude that probability ratio is a poor correlation measure.

If we retrieve 10 MFCIs with the simplified $\chi^2$-statistic, some sets just coincidently happen once or twice. From Table VII, the average support of the 10 MFCIs is 184.8 in Netflix, 94.8 in Website, and 165.9 in Careplan. Although the simplified $\chi^2$-statistic is still biased to low-support itemsets, it is better than probability ratio.

If we retrieve 10 MFCIs using leverage, most of them are as reasonable as likelihood ratio. However, it has a bias to high-support itemsets even if the correlation among them is not strong. On the other hand if the support of a itemset $S$ is no more than the threshold we specify, the itemset $S$ will never be retrieved even if items in $S$ are strongly correlated. For example, the 2nd set {My Favorite Brunette (1947) (code 12), The Lemon Drop Kid (1951) (code 90)} in Table V can be retrieved by likelihood ratio, but not by leverage. The occurrence of "My Favorite Brunette (1947)" is 248, of "The Lemon Drop Kid (1951)" is 313, and the two together is 103, from a total of 138805 records. If these two movies are independent, their expected co-occurrence is 0.56, far smaller than 103, so these two movies are strongly correlated. But the support of {My Favorite Brunette (1947), The Lemon Drop Kid (1951)} is less than 0.001, so leverage fails to find this pattern. A similar situation happens in Website with {1052, 1060} and in Careplan with {013, 087}. If the utilities of the items [13] are equal, we might be interested in the correlation result biased on high support items. In this case, leverage is better. But when the utility of items is different, we might be interested in each strong correlation regardless of the actual occurrence. In this case, likelihood ratio is better.

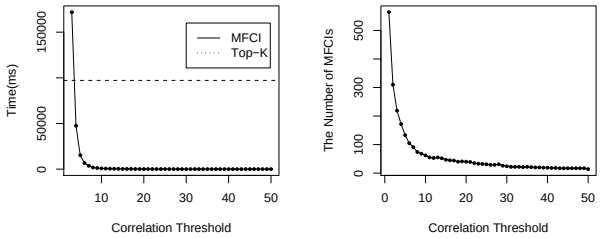
### C. Performance

The performance of our algorithm depends on the characteristic of the data set. In the extreme case, among all the $n$ items in the data set, if the first $(n-1)$ items always show up together and the remaining item appears alone, our algorithm will start with $\binom{n-1}{2}$ 2-itemsets and end with the only $(n-1)$-maximal fully-correlated itemset. In other words, all the $2^{(n-1)}$ possible combinations will be checked. It is still an NP-hard problem. But in reality, most data sets are sparse, so most of the search space can be pruned.

Since all three datasets have similar performance patterns, we only show the performance patterns on the Netflix dataset. The runtime of our algorithm on the Netflix data given different correlation thresholds are shown in Figure 1(a). The runtime decreases drastically as we increase the threshold. The runtime of the top-$k$ method is also shown. When running the top-$k$ method, we only checked the itemsets which occurred at least once in the dataset, so about 2 million itemsets were checked instead of all $2^{100}$ possible itemsets. In the worst case, our algorithm checks all the itemsets which occur at least once. Therefore, the runtime of our algorithm at least as good as the top-$k$ method if both use Apriori. Even though we use the prefix

