1 Product You Might Buy

(a) Because the confidence only tells you how much probability B occurs when A is already existed. It only measures the probability that the rule is correct in the context of A. However, the occurrence of the rule also has something to do with the frequency of B in the dataset. Therefore, a high confidence value can also be achieved if B occurs very commonly in the dataset, even if there is no association between A and B. And thus, A won't have a strong influence on B and the rule is not convincing. It will lead to the drawback of misleading and cause the rule to be uninformative, since people would believe that the occurrence of B would depend on the existence of A if the confidence is high.

Lift takes the probability of the occurrence of B to be the denominator. Therefore, less occurrence of B would lead to smaller $Pr(B)$, and the value of lift would be larger, suggesting that the reason A and B occur together is more because the rule is effective, but depends less on the occurrence frequency of B. Vice versa when the occurrence of B become larger. If lift is greater than 1, it indicates A may have a positive association with B and the rule may be effective. And if it's less than 1, it indicates A and B are not strongly related due to the low confidence of the rule or the probability of B is high enough such that the rule is not convincing. Consequently, lift eliminates the influence of the probability that B occurs, and therefore does not suffer from the drawback.

Interest eliminates the influence by calculating the difference between the confidence and the probability that B occurs. Like lift, less occurrence of B would also lead to smaller $Pr(B)$, and the value of interest would also be larger, suggesting that the reason A and B occur together is more because A is associating with B, and vice versa. Those with high positive or negative interest values are interesting rules, indicating the rule is more effective or more impossible to happen. Thus, interest also does not suffer from the drawback, by getting rid of the fraction caused by the high occurrence of B.

(b) **Confidence:** not symmetrical.

Counterexample: consider $Pr(B) = 0.4,\ Pr(A) = 0.2$ and $Pr(A \cap B) = 0.1$. Therefore, $conf(A \to B) = Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = 0.5$, while $conf(B \to A) = Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = 0.25$. $conf(A \to B) \neq conf(B \to A)$.

Lift: symmetrical.

Proof: $lift(A \to B) = \frac{conf(A \to B)}{Pr(B)} = \frac{Pr(B|A)}{Pr(B)} = \frac{Pr(A \cap B)}{Pr(B)Pr(A)} = \frac{Pr(A|B)}{Pr(A)} = \frac{conf(B \to A)}{Pr(A)} = lift(B \to A)$.

Interest: not symmetrical.

Counterexample: consider $Pr(B) = 0.4,\ Pr(A) = 0.2$ and $Pr(A \cap B) = 0.1$. Therefore, $interest(A \to B) = conf(A \to B) - Pr(B) = Pr(B|A) - Pr(B) = \frac{Pr(A \cap B)}{Pr(A)} - Pr(B) = 0.1$, while
$interest(B \to A) = conf(B \to A) - Pr(A) = Pr(A|B) - Pr(A) = \frac{Pr(A \cap B)}{Pr(B)} - Pr(A) = 0.05$.
$interest(A \to B) \neq interest(B \to A)$.

(c) Top 5 rules for pairs:

```
DAI93865 -> FRO40251: 1.0
GRO85051 -> FRO40251: 0.999176276771005
GRO38636 -> FRO40251: 0.9906542056074766
ELE12951 -> FRO40251: 0.9905660377358491
DAI88079 -> FRO40251: 0.9867256637168141
```

(d) Top 5 rules for triples:

```
(DAI23334, ELE92920) -> DAI62779: 1.0
(DAI31081, GRO85051) -> FRO40251: 1.0
(DAI55911, GRO85051) -> FRO40251: 1.0
(DAI62779, DAI88079) -> FRO40251: 1.0
(DAI75645, GRO85051) -> FRO40251: 1.0
```

2 Min-hashing

(a) Since for each permutation, we have

x	$h_1(x)$	$h_2(x)$	$h_3(x)$
0	1	2	2
1	3	5	1
2	5	2	0
3	1	5	5
4	3	2	4
5	5	5	3

Therefore, the signature of S_1 is $[5,\ 2,\ 0]^T$; the signature of S_2 is $[1,\ 2,\ 1]^T$; the signature of S_3 is $[1,\ 2,\ 4]^T$; the signature of S_4 is $[1,\ 2,\ 0]^T$.

(b) According to the chart above, h_3 produce true permutations.

3 Estimation via Sampling

(a) We choose (universityID, studentID) to be the key attribute. Since the question ask about the fraction of students, we use this key to distinguish each student uniquely, so it is unbiased. Sample construction: hash each data tuple's key uniformly into 20 buckets, take the data tuple if it hashes to the first bucket.

(b) We choose (universityID, courseID) to be the key attribute. Since the question ask about the fraction of courses, the university ID together with course ID can distinguish each course uniquely, so it is also unbiased. Sample construction: Hash each data tuple's key uniformly into 20 buckets, take the data tuple if it hashes to the first bucket.