| | | | Netflix | Website | Careplan |
|---|---|---|---|---|---|
| 10 MFCIs | Threshold | Likelihood Ratio | 100 | 230 | 31.2 |
| | | Leverage | 0.0018 | 0.018 | 0.009 |
| | | Simplified $\chi^2$ | 1450 | 5400 | 530 |
| | | Probability Ratio | 250 | 800 | 99.7 |
| | | All-confidence | 0.11 | 0.35 | 0.25 |
| | Average Support | Likelihood Ratio | 610.5 | 1159.7 | 703.3 |
| | | Leverage | 617 | 1342.9 | 1115.8 |
| | | Simplified $\chi^2$ | 184.8 | 94.8 | 165.9 |
| | | Probability Ratio | 9.4 | 1.4 | 1 |
| | | All-confidence | 1417.6 | 1085.1 | 1964 |
| | Average ItemsetSize | Likelihood Ratio | 2.14 | 2.38 | 3 |
| | | Leverage | 2.14 | 2.58 | 3.32 |
| | | Simplified $\chi^2$ | 2.38 | 2.77 | 2.49 |
| | | Probability Ratio | 2.38 | 2.77 | 2 |
| | | All-confidence | 2 | 2.77 | 2.38 |
| Top 10 Sets | Average Correlation | Likelihood Ratio | 207.45 | 698.73 | 483.41 |
| | | Leverage | 0.0044 | 0.036 | 0.056 |
| | | Simplified $\chi^2$ | 4.04 E31 | 3.25 E26 | 1.06 E13 |
| | | Probability Ratio | 4.04 E31 | 3.25 E26 | 1.06 E13 |
| | | All-confidence | 0.157 | 1 | 0.558 |
| | Average Support | Likelihood Ratio | 685.3 | 866.4 | 282.1 |
| | | Leverage | 1176 | 1849.1 | 3700.1 |
| | | Simplified $\chi^2$ | 1 | 1 | 1 |
| | | Probability Ratio | 1 | 1 | 1 |
| | | All-confidence | 1417.6 | 1 | 5844.1 |
| | Average ItemsetSize | Likelihood Ratio | 2.6 | 3.1 | 4.4 |
| | | Leverage | 2.4 | 2.2 | 3.3 |
| | | Simplified $\chi^2$ | 17.1 | 12.8 | 10.2 |
| | | Probability Ratio | 17.1 | 12.8 | 10.2 |
| | | All-confidence | 2 | 2.4 | 2.2 |

Table VII
SUMMARY OF VARIOUS CORRELATIONS




(a) Runtime for each threshold (b) MFCIs for each threshold

Figure 1. Performance results for Netflix

tree structure [14] to find the top-$k$ correlation itemsets, the runtime of our method is less than the top-$k$ method when the threshold is larger than 3. In addition, the number of maximal fully-correlated itemsets corresponding to each correlation threshold is shown in Figure 1(b). It shows even if we use a small threshold like 1, we still get a relatively small number of compact itemsets which is 564.

## V. CONCLUSION

This paper presents several key properties for choosing a good correlation measure and a useful definition of fully-correlated itemsets which decouples the correlation measure from the need for efficient search. Among the existing correlation measures, likelihood ratio and leverage are the best. Given our definition of fully-correlated itemsets, we can find more compact and useful information than top-$k$ correlation itemsets. Due to the desirable downward closure property of fully-correlated itemsets, we can discover strong patterns in a reasonable amount of time.

## REFERENCES


[1] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.

[2] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 1997, pp. 265–276.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 1997, pp. 255–264.

[4] E. R. Omiecinski, "Alternative interest measures for mining associations in databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 1, pp. 57–69, 2003.

[5] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 1993, pp. 207–216.

[6] G. Piatetsky-Shapiro, *Discovery, Analysis, and Presentation of Strong Rules*. AAAI/MIT Press, 1991.

[7] C. C. Aggarwal and P. S. Yu, "A new framework for itemset generation," in *Proc. 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1998, pp. 18–24.

[8] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[9] C. Jermaine, "Finding the most interesting correlations in a database: How hard can it be?" *Information Systems*, vol. 30, no. 1, pp. 21–46, 2005.

[10] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys*, vol. 38, no. 3, p. 9, 2006.

[11] K. Fujisawa, Y. Hamuro, N. Katoh, T. Tokuyama, and K. Yada, "Approximation of optimal two-dimensional association rules for categorical attributes using semidefinite programming," in *Proc. 2nd Int. Conf. on Discovery Science*, 1999, pp. 148–159.

[12] A. Asuncion and D. Newman, "UCI machine learning repository," http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007.

[13] Y. Liu, W.-K. Liao, and A. Choudhary, "A fast high utility itemsets mining algorithm," in *Proc. 1st Int. Workshop on Utility-based Data Mining*, 2005, pp. 90–99.

[14] C. Borgelt, "Efficient implementations of Apriori and Eclat," in *Proc. 3rd IEEE Int. Conf. on Data Mining, Workshop on Frequent Itemset Mining Implementations*, 2003.